In the Lab and Out in the Wild: Remote Web Usability Testing for Mobile Devices

Sarah Waterson, James A. Landay
Computer Science Division
U.C. Berkeley
Berkeley, CA 94720-1776 USA
{waterson, landay}@cs.berkeley.edu

Tara Matthews
Computer Science Department
Seattle University
Seattle, WA 98122 USA
matthewt@seattleu.edu

ABSTRACT
In this paper we discuss a pilot usability study using wireless Internet-enabled personal digital assistants (PDAs). We compared usability data gathered in traditional lab studies with a proxy-based clickstream logging and analysis tool. We found that this remote testing technique can more easily gather many of the content-related usability issues, but device-related issues are more difficult to capture.

Keywords
Remote Usability Testing, Clickstreams, Log File Visualization, Usability Evaluation, Mobile Internet.

INTRODUCTION
Understanding where and when users experience difficulties while performing tasks on a web site is critical to improving the design of a site. As more users enter the world of mobile computing, web designers are challenged even further by moving from the large screens and familiar input devices of the desktop computer, to the small, pocket-sized screens and limited interaction techniques of mobile devices, such as PDAs and WAP-enabled cellular phones [4]. Gathering good usability data is vital to making these interfaces, and subsequently these devices, successful.

Close observation and interviewing of a select few users in a usability lab is the most common technique for testing mobile devices, often involving elaborate set-ups for video-taping and capturing of user actions [1]. Another method for gathering usability information is through remote logging of clickstream data [2]. While traditional usability testing methods provide good qualitative feedback, they are often expensive and time consuming. Furthermore, setting up a mobile device in a lab, often clamped down in front of a camera, pushes the evaluation far from the realistic use environment. Gathering clickstream data often means more users can be tested outside of the lab, but lacks qualitative information, such as spoken comments, found in the lab.

Our pilot usability study was motivated by a need to find design implications for running better remote usability tests on mobile devices, and to improve our analysis tools. We also hoped to gain insights on how to better support user testing of mobile Internet devices in situations where users might be accessing the Internet on the go.

EXPERIMENT DESIGN
We had ten participants find safety and local dealer information for the latest Nissan Sentra model on the Edmunds.com PDA web site (http://pda.edmunds.com). The participants used a Handspring Visor Edge with an OmniSky wireless modem. Half of the participants performed the usability test in a traditional lab set-up, with the PDA fixed to a table and video cameras recording both audio and user interaction with the device. Additionally, an observer was present to record comments and answer any questions relating to test administration. The other five users were given the PDA and modem and asked to perform the task when and wherever they chose. For these participants, no observer was present, nor was any form of audio or video recording done. Instead, clickstream data was collected remotely.

All task directions, demographic data and follow-up questionnaire data, for both the lab and the remote participants, was presented and collected in forms on the device itself. As each of the ten participants performed the task, their clickstream data was logged.

The WebQuilt System
To gather the clickstream data, we used WebQuilt, a proxy-based clickstream logging and visualization system [3]. Logging is done through a proxy, capturing user traces of a given task without having to modify the participants software or have access to the web site server. These clickstreams are then aggregated and visualized in a zooming and filtering interface that shows the web pages people viewed, where they clicked on each page, the most common paths taken through the web site for the task, and the designer's designated path. Figure 1 shows the WebQuilt visualization interface and clickstream data for all ten participants in this experiment.

The Usability Data
To compare the usability data collected in the lab with WebQuilt data, we needed to classify and organize the usability issues found with each method. For the lab studies, we took all of the observations made by the tester, as well as the comments and questionnaire data of the

Figure 1: The WebQuilt visualization. Nodes represent visited web pages, and arrows represent the traffic between the pages. Entry pages are colored green, and exit pages cyan. Thicker arrows representing heavier traffic. Arrow color is used to indicate time spent on a page before transitioning, where the closer the arrow to red, the longer spent in transition. The designer's path is highlighted in blue. The slider interface along the left hand side allows the designer to zoom into the graph, viewing actual images of the pages users saw, and where they clicked.

participants, and categorized them. The categories were based on whether or not the issues were related to the testing framework, to the design of the web site, to the PDA web browser, or to interactions with the device itself. We also recorded when and where the issue occurred, the total number of participants encountering each issue, its implications, and finally assigned the issue a severity rating from 1-5, or a zero if the issue was only a comment. The same experimenter classified the WebQuilt data similarly, using the visualization and questionnaire data alone to identify issues.

FINDINGS AND ANALYSIS

For the Edmunds.com task, we found a total of 18 issues, summarized in Table 1. All 18 issues were found in the lab study, and only 7 of them were found with WebQuilt. However, 5 of the 7 design issues were found with WebQuilt, including three of the four higher severity issues. The high severity issue not found with the clickstream or survey data was the user attempts to interact with the screen before the page fully returned.

The WebQuilt methodology also revealed information on the test design and problems with the device. Survey data, for both the lab and the remote participants, indicated difficulty remembering the task description, as well as difficulty scrolling. Because of the nature of these remote tests, browser and device related issues are difficult to identify with WebQuilt unless specifically addressed in the questionnaires or reported as a comments by the participants. The browser and device issues found in this experiment account for one third of all the issues.

In further tests of the system, a number of other concerns about remote usability testing were raised. In particular, problems with running the test are difficult to resolve remotely, as are understanding user motivations such as

Browser	Device
• Interact before load (3) • No forward button (2)	• Difficulty with input in questionnaire (3) • **Difficulty scrolling** (2) • Device errors unrelated to testing (1) • Tried writing on screen (0)
Site Design	**Test Design**
• **Falsely completed task** (4) • **Long download times** (4) • **Ping-pong behavior** (3) • Interact before load (3) • **Too much scrolling** (2) • Save address functionality not clear (1) • **Back button navigation** (0) • Would like more features (0) • Finds site useful (0)	• **Falsely completed task** (4) • **Difficulty remembering task description** (3) • Difficulty with input in questionnaire (3) • Questionnaire wording problems (3) • Forgot how to end task (1) • Confusing task description (1)

Table 1: Each cell in this table is an issue category. Each issue is listed, along with its severity in parenthesis, with those in bold indicating issues found in the remote usability tests. Some issues appear in more than one category.

curiosity to explore other areas of a site. With WebQuilt, curiosity could easily be interpreted as confusion. WebQuilt can not easily capture any interactions with the device that do not cause communication with the server, thus actions such as clicks on non-links or clicks on the screen before the content is fully loaded are not recorded. Other problems occur when participants misinterpret the actual tasks and questions asked. For PDAs, an added difficulty is gathering text input from the user.

FUTURE WORK AND CONCLUSIONS

We plan to extend our user studies to WAP phones and other mobile devices. We are also investigating better remote recording methods for qualitative data, including audio. We plan on including professional web designers, testing more sites under design, and performing more rigorous experimental designs in future evaluations.

Two main conclusions can be drawn here. The first is that clickstreams and remote testing seem to be pretty good at finding usability issues related to the *web content* displayed on mobile devices. However, the second conclusion is that this methodology is a poor substitute for finding usability issues with the *device* itself. Understanding how to better design for mobile devices requires knowledge of both these aspects: content design and device design. We have shown promising methods for gathering good content-related usability information remotely, and are investigating ways to improve and enhance these methods.

REFERENCES

1. *Wireless Phone Usability Research*. 2001, Usable Products Company: New York City. p. 2-3.
2. Brainard, J. and B.Becker. *Case Study: E-Commerce Clickstream Visualization*. In *IEEE Proceedings of Information Visualization*. 2001. San Diego, CA.
3. Hong, J.I., et al., *WebQuilt: A Proxy-based Approach to Remote Web Usability Testing*. To appear in ACM Transactions on Information Systems, 2001.
4. Nielsen, J. *Mobile Devices Will Soon Be Useful*. http://www.useit.com/alertbox/20010916.html. 2001.