An Empirical Comparison of Lab and Remote Usability Testing of Web Sites

Tom Tullis, Stan Fleischman, Michelle McNulty, Carrie Cianchette, and Marguerite Bergel
Human Interface Design Dept,
Fidelity Investments
82 Devonshire St., V4A
Boston, MA 02109 USA
Contact: tom.tullis@fmr.com

ABSTRACT
This paper presents the results of two studies comparing traditional lab-based usability testing and remote Web-based usability testing of Web sites. Two sites were tested: an employee benefits site and a financial information site. The remote tests used an automated technique whereby the users participated from their normal work locations using their normal browser, and there was no real-time observation. Tasks were presented to the user in a small browser window at the top of the screen that was also used to capture their input. Results showed high correlations between lab and remote tests for the task completion data and the task time data. Subjective ratings of the sites did not correlate as well. The most critical usability issues with the sites were identified by both techniques, although each technique also uniquely uncovered other issues. In general, the results indicate that both the lab and remote tests capture very similar information about the usability of a site. Each type of test appears to offer its own advantages and disadvantages in terms of the usability issues it can uncover.

Keywords
Usability testing, evaluation, remote testing, Web usability, Web-based testing.

INTRODUCTION
Laboratory-based usability tests have been around for quite a few years. Their effectiveness as a way of uncovering usability problems with Web sites and other applications is widely accepted [e.g., 1,2]. A typical lab-based usability test involves a relatively small number of representative users (e.g., six to eight) [3,4,5] coming to the lab individually and performing a series of tasks using an application or prototype. One or more observers in the same room or an adjacent room commonly record the time it takes the users to complete each task, whether they were successful, and any significant comments or problems. Subjective feedback (rating scales, written comments) is usually collected as well. All of this information is then used to develop a list of "usability issues" or potential problem areas in the application.

In recent years, various techniques for extending usability testing beyond the lab have emerged [e.g., 6,7]. One reason for this has been the cost associated with traditional lab-based testing, which tends to be rather expensive both in terms of facilities and the staff required for testing. Other reasons have centered on perceived shortcomings of the traditional lab-based techniques, such as the limited number of test participants commonly used (again, primarily due to costs and lab time).

The earliest remote usability testing techniques used the same basic techniques as lab tests, but allowed the test users and the observers to be in two different locations, geographically separate. They used special software or video connections that allow the observer to see what is happening on the user's screen, along with the use of a telephone for audio communication. This can be thought of as a "video conference" approach to usability testing. One study [8] has demonstrated that this technique yields comparable results to a lab-based test of the same application. While this technique potentially saves some travel and facilities costs, it is still a very labor- and time-intensive process, with the observers involved full-time for each test user's session.

More recently, a different form of remote usability testing has emerged, particularly for testing Web sites and applications. The key difference with this approach is that the observer is no longer "in the loop" in real time (remotely or locally) during the usability test. All data collection during the test is done automatically and stored for future analysis. The key advantage this technique offers is that many more test users can participate (in parallel), with little or no incremental cost per participant.

Several different approaches to this "unattended" remote usability testing could be taken. For example, Vividence Corporation [9] offers a type of remote usability testing in which the test users must download and use a specially instrumented browser that can capture the user's "clickstreams" as well as screen shots, and transmit those back to the host site for analysis. While this approach provides a very rich set of data, we decided that we could not ask our

test users to download a special browser. We also felt that our test users would be concerned about the potential for transmission of sensitive information if the site being tested included such items as personal financial data. For those reasons, we decided to take an approach which utilizes the standard HTML and Javascript capabilities of the normal browsers, and which clearly limits the information being returned to the host.

OUR APPROACH TO REMOTE USABILITY TESTING

In our approach to remote testing, the entire test is conducted over the Web, with the users participating from their normal work (or home) locations using their own computer. These users could be anywhere, as long as they can access the Web. Typically, participants are recruited via an email message that includes a link to a "Welcome" page explaining the characteristics of the test. After starting the test via a link on that page, two independent browser windows are opened, as shown in Figure 1:

- A small browser window across the top of the screen is used to present tasks to the user, individually, and to capture user input, feedback, and ratings.
- The main browser window, which fills the majority of the screen, is used to present the Web application or site being tested.

It's important to realize that there is no communication from the main browser window to the small task window (e.g., the task window gets no information about pages visited in the main window). The site in the main browser window does not have to be modified in any way for the test. It can be a production site or a prototype. None of the code for conducting the remote test is contained within the main site.

The tasks presented to the users typically require them to find the answer to a relatively specific question by using the site in the main browser window. While trying to find the answer, the users are encouraged to provide any comments they wish using a comment field in the task window.

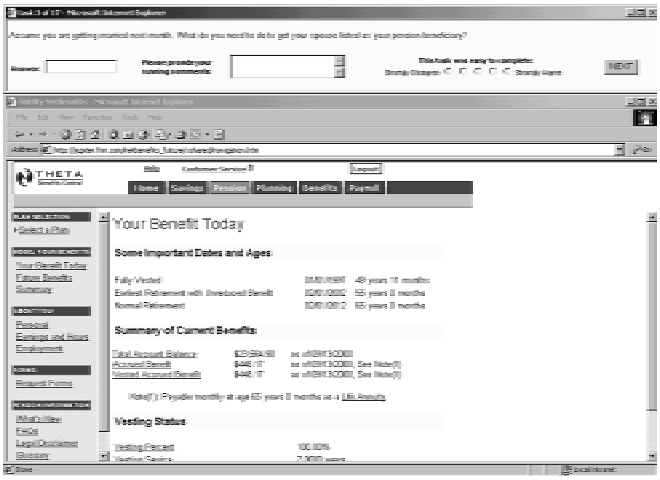

Figure 1. Screen shot showing the configuration of browser windows used in the remote usability tests.

Once they have found the answer, they type it into an "answer" field in the task window, provide any additional comments, and typically also are asked to rate the difficulty of completing the task. Clicking a "Next" button displays the next task. All of their entries, plus the elapsed time from display of the task until clicking the "Next" button, are automatically recorded in a file on the Web server hosting the study.

The information that can be collected using this technique is limited. Since the two browser windows are basically independent of each other, it is not possible to detect what pages the user visits in the main browser, or any interactions with those pages. Our information is limited to what the users report to us in the small task window, plus the elapsed time.

After going through the tasks in a remote test, the user is typically then asked to provide a subjective evaluation of the site or application using a set of rating scales and open-ended questions

OVERVIEW OF THE STUDIES

The overall goal of our studies was to evaluate the effectiveness of this remote usability testing technique, particularly in comparison to more traditional lab-based testing. Specifically, we wanted to determine whether the results from lab and remote tests of the same Web sites would yield similar results. Further, we wanted to assess the strengths and weaknesses of both techniques.

During the first half of 2001, we conducted both lab and remote usability tests of two prototype Web sites. For both sites, the lab and remote tests were conducted by different teams.

EXPERIMENT 1

The first study involved a prototype of a Web site for providing the employees of a company with access to information about their own benefits, including retirement savings information, pension information, medical and dental coverage, payroll deductions and direct deposit, and financial planning. It was a combination of an informational site and a transactional site where, for example, the users could change their payroll deductions.

Seventeen tasks were developed for this usability test. Some example tasks included the following:

- In your 401(k) plan, how much money do you have in the <fund name deleted> fund?
- If you were to stop working today, what monthly payment would you get from your pension plan if you retire at age 65?
- Have $200 automatically deposited each month from your paycheck into your savings account.

In both the lab and remote tests, the users were given a test ID for logging in to the system, so they all saw the same data for a hypothetical employee. Most of the tasks, such as the first two above, had definitive answers. But some tasks, such as the last one above, involved a simulated

transaction. In these cases, the users in the remote test were asked to report the title of the page they reached after completion of the task, so that we could determine whether they had actually completed it.

After the tasks, for both types of tests, each user was asked to provide subjective feedback about the Web site. In the lab tests this survey was presented on paper, while in the remote tests it was presented online. The survey consisted of the following nine statements to which the users rated their level of agreement:

1. This Web site is visually appealing.
2. It was easy to get around the site (moving from one page to another).
3. The information contained in this site is organized in a logical way.
4. Individual pages are well formatted.
5. Terminology used on the pages is clear.
6. The content of the Web site met my expectations.
7. I would be likely to access this Web site in the future.
8. I was able to complete my tasks in a reasonable amount of time.
9. Overall, the site is easy to use.

Each statement was accompanied by a scale of -3 to +3, ranging from "Strongly Disagree" to "Strongly Agree". On this scale, "0" provides an obvious neutral point. Each rating scale was also accompanied by an open-ended comments field. This is the same instrument for subjective assessments that we have been using in lab usability tests for several years.

Lab Test

We used our normal procedures for conducting the lab-based test. A total of eight users participated in the test individually, averaging about 1.5 hours per session over two days. All users were employees of our company recruited by random selection from the internal phone directory. Their incentive for participating was free movie tickets. Following a standard briefing about usability testing, the tasks were presented to the user on paper. The user was asked to read the task aloud and to think aloud while working through the tasks. The moderator for the test, who handled all interaction with the user, was in an adjacent room connected by one-way glass. The moderator also had two-way audio contact with the user, a monitor slaved to the user's monitor, and a remote-control video camera typically aimed at the user's face. The moderator was assisted by a data logger who used a laptop to record task start and finish times, whether the tasks were completed successfully, and any significant comments or issues. The entire session was videotaped, including a scan conversion of the user's screen, for subsequent review if needed.

The users were basically left to their own devices to figure out how to complete the tasks (all of which were possible to accomplish). If asked for assistance, the moderator generally just referred the user back to the task description. The criterion stated by the moderator for giving up on a task was whether the participants had reached the point where, if they were doing this for real, they would either give up or pick up the phone and call someone for assistance.

After all the sessions were completed, the moderator and data-logger analyzed the task completion data, task times, and subjective ratings. They also reviewed all of their notes from the test sessions, and, if needed, reviewed any of the videotapes. Based upon all of this information, they developed a master list of usability issues uncovered by the test.

Remote Test

For the remote test, participants were recruited via an email message sent to 100 randomly selected employees of our company. The incentive for participating was the same as in the lab—free movie tickets. A total of 38 people chose to participate, with 29 completing the entire study (a 24% drop-off rate). They were allowed to participate in the test whenever they wanted over the course of a week. However, they were instructed to complete the test in one sitting. The majority of the people chose to participate within 24 hours of receiving the email message.

The email message contained a link to the "Welcome" page of the study, which explained the basic purpose of the study and set some expectations (e.g., that it should take about 45 minutes, which turned out to be a conservative estimate). It also explained the mechanics of the test, encouraged the users to freely enter comments in the task window, and explained that the tasks were being timed. After they entered their corporate ID number (mainly for data correlation purposes) and clicked a "Go" button, the two windows of the remote test were opened, the first task was presented, and the logon screen of the prototype was displayed.

All user inputs in the task window were automatically recorded, as were the times to complete each task. (We also asked the users to rate the difficulty of each task, but since we did not capture these ratings in the lab, they will not be discussed.) Results from the subjective survey at the end of the tasks were also automatically recorded.

After all participants had finished, a different team from the one that had conducted the lab test analyzed the data. They had no knowledge of the results from the lab test, which was completed before the remote test. A key part of the analysis was determining whether the answers that the users had given for each task were correct, indicating successful completion of the task. Typographical errors were allowed as long as it was obvious that the user had found the right answer. Comments that the users gave for each task, which were surprisingly extensive, were also studied for any indications of usability issues. In analyzing the time data, individual task completion times under 5 sec (indicating the user did not seriously attempt the task) or over 1,000 sec (indicating the user was probably interrupted)

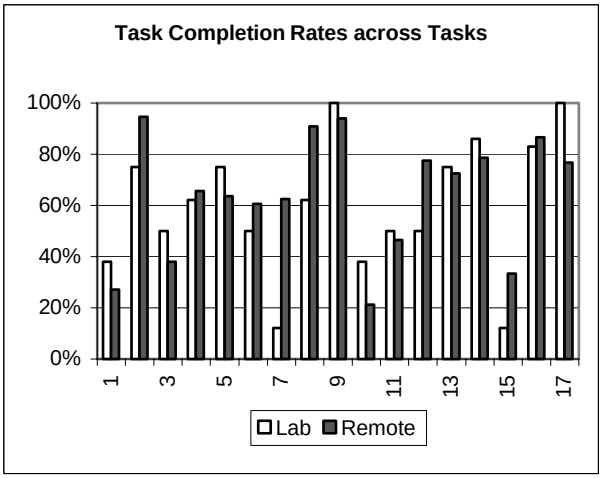

Figure 2. Experiment 1: Comparison of percentages of users who successfully completed each of the 17 tasks for both the lab and remote tests. Averages: Lab=60%, Remote=64%, *r* =0.70.

were discarded. There were only a few of these for each task. Based on all of this information, the team developed a master list of the usability issues that were uncovered.

Results

Four main types of data will be presented for both tests: successful task completion rates, task completion times, subjective ratings, and usability issues identified.

Task Completion Data

Figure 2 shows the percentage of users who successfully completed each of the 17 tasks, for both the lab and remote tests. Overall, the lab users completed 60% of the tasks while the remote users completed 64%. The differences were not statistically significant (t-test, p=0.39). The correlation coefficient between the two sets of completion rates was 0.70. One common use of task completion data in a usability test is to identify those tasks that the users had the most difficulty with, as an aid to identifying usability issues. The data from both tests pointed to tasks 1, 3, 10, 11, and 15 as having been particularly difficult (based on a 50% or lower completion rate). In addition, the lab test also indicated tasks 6, 7, and 12 as being difficult.

Task Time Data

Figure 3 shows the average time spent on each task for both the lab and remote tests. This includes tasks successfully completed as well as not. Overall, the lab users spent an average of 147 sec per task while the remote users spent 164 sec. The differences were not statistically significant (p=0.30). The correlation coefficient was 0.65. As with the task completion data, task times can be used to help identify potential usability issues, with unusually long task times perhaps indicating problem areas. The data from both tests pointed to tasks 1, 4, 6, 10, and 13 as problematic (based on an average task time of 175 sec or more). In addition, the remote test also indicated tasks 3 and 11 as problems, while the lab test also indicated task 7.

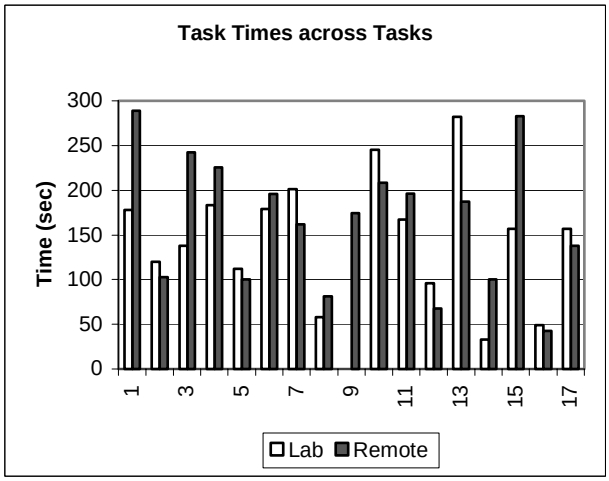

Figure 3. Experiment 1: Comparison of average times spent on each of the 17 tasks for both the lab and remote tests. Averages: Lab=147 sec, Remote=164 sec, *r*=0.65. (Time data missing for task 9 in the lab.)

Subjective Ratings

Figure 4 shows the average subjective ratings given by the users on each of the nine rating scales (visual appeal, ease of navigation, etc.) for both the lab and remote tests. The overall average for the lab users was 1.6 (on a scale of -3 to +3) while the overall average for the remote users was 0.7. For all but one question (#5), the remote users gave lower (worse) ratings than the lab users. The correlation coefficient was 0.49. The data from both tests pointed to questions 5 and 8 as being problem areas (based on an average rating less than 1.0), while the remote test also pointed to most other questions.

Usability Issues

The identification of issues, or potential usability problems, is always the most difficult part of any usability test. It is admittedly a somewhat subjective process, and influenced to a significant extent by the skill and experience of the

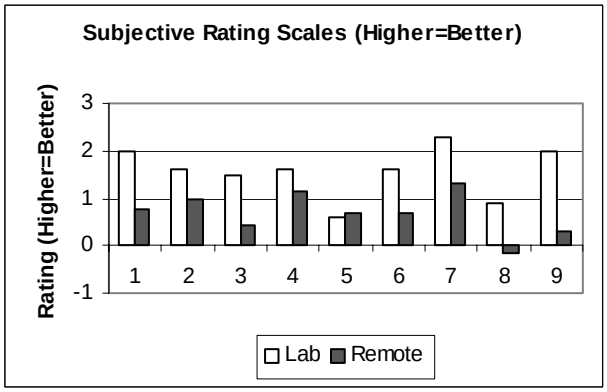

Figure 4. Experiment 1: Average subjective ratings given on nine rating scales (visual appeal, ease of navigation, etc) for both lab and remote tests. Scale: -3 to +3, where higher ratings are better. Averages: Lab=1.6, Remote=0.7, *r*=0.49.

individuals identifying the issues. In both the lab and remote tests, the individuals identifying the issues were experienced usability professionals. In both cases, they were careful to limit themselves to issues that could be directly traced to some observed behavior on the part of the users, as reflected by the data collected, including user comments. The process of deriving issues was perhaps more difficult for the remote test simply because our team had less experience with that type of test.

For the lab test, a total of 26 usability issues were identified. For the remote test, a total of 16 usability issues were identified. The two lists were compared to determine how much overlap there was between them. Eleven issues appeared in basically the same form on both lists. There was general agreement that the two most serious problems (confusion over the contents of the tabs defining the main content areas and general problems with terminology) were clearly identified by both tests. However, the lab test identified 15 issues not found in the remote test, and the remote test identified 5 issues not found in the lab test.

An example of an issue identified only by the lab test was most users initially failing to see a bar across the top of a table of numbers giving totals for the columns of the table. It was apparent from observing their behavior during the session that they initially overlooked this bar, although most eventually found it. From the remote test, the primary data indicated simply that the users generally did find the information in that bar; none of the users commented that they had failed to see it initially.

On the other hand, an example of an issue identified only by the remote test was a problem with fonts being too small and difficult to read in certain areas. An important point to realize is that the users in the remote test used their normal system configuration, including whatever screen resolution they normally use. It turned out that all used either 1024 x 768 pixels or 1280 x 1024. (In the remote test, certain characteristics of the user's system configuration are automatically captured.) In the lab test, the screen resolution was set to 800 x 600. Consequently, the lower resolution in the lab caused the fonts to appear larger, thus failing to find that problem. One could also postulate that the larger pool of users in the remote test was more likely to include some older people who may be experiencing some vision problems.

Discussion

Overall, the task completion data and the task time data from the two usability tests were surprisingly similar. Users generally had the most difficulty with the same tasks in both the lab and remote tests. This points to a consistency in the way we are capturing the user experience across both types of tests. Another pleasant surprise was the richness of the comments that users typed in the remote test, which helped to take the place of the direct observation available in the lab.

The subjective ratings, however, failed to show very much consistency, with the remote test participants generally giving more negative ratings. We have come up with two possible explanations for the difference. The first is simply that it is an artifact of the variability of the subjective ratings and the difference in the sample sizes (8 for the lab vs. 29 for remote). To test this hypothesis, we analyzed randomly selected sub-samples of size 8 from the remote data to see how many of them were statistically indistinguishable from the lab data. We found that about 25% of them were similar to the lab data. The second hypothesis is that users in the remote test may feel more anonymous, and thus more willing to be critical, since they are not physically in the same place as the people conducting the test. (Users in this remote test were not actually anonymous, although they could have been.) In addition, there is no social dynamic between the test user and the moderator which might cause the user to give more positive ratings. In spite of being told otherwise, some lab participants appear not to distinguish between the moderator and the developers of the site.

The comparison of the usability issues uncovered by both tests was reasonably encouraging, since the issues that we judged to be most significant were clearly identified in both tests. The issues that were uniquely captured by either the lab or remote test appeared to reflect the specific strengths or weaknesses of the technique.

EXPERIMENT 2

The purpose of the second set of usability tests was to validate some of the findings from the first set using a different Web site and different tasks, and to help clarify the reasons for the differences in the subjective ratings. In addition, we hoped to learn more about the strengths and weaknesses of each technique.

The second experiment involved a prototype version of a site for providing general financial information, such as stock quotes, company news, research, and information about investment strategies. It was an informational site only, with no transactional aspects or logon required.

Thirteen tasks were developed for this usability test, including the following:

- What is Step 2 of the guided tour of this Web site?
- You want to find more information on a fiber optics company named <company name deleted>. Which of their customers account for 85% of their revenue?
- What is the 52-week high price for a share of <company name deleted>?

Unlike Experiment 1, we designed all the tasks to have reasonably definitive answers, rather than relying on the user to report the name of the page reached upon completion of the task. This was primarily due to the fact that these tasks were designed from the beginning to be used in both the lab test and the remote test. In the first experiment, the remote test was added after the lab test was complete.

After completing the tasks, users in both the lab and remote tests were asked to provide subjective feedback about the site, using the same nine rating scales as in the first experiment.

Lab Test

A total of eight users participated in the test individually, averaging just over one hour per session. The sessions were conducted over two days. All users were employees of our company. They were recruited from a list of 403 employees who had attended an internal seminar on Web design. All other features of the lab test procedure were the same as in Experiment 1.

Remote Test

The same list of 403 employees who had attended the Web design seminar was used for recruiting the participants for the remote test, minus the 8 people who had participated in the lab test (which had been completed). A total of 108 people chose to participate in the remote test, with 88 of them completing the entire study. This drop-off rate of 18% is slightly lower than the 24% drop-off rate from Experiment 1, perhaps because this test was a little shorter. The participants were told that the study should take about 45 minutes, which appears to have been reasonably accurate.

All other features of the remote test procedure were the same as in Experiment 1, except that in this test we added a "Pause" button to each task in the small task window. This allowed the user to "stop the clock" on a task if they were interrupted or just wanted to take a break. This resulted in a dialog box centered on the screen with instructions to click "OK" when they were ready to continue with the task.

Results

As in Experiment 1, four types of data were analyzed for both tests: successful task completion rates, task completion times, subjective ratings, and usability issues identified.

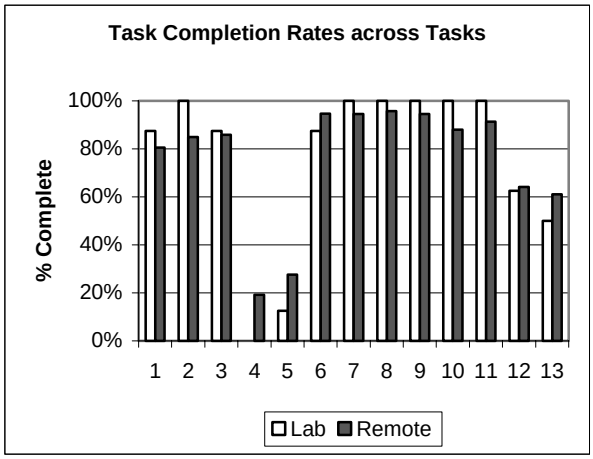

Figure 5. Experiment 2: Comparison of percentages of users who successfully completed each of the 13 tasks for both the lab and remote tests. Averages: Lab=76%, Remote=76%, *r*=0.98. (No lab users completed task 4.)

Task Completion Data

Figure 5 shows the percentage of users who successfully completed each of the 13 tasks, for both the lab and remote tests. The average task completion rate was 76% for both the lab and remote tests. The correlation coefficient was 0.98. Obviously, the two sets of task completion rates tracked closely with each other. From a practical standpoint, both tests showed that tasks 4 and 5 were the most difficult, followed by tasks 12 and 13, with the remaining tasks being relatively easy.

Task Time Data

Figure 6 shows the average time spent on each task for both the lab and remote tests. This includes tasks successfully completed as well as not. Overall, the lab users spent an average of 155 sec per task while the remote users spent 161 sec. The differences were not statistically significant (p=0.56). The correlation coefficient was 0.94. Similar to the task completion data, the two sets of task times tracked closely with each other. Both tests also showed that tasks 4 and 5 were the most problematic, followed by tasks 12 and 13.

Subjective Ratings

Figure 7 shows the average subjective ratings given by the users on each of the nine subjective rating scales for both the lab and remote tests. The overall average for the lab users was –0.01 (on a scale of –3 to +3) while the overall average for the remote users was 0.29. There was essentially no correlation between the ratings (r=0.04). In sharp contrast to the subjective ratings from Experiment 1, the remote users appeared to be slightly more positive than the lab users.

Usability Issues

The techniques for identifying usability issues were the same as those used for Experiment 1, in which two different teams independently derived the usability issues from

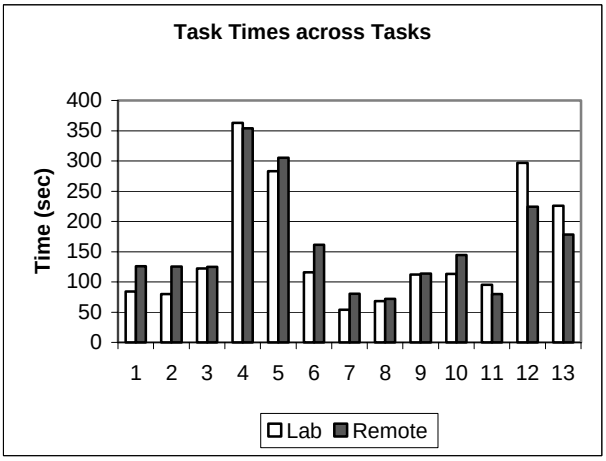

Figure 6. Experiment 2: Comparison of average times spent on each of the 13 tasks for both the lab and remote tests. Averages: Lab=155 sec, Remote=161 sec, *r*=0.94.

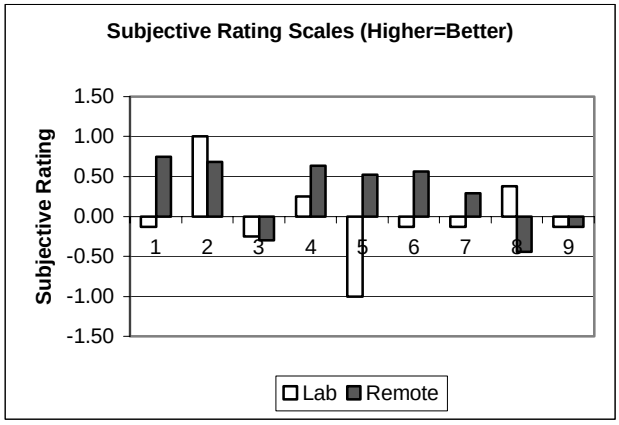

Figure 7. Experiment 2: Average subjective ratings given on nine rating scales (visual appeal, ease of navigation, etc) for both lab and remote tests. Scale: -3 to +3, where higher ratings are better. Averages: Lab=-0.01, Remote=0.29, *r*=0.04.

each of the tests. For the lab test, 9 usability issues were identified. For the remote test, 17 usability issues were identified. In comparing the two lists, we found that 7 very similar issues were on both lists. Thus, the lab test uniquely identified 2 usability issues while the remote test uniquely identified 10. Similar to Experiment 1, there was general agreement that the three most significant usability issues were clearly on both lists (overloaded home page, a general terminology problem, and a navigation problem caused by unclear categories on the home page).

Once again, at least some of the issues uniquely identified by each of the tests appeared to reflect certain important characteristics of the techniques. In the lab, an issue associated with excessive vertical scrolling on the home page was identified. As in Experiment 1, the lab test was run in 800x600 resolution, since that is the most dominant resolution on the Internet. In the remote test, only one user ran in that resolution; the vast majority ran in 1024x768 resolution, which required less scrolling on the home page. (The remote test was run on our company's Intranet, where 1024x768 is by far the dominant resolution.) Many of the issues identified uniquely by the remote test did not appear to have an origin in characteristics of the technique (e.g., titles of articles not matching links that led to them). They may have been found simply because of the far larger number of users (108 remote vs. 8 lab). A few of the unique issues once again appeared related to the technique. As in Experiment 1, small fonts were identified as being an issue in some areas of the site, probably due to the higher resolution most remote users were running in.

Discussion

Even more so than in Experiment 1, the task completion data and the task time data correlated extremely well. And, once again, the remote users provided very rich comments.

The lack of any correlation in the subjective ratings was somewhat surprising at first, as was the fact that they appeared more positive for the remote test than the lab test, which was the opposite of Experiment 1. It appears that the source of the apparent difference between the remote and lab ratings is simply the variability of the ratings and the much smaller sample size in the lab. In analyzing the data from randomly selected sub-samples of size 8 from the remote data, we found that 65% of them were statistically indistinguishable from the lab data. Coupled with the findings from Experiment 1, we believe this makes it clear that the subjective ratings from only 8 users in a typical lab test are simply not reliable.

Although it was reassuring that both tests uncovered the major usability problems, it was clear that the remote test uncovered more (17 vs. 9). We can think of at least two likely reasons for this. An obvious one is the much larger number of test users. One effect of the larger sample is greater diversity among the test users, thus increasing the likelihood that an issue specific to a particular class of users (e.g., elderly users) or a particular usage environment (e.g., certain screen resolutions) might be uncovered. A very different explanation could simply be variability in the interpretation of the data and identification of usability issues by the two teams. This explanation is consistent with the conclusions of Molich and his associates [10,11], who have found that different teams conducting usability tests of the same site yielded surprisingly different results. We believe that both reasons probably contributed to the differences we found in usability issues.

CONCLUSIONS

Some conclusions are clear from the findings of these two sets of usability tests:

- The behavior of test users is strikingly similar in lab and remote usability tests, as evidenced by their task completion rates and task times. Our users encountered similar problems in completing the tasks and devoted similar amounts of time to them. This is reassuring, and indicates that the different environments do not lead to different kinds of behavior.
- The users in these remote tests almost universally provided very rich typed comments. In most cases, this compared favorably to the kind of data we could get in the lab via direct observation. In our approach to remote data collection we had decided to limit ourselves to normal browser capabilities. If the situation allows for an instrumented browser (e.g., for capturing clickstreams), it may be possible to come even closer to the rich set of data available in the lab.
- The larger number of users that can be practically run in this type of remote test definitely offers some advantages. The most obvious advantage is that the users are more diverse, thus increasing the likelihood of uncovering problems unique to specific types of users. Since the users perform the tasks at their own computers, another advantage is that problems unique to specific environments are also more likely to be captured.

- It seems clear that these remote tests provided much more reliable subjective assessments of the Web sites, due to the larger numbers of users involved. Because of the inherent variability of subjective ratings, the small number of users typical of lab tests is simply not sufficient to draw any reliable conclusions about subjective reactions.
- The lab (or at least direct observation) still appears to be uniquely suited for capturing certain kinds of usability issues. We saw evidence of certain kinds of user behaviors in the lab (e.g., excessive scrolling, failure to see certain elements on the screen at first) that were less likely to be captured in the remote tests.
- Both types of tests appear to capture the most significant usability issues for a Web site. There were two or three major usability issues with each of these sites which both techniques very clearly captured. Thus, if you only care about capturing the major problems, either technique could be used. However, we believe that the most thorough assessment of the usability of a Web site would involve both lab and remote tests.

REFERENCES

1. Rubin, J. *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests.* John Wiley & Sons, 1994.
2. Dumas, J. S., and Redish, J. C. *A Practical Guide to Usability Testing, Revised Edition*. Intellect, 1999.
3. Nielsen, J, and Landauer, T. K. A mathematical model of the finding of usability problems. *Proceedings of ACM INTERCHI'93 Conference* (Amsterdam, The Netherlands, 24-29 April 1993), pp. 206-213.
4. Virzi, R. Refining the test phase of usability evaluation: how many subjects is enough? *Human Factors, 34*, p. 457.
5. Nielsen, J. Why you only need to test with 5 users. *Alertbox* (http://www.alertbox.com), March 19, 2000.
6. Hammontree, M., Weiler, P., and Nayak, N. Remote usability testing. *Interactions*, July 1994, pp. 21-25.
7. Castillo, J. C., Hartson, H. R., and Hix, D. Remote usability evaluation: Can users report their own critical incidents? *Proceedings of CHI'98 Conference*, pp. 253-254.
8. Hartson, H. R., Castillo, J. C., Kelso, J., and Neale, W. C. Remote evaluation: The network as an extension of the laboratory. *Proceedings of CHI'96 Conference*, pp. 228-235.
9. Vividence Corporation, 1850 Gateway Drive, San Mateo, CA. http://www.vividence.com.
10. Molich, R., Bevan, N., Butler, S., Curson, I., Kindlund, E., Kirakowski, J., and Miller, D. Comparative evaluation of usability tests. *Proceedings of UPA'98 Conference* (Usability Professionals Association), pp. 189-200.
11. Molich, R., Thomsen, A. D., Karyukina, B., Schmidt, L., Ede, M., van Oel, W., and Arcuri, M. Comparative evaluation of usability tests. *Proceedings of CHI'99 Conference, Extended Abstracts*. Also see http://www.dialogdesign.dk/cue.html.