Comparing Synchronous Remote and Local Usability Studies for an Expert Interface

Morgan Ames[+], A.J. Bernheim Brush[*], and Janet Davis[*]

EECS Department[+]
University of California, Berkeley
morganya@cs.berkeley.edu

CS&E Department[*]
University of Washington, Seattle
{ajb, jlnd}@cs.washington.edu

ABSTRACT

Synchronous remote usability studies can be a convenient and cost-effective alternative to conventional local usability studies. Although they are common in the field, there has been little research comparing synchronous remote usability studies with local studies. In our comparison of remote and local studies for an expert interface, the primary differences were in the participant's and facilitator's qualitative experience. The number of usability issues found, their type, and severity were very similar. While more research is needed, our experience suggests that evaluators of expert interfaces will have comparable success identifying usability issues with either remote or local studies.

Keywords

Remote evaluation, usability research, usability testing and evaluation, user studies.

INTRODUCTION

Usability studies are an important part of the software development process. Many usability studies are conducted in a lab setting in which a user completes a set of tasks as a usability specialist looks over the user's shoulder or watches from an adjacent room. However, the users of some applications are in remote or distributed locations, and the travel expenses for in-person evaluation with the remote users of these systems can be prohibitive. Moreover, if the software is for specialists or the culture of the target users differs significantly from the local culture, it may not feasible to recruit local "representative users" to participate in place of the target users. For instance, UrbanSim [19], the land use and transportation simulator that we evaluated in this study, has a distributed user base including urban planners in Washington, Oregon, Utah, Texas, and Hawaii. In these cases, synchronous remote usability studies, where the study facilitator and participant are not co-located but interact over a computer or telephone network, can be more cost-effective than local studies.

Remote usability studies can also potentially provide data from large numbers of participants [14]. In addition, they allow participants to remain in their normal setting, yielding a more realistic test of the interface. However, some warn that a remote study facilitator may miss contextual information and subtle cues such as facial expressions, making the results of remote studies more difficult to interpret [8, 11, 15].

Hartson *et al.* divide remote usability studies into synchronous and asynchronous studies [10]. Synchronous remote usability studies simulate in-person local studies, in which a participant and a study facilitator are directly communicating in real-time. In asynchronous studies, participants are both spatially and temporally remote from the facilitator. Participants provide comments or have their actions logged, but have little or no contact with the researchers running the study.

Asynchronous remote usability methods, such as critical-incident reporting [e.g., 10] and automated data collection [e.g., 11], are well researched, but the data, such as mouse clicks and web page viewing times, are often low-level and difficult to put into context [11, 17, 18]. In contrast, synchronous remote usability studies provide results that are easier to interpret. They are often used in the field, but have not been as well investigated.

In our comparison, we focus on remote studies as they are often done in the field today, with commonly-used software in the participant's workplace. Furthermore, the interface evaluated in this study is a new graphical interface to UrbanSim, a tool that is part of or closely related to the participants' work process. We will discuss:

Usability Issues Found: How do the types, number, and severities of issues found differ between remote and local studies?

Participant's Experience: How does the participant's experience differ between remote and local studies? Do participants prefer one type of study?

Facilitator's Experience: How does the facilitator's experience differ? How do study time and effort differ between remote and local studies?

The results of our comparison are a valuable first step toward answering these questions and building an understanding of the tradeoffs between remote and local usability studies, especially for expert interfaces.

RELATED WORK

Remote usability encompasses a wide range of practices, including automated data collection and mining [e.g., 11], remote questionnaires, asynchronous user-initiated critical-incident reporting [e.g., 10], synchronous telephone conferencing [e.g., 15], and synchronous video conferencing, which simulates a local study as closely as possible. In this section we focus on research on synchronous usability studies.

Synchronous remote usability studies use a variety of media, including telephone, recorded audio or video, networked audio or video, online text chat, and screen sharing [14]. Common tools for synchronous remote usability studies include VNC and other screen sharing programs; WebEx, NetMeeting and other conferencing tools; and WebCat, a web-based category testing tool [8, 14, 16]. Synchronous usability studies have been used for evaluating a variety of interfaces, from low-fidelity HTML or Flash prototypes to finished interfaces [14].

Although several works offer best practices for synchronous remote studies [5, 8, 12, 14], we are aware of only one other experimental comparison of synchronous remote and local studies. In 1996, Hartson *et al.* found no significant difference, in terms of number of usability problems found or participant experience, between synchronous remote usability studies and local (next-room) studies of a commercial web site with eight participants [10].

Our comparison differs from that of Hartson *et al.* in several ways. First, we evaluate synchronous remote usability studies as they are often done in the field today, with commonly-used software rather than special hardware such as high-frame-rate scan converters. This allows remote participants to work from their desks rather than a dedicated satellite usability lab. Second, 8 of our 20 participants participated in both a remote and local usability study, allowing a within-groups comparison of their experiences. Finally, we evaluated an interface intended for experts, rather than a general audience.

METHOD

The 20 participants in our comparison performed tasks using the UrbanSim interface. Each usability study took between 1 and 1.5 hours. To control for facilitator variation, the same facilitator performed all the studies.

We tested two study conditions in our comparison:

Local: The participant came to our usability lab and completed tasks related to the UrbanSim interface. The study facilitator sat beside the participant taking notes, and an observer seated in the room also took notes. The participant and facilitator interacted using the Boren-Ramey think-aloud protocol [1]. The participant's voice and computer screen were recorded.

Remote: Before their study session, remote participants downloaded (but did not install) Eclipse [6], to avoid long downloads during the study. They also installed supporting software, such as Java, if necessary. The study facilitator called the participant at work at the specified time of the study. The facilitator then helped the participant install Glance, a VNC-based screen-sharing program [7]. Once the tasks began, the participant and facilitator interacted over the phone using the think-aloud protocol, while the facilitator and an observer took notes. The participant's voice and computer screen were recorded.

Setup

To evaluate the differences between the remote and local conditions, we conducted 12 remote and 8 local usability studies. We found it much easier to recruit remote participants, and chose to schedule as many remote studies as was feasible. In contrast, it was a challenge to find 8 local participants.

Participants worked with the graphical interface for UrbanSim, developed as a plug-in to the Eclipse platform. Participants installed Eclipse and UrbanSim and then created an UrbanSim project representing a small city. Next they ran a simulation of the city's development, interpreted the results, and turned on additional logging for the simulation. After participants completed the tasks, the facilitator elicited further comments and reflection from participants, supported by task descriptions and, in the local condition, the screen recording. Before and after the tasks, the facilitator read from a script to ensure consistency between studies.

To allow a within-groups comparison of participants' experiences in the two conditions, 8 participants returned for a second usability study a day or two after their first study. Four remote participants came to our lab for a local study, and we called 4 local participants for a remote study. Thus, 8 of the 20 participants experienced both conditions, for a total of 28 studies. In the second study, participants installed Eclipse and UrbanSim again, and then completed tasks comparable in difficulty to the first study.

In addition to the 28 usability studies completed, 3 other remote studies were canceled due to technical difficulties.

Participants

Our participants were professional urban planners and urban planning students. Twelve participants were from Seattle, while the other 8 participants were from Indianapolis, Salt Lake City, Boston, and elsewhere in the

United States. Five of the participants had used UrbanSim previously, but none had ever seen the graphical interface under evaluation. Participants were compensated with a $15 online gift certificate for one session or a $20 gift certificate for two sessions.

Protocol

For all usability studies, we used the Boren-Ramey think-aloud protocol [1]. In this protocol, which is based on speech communication theory, interaction between facilitator and participant is viewed as a conversation. While the participant completes tasks and thinks aloud, the facilitator gives the participant various conversational cues to indicate attentiveness. We scripted these cues and other allowed responses to various situations, such as direct questions, significant frustration, and system failures, to ensure consistency across all participants.

Data Collection

Participants completed a demographic survey before starting any tasks, and another survey after the study was finished. Those who participated in a second study, as part of our within-groups comparison, completed an additional post-study survey as well as a survey asking them to compare the two studies.

During the tasks, the facilitator and a single observer took notes, and the participant's voice and computer screen were recorded using Camtasia Experience Recorder [3]. All reminders, encouragements, interventions, and suspensions were recorded.

Pilot Studies

To test our experimental setup and refine our procedures, we conducted 18 pilot studies with 9 computer science students and researchers, each of whom participated in one local and one simulated remote study. We simulated the remote condition by setting up a satellite usability lab several rooms away from the lab where the study facilitator and note-takers ran the study. We felt that this simulation would give us the opportunity to test the mechanics of remote studies while giving us control over the computer configuration and the ability to troubleshoot in-person, if necessary.

The experiment was counter-balanced, with 4 participants beginning with the local study, and 5 beginning with the simulated remote study. The results of the pilot studies indicated that screen sharing and communicating using the phone was feasible, and did not appear to make the simulated remote studies less effective than the local studies. The number of usability problems found, the participant experience, and the facilitator experience were similar to the results of our later studies.

USABILITY PROBLEMS FOUND

Our 20 participants experienced a total of 243 usability issues, from which we identified 94 unique issues. This does not include any issues found in the second study sessions completed by 8 of the participants, as there were a different set of tasks for those sessions.

Table 1 shows the issues broken down into five categories: (1) installation, (2) the entire interface, (3) a single dialog or element, (4) documentation, and (5) other software (such as WinZip). To determine the category for an issue, we each independently coded the issues and then resolved differences through discussion.

Initially, we thought it might be harder to observe issues in the remote condition since we only had screen sharing and a phone connection with participants. However, as Table 1 shows, the median number of issues found in the two conditions are very similar, both overall and broken down by categories. Mann-Whitney U tests showed none of the medians are significantly different (all $p > 0.1$) between the two conditions. While the median number of issues found did not differ significantly, some installation issues related to proxy servers and firewalls that were found in remote studies could not have been found in local studies.

We each independently rated the severity of the issues using Nielsen’s severity rating scale [13], and then averaged the three sets of severity ratings. A Mann-Whitney U test showed there is no significant difference between the median severity of issues found by

Issue Categories	Unique Issues	Total Issues Experienced	Median number of issues experienced (Avg., SD)		
			Remote, N=12	Local, N=8	Significance
1. Installation	16 (17%)	33 (13.5%)	1.5 (1.8, 1.2)	1.5 (1.5, 1.5)	p = 0.678
2. Entire interface	33 (35%)	89 (37%)	4 (4.1, 1.8)	4.5 (5, 2.3)	p = 0.427
3. Single dialog or element	31 (33%)	88 (36%)	4 (4.3, 1.4)	4 (4.6, 1.8)	p = 0.851
4. Documentation	6 (6%)	22 (9%)	1 (1.1, 0.67)	1 (1.1, 1.1)	p = 0.970
5. Other software	8 (9%)	11 (4.5%)	1 (0.75, 0.75)	0 (0.3, 0.46)	p = 0.181
Total	94	243	12 (11.9, 2.8)	14 (12.5, 2.8)	p = 0.571

Table 1. Issues found by study participants. Mann-Whitney U tests show no significant differences in the median number of issues experienced in the remote and local conditions for any issue category.

Question	About equal	Remote	Local
Q1. In which study were you more comfortable talking to the evaluator? (N=8)	6 (75%)	0	2 (25%)
Q2. In which study was it easier to remember to “think aloud”? (N=7)	5 (71%)	1 (14.3%)	1 (14.3%)
Q3. In which test was it easier to remember and discuss what you were thinking during each task? (N=8)	7 (87.5%)	0	1 (12.5%)
Q4. In which study was it easier to concentrate on the tasks? (N=8)	4 (50%)	1 (12.5%)	3 (37.5%)
Q5. In which study did you feel like you have contributed something important to the redesign of the UrbanSim interface? (N=8)	7 (87.5%)	0	1 (12.5%)
Q6. Which study was more convenient for you? (N=8)	1 (12.5%)	7 (87.5%)	0
Q7. Which kind of study would you rather participate in if you were asked to do a usability study in the future, either for UrbanSim or for other projects? (N=8)	4 (50%)	4 (50%)	0

Table 2. Selected questions from the comparison survey given to the 8 participants in both the local and remote conditions.

participants in the two conditions (Z = -0.046, p = 0.970).

In addition to usability issues, we also identified issues participants experienced during the studies related to questions or confusion about the assigned tasks, technical difficulties such as network outages, and issues with software setup. Mann-Whitney U tests showed there are no significant differences between the median number of these types of issues found by participants in the two conditions.

PARTICIPANT’S EXPERIENCE

We were very interested in understanding participants’ qualitative experiences of local and remote usability studies. Table 2 summarizes the results of the comparison survey answered by the 8 participants who experienced both conditions.

We had initially hypothesized that participants would be more comfortable talking to the facilitator and would find it easier to think aloud and concentrate on tasks in the local condition. However, 75% of participants thought that their comfort level talking to the facilitator was about equal in both conditions (Q1), and 71% felt that it was equally easy to remember to “think aloud” in both conditions (Q2). All but one participant thought it was equally easy in both conditions to remember and discuss what they were thinking during the tasks (Q3). One difference between conditions was that three of the participants (37.5%) felt it was easier to concentrate on the tasks in the local condition (Q4).

In both conditions, participants felt that their contributions to the redesign of the UrbanSim interface were about equal (Q5). The majority of participants felt that the remote condition was more convenient (Q6) and half would prefer to be involved in remote studies over local studies in the future, while none preferred local over remote (Q7).

We were interested in what types of monitoring participants would accept during a remote study. In the surveys given after the remote studies, we asked participants how acceptable different types of monitoring would be in a remote study. Table 3 shows the responses from the sixteen participants in the 12 remote studies and the 4 additional remote studies that we conducted for the within-groups comparison. As Table 3 shows, all or most participants were willing to accept further monitoring of the types that were used in this study, as well as recording of input events. However, many were unwilling to accept eye tracking or videotaping of the hands or face. This suggests that the use of eye tracking or videotaping would limit the pool of available participants for remote usability studies.

Type of monitoring	Acceptable (N=16)
Recording the telephone call	16 (100%)
Remotely viewing your computer screen (as in this study)	16 (100%)
Recording your computer screen	15 (94%)
Recording input events (mouse movements, keystrokes)	16 (100%)
Eye tracking	7 (44%)
Recording your screen and hands with a video camera	9 (56%)
Recording your facial expressions with a video camera	7 (44%)

Table 3. Remote participants were asked, “If you participated in a remote study in the future, what kinds of monitoring during the study would you accept?”

FACILITATOR’S EXPERIENCE

In this section we compare our experience preparing for and facilitating remote and local studies.

Before the Studies: It took more effort for us to prepare for the remote usability studies. This included ensuring each participant’s computer met our minimum configuration requirements and setting up a password-protected website with study materials. We found

Study Segment	**Remote, N=12** min. (Avg., SD)	**Local, N=8** min. (Avg., SD)	**Sig.**
Setup	15 (16, 2)	13 (13, 2)	p = 0.02*
Tasks	45 (42, 9)	42 (47, 14)	p = 0.678
Discussion	7 (7, 3)	13 (13, 6)	p = 0.005*
Suspensions	1 (3, 5)	1 (2, 3)	p = 1.0
Wrap-up	9 (10, 3)	4 (4, 1)	p ≤ 0.001*
Total	81 (77, 12)	73 (80, 20)	p = 0.678

Table 4. Median length of study segments in minutes. *Medians are significantly different with p < 0.05 based on a Mann-Whitney U Test.

password protection crucial for preventing people from looking over materials before the study, as several people mentioned they had tried to look ahead of time.

We found that recruiting remote participants was much easier than recruiting local participants. One email to the urbansim-users mailing list resulted in many more responses than we needed, while multiple emails and requests were necessary to find enough local participants. The ease of finding remote participants proved useful when three remote studies were canceled due to technical difficulties such as firewall restrictions, slow connections, and server failures.

During the Studies: We felt that it was just as easy to observe issues in the remote condition as in the local condition, once the screen sharing connection was established. Furthermore, the participant's tone of voice was enough to let us sense frustration.

From our experience helping participants before we could see their screen, we consider screen sharing to be a crucial element in successful remote usability studies. We found that without the visual cues afforded by screening sharing, it very difficult to understand what the participant was doing, and to assist if problems arose.

Coping with problems that required a suspension of the study, such as network failures and software crashes, was much more challenging in the remote studies. We had to guide the participant through diagnosing and fixing the problem, rather than asking the participant to take a break while we resolved the problem ourselves.

Finally, we experienced 9 short external interruptions (such as email arrival notifications) in the remote condition only. These interruptions had little impact on our studies, but could be significant for time-sensitive tasks.

Study Length: The median study length in both conditions was not significantly different based on a Mann-Whitney U test. However, as Table 4 shows, remote studies required slightly more time for setup and wrap-up (as expected), while local participants spent longer discussing their experience. Mann-Whitney U tests showed the median length of time spent on setup, wrap-up and discussion was significantly different at the $p < 0.05$ level. However, none of the medians differ by more than six minutes.

Although remote studies took a little bit longer than local studies, the facilitator found the remote studies to be far less stressful. Although both local and remote studies require the facilitator to maintain a neutral conversational style, the facilitator and observers in remote studies can move around during the study, which gives the facilitator and observers the ability to stretch and to quietly consult one another without distracting the participant.

CONCLUDING REMARKS

In our comparison, we found primarily qualitative differences between the remote and local study conditions. We saw no significant differences in terms of the number of usability issues found, their types, or their severities, consistent with the findings of Hartson *et al.* [10] in a different setting. However, half the eight participants who experienced both conditions would prefer to participate in remote studies in the future, and none would prefer local studies. As study facilitators, we needed to recruit more remote participants due to technical difficulties, but found this was not hard. We were also pleasantly surprised by how well we could recognize usability issues through screen sharing and the phone connection.

In choosing between local and remote usability studies, there is a tradeoff between control and realism. On one hand, our experience suggests that remote studies would not work well for relatively new software because the study facilitator has so little control of the remote user's environment. We had problems with network speed, firewalls, and web servers that did not teach us anything useful about the software being evaluated. On the other hand, in the remote cases we found important usability problems relating to supporting software and web proxy configurations that we never would have found in our lab.

FUTURE WORK

Our results suggest that evaluators of expert interfaces can choose to do remote or local studies and obtain comparable results. In the future, we plan to conduct primarily remote studies, allowing us to easily evaluate UrbanSim with geographically dispersed participants.

We are particularly interested in further exploring issues of comfort level and trust of the facilitator for participants in remote studies. In our comparison, we found that 25% of participants who experienced both conditions (2 of 8) felt more comfortable talking with the facilitator in the local condition. Since we recruited participants who had some connection to or knowledge of UrbanSim before the study, it would be helpful to understand whether a participant's comfort level in the remote condition is lower if they have a weaker interest in the software or are unfamiliar with it.

We are also interested in investigating the effectiveness of synchronous remote studies in which the participant does not speak the facilitator's language fluently. In our studies, all our participants were in the United States and spoke English fluently, and we had no trouble conversing. However, remote usability studies are cited as a cost-effective means to evaluate software with an international user base [4, 9], so such an investigation would be useful. Dray and Siegel [4] in particular discuss possible benefits and drawbacks of international remote usability studies, suggesting several interesting avenues for future research.

While our comparison is a valuable first step, we encourage other comparisons that evaluate different interfaces and other choices for configuring the remote and local conditions. Further experiments are critical for building a knowledge base of research to understand the tradeoffs between remote and local studies.

ACKNOWLEDGMENTS

We gratefully acknowledge the participants in our studies. We also thank Alan Borning, James Landay, Judy Ramey, and David Gibson for their assistance. This research has been funded in part by National Science Foundation grants EIA-0090832 and EIA-0121326, and the CRA-W Distributed Mentorship Project.

REFERENCES

[1] Boren, T., and Ramey, J. Thinking Aloud: Reconciling Theory and Practice. *IEEE Transactions on Professional Communication*, pages 261-278. September 2000.

[2] Brush, A.J., Ames, M., and Davis, J. A Comparison of Synchronous Remote and Local Usability Studies for an Expert Interface. *Proceedings of CHI 2004, ACM Conference on Human Factors in Computing Systems*, pages 1179-1183. April 2004.

[3] Camtasia. http://www.techsmith.com/products/studio/

[4] Dray, S. and Siegel, D. Remote possibilities? International usability testing at a distance. *Interactions*, pages 10-17. March/April 2004.

[5] Ebling, M.R., and John, B.E. On the Contributions of Different Empirical Data in Usability Testing. *Proceedings of DIS*, August 2000.

[6] Eclipse Project. http://www.eclipse.org

[7] Glance Networks. http://www.glance.net

[8] Gough, D. and Phillips, H. Remote Online Usability Testing: Why, How, and When to use it. http://www.boxesandarrows.com/archives/remote_online_usability_testing_why_how_and_when_to_use_it.php

[9] Hammontree, M., Weiler, P., and Nayak, N. Remote Usability Testing. *Interactions*, pages 21-25. July 1994.

[10] Hartson, H.R., Castillo, J.C., Kelso, J., and Neale, W.C. Remote Evaluation: The Network as an Extension of the Usability Laboratory. In *Proc. CHI 1996*, pages 228-235.

[11] Hilbert, D.M., Redmiles, D.F. Separating the Wheat from the Chaff in Internet-Mediated User Feedback. *ACM SIGGROUP Bulletin*, Volume 20, Issue 1, pages 35-40. April 1999.

[12] Jacques, R. and Savastano, H. Remote vs. Local Usability Evaluation of Web Sites. *Proceedings of IHM HCI 2001*.

[13] Nielsen, Severity Ratings for Usability Problems. http://www.useit.com/papers/heuristic/severityrating.html

[14] Olmsted, E. and Horst, D. Remote Usability Testing: Practices and Procedures. Workshop at Usability Professionals Association conference, June 2003.

[15] Ratner, J. Learning About the User Experience on the Web With the Phone Usability Method. *Human Factors and Web Development*, 2nd edition. October 2002.

[16] Scholtz, J. Adaptation of Traditional Usability Testing Methods for Remote Testing. In *Proceedings of Hawaii International Conference on System Sciences*. January 2001.

[17] Tamler, H. High-Tech vs. High-Touch: Some Limits of Automation in Diagnostic Usability Testing. *User Experience*, pages 18-22. Spring/Summer 2003.

[18] Tullis, T., Fleischman, S., McNulty, M., Cianchette, C., and Bergel, M. Empirical Comparison of Lab and Remote Usability Testing of Web Sites. In *Proceedings of Usability Professionals Association Conference*, July 2002.

[19] Urbansim Project. http://www.urbansim.org