A Comparison of Synchronous Remote and Local Usability Studies for an Expert Interface

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

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

EECS Department[+]
University of California, Berkeley
morganya@uclink.berkeley.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 of an expert interface, the primary differences were in the participant's and facilitator's qualitative experience. The remote and local studies agreed closely (with no significant differences) in terms of the number of usability issues found, their type, and their severity. While our comparison focuses on an expert interface and more work is needed to understand remote studies in general, our experience suggests that evaluators of expert interfaces will have comparable success identifying usability issues with either remote or local studies.

Categories & Subject Descriptors
H.5.2 [User Interfaces]: Evaluation/methodology, User-centered design, Theory and methods; H.1.2 [User/Machine Systems]: Human Factors

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 be feasible to recruit local "representative users" to participate in place of the target users. For instance, UrbanSim [11], 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 [10]. 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 [5,7,10].

While asynchronous remote usability methods, such as critical-incident reporting [e.g., 6] and automated data collection [e.g., 7], are well researched, synchronous remote studies – where a participant and study facilitator communicate directly in real-time – have not been as well investigated. In our comparison of remote and local studies we focus on:

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?

Though we cannot answer these questions definitively, the results of our comparison are a valuable step toward building an understanding of the tradeoffs between remote and local usability studies.

RELATED WORK
Although several works offer best practices for synchronous remote studies [2,5,8], we are aware of only one other experimental comparison of synchronous remote

CHI 2004, April 24–29, 2004, Vienna, Austria.
ACM 1-58113-703-6/04/0004.

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 [6].

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, eight of our twenty participants participated in both a remote and local study, allowing a within-groups comparison of their experiences. Finally, we evaluated an interface intended for experts, rather than a general audience.

STUDY METHOD

The 20 participants in our comparison performed tasks using the UrbanSim interface. Each 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 [3], 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 [4]. 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 studies and 8 local 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.

To allow a within-groups comparison of participants' experiences in the two conditions, 8 participants returned for a second 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 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 across 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.

USABILITY ISSUES 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 [9], 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 participants in the two conditions ($Z = -0.046$, $p = 0.970$).

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 eight participants that experienced both the local and remote conditions.

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 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 to remember and discuss what they were thinking in both conditions (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).

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 studies. This included setting up a password-protected website with study materials and ensuring each participant's computer met our minimum configuration requirements.

We found that recruiting remote participants was easier. 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.

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.

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

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 3 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.

CONCLUDING REMARKS

In our comparison, we found primarily qualitative differences between the remote and local study conditions. We saw no differences in terms of the number of usability issues found, their types, or their severities, consistent with the findings of Hartson *et al.* [6] in a different setting. However, half the 8 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 challenging. We were also pleasantly surprised by how well we could recognize usability issues through screen sharing and the phone connection.

Our experience suggests 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.

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 studies 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, and Judy Ramey for their assistance. This research has been funded in part by National Science Foundation grants EIA-0090832 and EIA-0121326.

REFERENCES

1. Boren, T., and Ramey, J. Thinking Aloud: Reconciling Theory and Practice. *IEEE Trans. on Professional Communication*, pg. 261-278. Sept. 2000.
2. Ebling, M.R., and John, B.E. On the Contributions of Different Empirical Data in Usability Testing. *Proceedings of DIS*, Aug. 2000.
3. Eclipse Project. www.eclipse.org
4. Glance Networks. www.glance.net
5. Gough, D. and Phillips, H. Remote Online Usability Testing: Why, How, and When to use it. www.boxesandarrows.com/archives/remote_online_usability_testing_why_how_and_when_to_use_it.php
6. Hartson, H.R., Castillo, J.C., Kelso, J., and Neale, W.C. Remote Evaluation: The Network as an Extension of the Usability Laboratory. *CHI 1996*, pg. 228-235.
7. Hilbert, D.M., Redmiles, D.F. Separating the Wheat from the Chaff in Internet-Mediated User Feedback. *ACM SIGGROUP Bulletin*, V. 20, Issue 1, pg. 35-40. April 1999.
8. Jacques, R. and Savastano, H. Remote vs. Local Usability Evaluation of Web Sites. *IHM HCI 2001*.
9. Nielsen, Severity Ratings for Usability Problems. www.useit.com/papers/heuristic/severityrating.html
10. Ratner, J. Learning About the User Experience on the Web With the Phone Usability Method. *Human Factors and Web Development,* 2nd edition. Oct. 2002.
11. Urbansim Project. www.urbansim.org