Human-Moderated Remote User Testing: Protocols and Applications

Asimina Vasalou, Brian David Ng, Peter Wiemer-Hastings and Lidia Oshlyansky

DePaul University, School of CTI
243 S. Wabash Avenue, Chicago, IL 60604
minav@luminainteractive.com

Abstract. Local user testing accrues high costs when solely relied on during the iterative cycle of user-centered design. A number of automatic evaluation methods have emerged with the aim of reducing costs incurred by repeated local tests, international user testing and post-deployment assessments. When considering costs, it is important not to overlook the benefits provided in human-moderated user testing. This paper focuses on the benefits of human moderation, provides test administration protocols for human-moderated remote testing and examines those protocols in real life settings.

Keywords
Remote, human-moderated, user testing, protocols, test administration

1. INTRODUCTION

Local user testing, at a portable or established laboratory, has been the customary method used in determining the effectiveness of a product. Local user tests are conducted throughout the product development lifecycle, each time with different objectives.

Depending on those objectives, several types of tests are implemented. Exploratory tests validate the product, determine users' mental models and suggest an appropriate design strategy. Assessment tests verify the soundness of the exploratory test. Task completion and ease of use are examined. Finally, validation tests serve as verification of a product. Performance measures are evaluated to ensure that the product objectives have been met [Rubin, 1994].

In a market spreading steadfast across the globe, it is important to involve the right mix of users from the appropriate constituencies. This conscious choice results in a product that is easier to use, more useful, efficient and desirable to its intended audience. The requirements of a global market lead us to this new challenge which can consequently result in a product's success or failure. Remote user testing allows development teams to reach a greater number of users regardless of *location* by addressing the need for cost-effective and time efficient evaluation. Remote user testing provides answers to a number of difficulties encountered when planning a user test.

1.1 International user testing

When considering global product development, cost-effective user tests have emerged as a new challenge. Global products necessitate the most diverse user base be tested. Examples of compromised usability have been found in cases where the full spectrum of end-users was not adequately represented [Cleary, 2000; see also Kawaguchi, 2000]. Remote testing eliminates travel expenses and costs incurred by additional team members traveling to client sites by placing users in closer reach. In addition, it allows a single evaluation team to conduct all tests, ensuring consistency in the evaluation technique and results.

1.2 Offshore development

Opportunities for remote evaluation also present themselves in the business of offshore development. Offshore development has responded to a weakened economy by raising talent in countries such as India, Russia and China. Large teams are employed offshore and although costs are minimized, quality expectations remain the same. Provided that offshore teams receive appropriate training, the network provides an ideal means of transport for user testing [Nielsen, 2002] by allowing users, and the offshore team to work together regardless of distance.

1.3 Users' natural environment

Traditionally conducted in a laboratory, usability testing presents difficulties in reproducing the user's environment. In many cases it is an arduous task to even find representative users [Hartson & Castillo, 1998]. User environment and targeted users are essential in the success of a user test. Remote methods provide a gateway to accessing users in their natural environment.

1.4 Post-deployment

Product development is a progressive process. New versions of products are introduced as they evolve over time. Introducing improvements and new features requires post-deployment usability data. Although post-questionnaires and surveys are a valid source of information about customer satisfaction, they are in no way a substitute for user testing. Remote testing is a natural solution to the low-cost requirements of post-deployment usability testing [Hartson & Castillo, 1998].

In conclusion, remote testing is a suitable solution to a variety of scenarios, all of which have resulted from an ever-growing global market. It successfully supports the requirements brought forth by universal design.

The majority of the research available has been directed towards remote automated methods. Those methods can be applied without the presence of an evaluation team. In fact, the essence of those applications lies in quantification, which can be extracted by software applications as efficiently as by humans [e.g. Scholtz, 2001; Rodríguez & Gutierrez, 2000; Hartson & Castillo, 1998]. Our interests lie in *human-moderated* remote testing where although testing is conducted remotely, a facilitator and an evaluation team are present.

For the purpose of this paper we will be referring to user testing in three ways. First, as 'local user testing' or 'user testing', meaning a test that is conducted with the evaluation team present. Second, as 'remote automated testing', which is user testing with the assistance of software and without the evaluation team present. Third, as 'human-moderated remote testing' referring to user testing where the evaluation team is present but not located in the same premises as the user during the time of the test. This type of testing takes place with the help of videoconferencing software. Finally, the term 'remote testing' includes both automated and human-moderated remote testing.

2. HUMAN-MODERATED REMOTE USER TESTING

Previous attempts to include a moderator in the various types of user testing have been made.

Open University introduced a human moderator in LCR, a method strongly based on contextual inquiry, implemented in the exploratory stage of user testing. Their objective was to facilitate a discussion leading to more subjective findings [Rapanotti, Dunckley & Hall, 2002].

Sun Microsystems, Hewlett Packard and IBM among others currently use videoconferencing tools for exploratory, assessment and validation tests [Hammontree, Weiler & Nandini, 1994; see also Bartek & Cheatham, 2003].

Hartson, Castillo, Kelso, and Neale [1996] present a case study comparing local tests to human-moderated remote validation tests. Their informal study concludes that the two methods appear to produce similar qualitative and quantitative results.

Although remote user tests have been examined in the context of facilitation, our contribution is aimed in defining remote test administration protocols and determining their validity in practice.

Specifically, our objectives are the following:

- Determine strengths and weaknesses of proposed human-moderated remote protocols
- Compare and contrast local to human-moderated remote testing

2.1 Protocols

The protocols proposed in this section will extend local user testing into human-moderated remote testing. The guidelines outlined give us the opportunity to define human-moderated remote testing and replicate it.

Although a variety of evaluation methods exist, we have chosen to extend user testing on the grounds of its proven success in identifying usability problems [Nielsen, 1993]. However, we will emphasize that the majority of suggested protocols can be adapted to most evaluation methods.

Cultural Considerations in Team & Test Materials Preparation

Remote testing is frequently implemented for international user testing. The distance involved in this situation is not merely of a physical sort, but one, which concerns the differences in the users' language, mental models and social habits. These potential issues should be considered first when assembling the evaluation team. The evaluators should possess cultural sensitivity with respect to the intended users in order to accurately respond to users and collect viable results, which will lead to targeted interpretations [Westat, 2002]. Another key consideration is the adaptation of test materials. Usability documentation should always be translated into the user's native language [Trillo, 1999]. Besides the language barriers, test objectives and data collection methods should reflect the users' cultural background [Westat, 2002]. As regards legal requirements, laws pertaining to human subject treatment vary from country to country. In the context of international testing, it is important to know which principles have global and which local application [Burmeister, 2001].

Training

The nature of remote, as opposed to local user testing demands that the user participate more actively in the test preparation process. The user is often required to install software and operate it prior to the test [Hartson et al., 1996], which leads to the need for training and support. The training approach taken is dependent on whether the usability team will assign the user to an active role in the test or if he/she will be mostly guided by the moderator. User experience is also essential in deciding the extent and nature of training needed. Novice users are in need of more guidance, both in the software installation process but also as regards how comfortable they feel about their participation.

Software

Videoconferencing software such as Microsoft NetMeeting, Lotus SameTime and many others, offer the connecting line between user and evaluation team with features such as application sharing and video transfer. When choosing a software package, it is good to consider using software already installed on the participants' machine. This approach reduces the total test time and makes it more likely that the user will already be familiar and comfortable with the specific software package.

Pre-test administration

Installing the required videoconferencing software may add extra time to the user test. As a result, it is common practice to administer the pre-test questionnaire and informed consent form at a time prior to the test. This aims to reduce the overall test time, give the users the opportunity to read the materials at their leisure and eliminate any unneeded stress that derives from the impersonal nature of remote testing. The pre-test materials should stress the evaluation teams' willingness to discuss potential questions during the user test.

Several readings suggest that the materials be online for issues of time management [Bartek & Cheatham, 2003]. It is important to take into account the users' experience with computer interfaces. Online questionnaires may require additional completion time for novice users or for those who simply prefer printed materials.

Communication

During the test session, the user and the team have three avenues of communication. First, application sharing allows the evaluation team to observe and appropriately respond to the user's onscreen behavior. This feature simulates real-time screen observation. Second, a phone connection serves as auditory feedback and provides a primary vehicle for qualitative data extraction. Using the built-in phone capacity of Video Conferencing software is not recommended due to observed low sound quality, which may negatively impact on the user test [Rapanotti et al., 2002]. Third, video transfer suggests two possible benefits. It establishes a more personal and friendly environment between team/user and also contributes to the qualitative data collection. We caution those who might employ video transfer that it should not be solely depended on for qualitative feedback due to the possibility of low quality data transfer.

Task Distribution

Task distribution is usually given sequentially. In a remote setting it can be simulated through the Whiteboard or Chat tool. Each task is presented one at a time. New tasks are not presented to the user until the current task is completed. User biasing is avoided by ensuring that tasks are not read prior to the test [Hammontree et al., 1994]. Also, cognitive overload is a possible reaction in the event that the user reads all tasks simultaneously.

Think-aloud

Usability problems are revealed during user testing by collecting a number of measures such as verbal, non-verbal cues and task-related benchmarks. Remote testing presents a new challenge in error interpretation due to the lack of non-verbal cues. In overcoming this limitation, a special emphasis on the think-aloud protocol is imperative.

The most commonly referenced think-aloud methodology, Ericsson and Simon's, is designed to capture the contents of short-term memory. Their efforts focus on determining the roots of human problem solving [Ericsson & Simon, 1984]. However, we will deviate from this application towards the speech communication approach as discussed by Boren and Ramey [2000; see also Ramey & Boren, 2001].

Boren and Ramey expand the think-aloud model, originally created by Ericsson and Simon, into an alternative theory based on speech communication. Their approach has several advantages over the original model. First, the listener and speaker roles are acknowledged and set in the beginning of the user test. The participant is established as the work domain expert and primary speaker. The test moderator on the other hand takes the role of the listener. Second, the human moderator can intervene at various parts of the user test. Such interventions are carefully planned in the form of acknowledgements and reminders. They consist of neutral language so that the results are not biased in any way. Third, special circumstances that may arise during a user test are taken into account. Among others, these include dealing with system crashes, a user's inability to continue with the task given and a user's confusion on whether a task has been completed [Boren & Ramey 2000].

We feel that a communication approach, as against one of more passive observation, is essential in reducing the physical distance between evaluation team and user.

Post-questionnaire and Debriefing

To conclude the testing session, a post-test questionnaire is administered, with a choice between a printed or online format. Following the test, the participant and moderator reconvene through the existing phone and video connection for the debriefing.

2.2 A pilot study

We conducted a comparison test between human-moderated remote and local user testing to measure the validity of our protocols.

We selected to test the web site of a local Chicago community bank. The bank operates four branches, all located in the midst of Chicago's most culturally diverse neighborhoods. The staff employed at those banks is carefully chosen to reflect the cultural background of its customers who are primarily Greek, Arabic, Hispanic and Polish. We chose to test our assumptions outside an academic environment with the aim of achieving a realistic application with targeted users.

Our users were recruited by advertising posted throughout all the bank branches. Word of mouth also sparked interest in potential participants.

Our test objectives were to determine how often customers and internal personnel accessed the bank web site, how easily they were able to locate information and what additional tools

they might need in the future. The tasks administered tested the most problematic and frequently used pages as reported by the banks' customer service.

We decided that it was more effective to conduct the testing with a remote terminal located in the bank. A significant number of customers visit the bank for their transactions, which lends itself to our use. Our users also belonged to all ages and varied in the level of their computer experience. Bringing the hardware to them decreased extra training time required while eliminating technical issues that might arise in an uncontrolled environment.

Our software of choice was Microsoft Netmeeting. All the necessary hardware and software was pre-installed on the users' and evaluation teams' computers. Pilot tests were run to ensure the test and computer environment met expectations.

Local user testing sessions took place in a conferencing room located in the bank. The moderator welcomed the participant, introduced him/her to the two observers and began testing soon after. In the remote setting, the user was greeted by the branch manager, directed to the conferencing room and given the test packet. The user was then left alone and the test began. Our comparison test was run with a total of twelve participants.

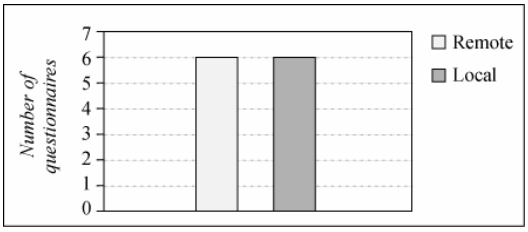
Our evaluation team was assembled with cultural considerations in mind. The bank is very diverse in both its customers and internal staff. A contributing factor to team member selection was their cultural origin and/or background.

The data we chose to collect in our comparison test is meant to measure the most important components of user testing.

2.3 Results

Questionnaire Completion

In both tests the questionnaires were administered in paper format. We expected remote users to demonstrate few, if any, difficulties during questionnaire completion. Our assumption was reinforced by the fact that the questionnaires were in the traditional paper form.

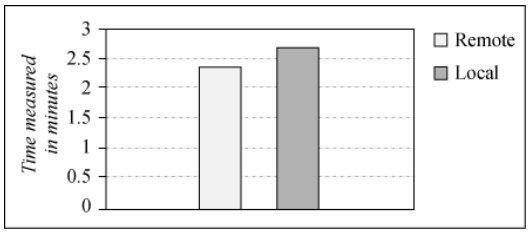


Our expectations were validated by our results. Both user groups were able to complete the questionnaires successfully with and without the physical presence of the evaluation team.

Think-aloud

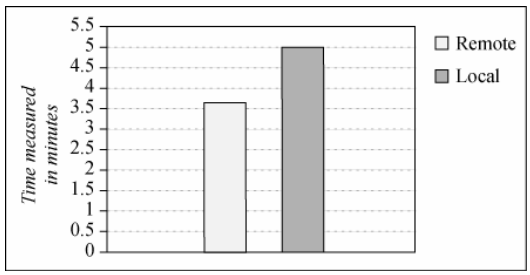
Thinking aloud was 'induced' by two methods, user training in the beginning of the test and continuous prompting by the moderator. Taking into account the distance factor between team and user, we were led to believe that remote users would strictly adhere to the think-aloud protocol as a means of eliminating that distance. We measured this assumption by timing users of both tests as they thought aloud. Timings were collected by extracting the think-aloud instances from the total test time. 2-second separator delays were added between each instance, simulating natural speech.

Our findings were quite contrary to our predictions. An unpaired t-test concluded P=0.704. The results produced no significant difference between the local and human-moderated remote test.



Qualitative data

Due to the communicative design of the human-moderated remote test, we saw no reason for the qualitative data yielded by both tests to be dissimilar. We recognize that qualitative data depends on the breadth of the information received. However, interpretation is a subjective matter and as a result the quality of subjective data remains a challenge to measure. Taking that into account, in addition to using the data for error interpretation, we used timing as a way to objectively evaluate the amount of qualitative data we received from each user. Each user was timed throughout the test, and later the qualitative data provided was extracted from the total time. As with the think-aloud, delays were added in between the sentences.

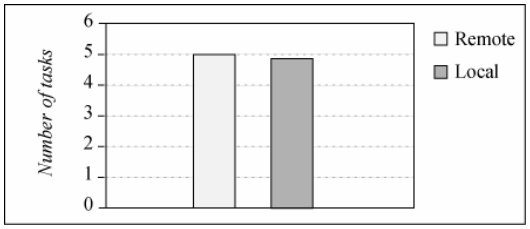


Our results demonstrated no difference between the two methods (P=0.204.) There was however, an observed reduction in the amount of qualitative data given in the human-moderated remote test.

Task Completion

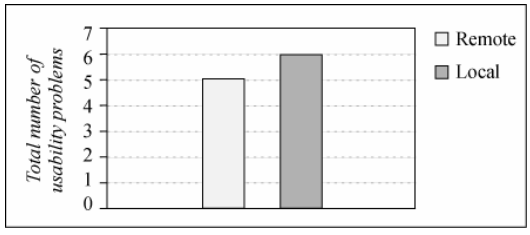
Tasks were administered in paper format. The moderator walked remote users through the tasks sequentially as would typically happen in a local test. We expected users to encounter little difficulty in task completion.

Our expectations were confirmed. Participants of both tests were able to complete approximately the same number of tasks (P=0.734).



Number of Usability problems found

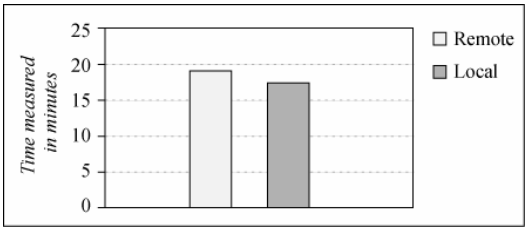
Data collected in both tests was measured in qualitative and quantitative terms. We collected two types of usability problems, ones that were detrimental to in task performance and ones that contributed to performance deterrence. We expected an equivalent number of usability problems to be discovered in both tests.



Our findings confirmed our assumptions. There was no statistically significant difference (P=0.418) between the two methods.

Total test time

We expected the total time of the human-moderated remote test to exceed that of the local test. Our assumption was supported by the amount of training required prior to the test. Remote users were given more detailed introductions and debriefings. We also expected that technical problems encountered would take longer to resolve due to the distance.



Contrary to our beliefs, both tests were similar in length of time (P= 0.621). However, the average remote test time was proportionately higher than the local one. Time spent on user training, during one of our remote sessions, contributed to a significant increase in the total test time.

Efficiency of communication

In the human-moderated remote test, communication between the team and user was conducted by application sharing using Netmeeting, through a phone connection and by a web cam, which allowed both sides to see each other. Although there were several seconds of delay in the video transfer, we believed that users would feel increased comfort by putting a face to the team members. We tested the success of our communicative approach by administering a post-test questionnaire to all remote users. Our questions were designed to determine whether the remote test setup was easy to use, whether the environment was maintained friendly and whether there was sufficient assistance given when problems arose. Users were asked to state how comfortable they were at the beginning and the end of the test session.

The results yielded were promising. All users with the exception of one rated the aspects of remote testing high as concerns ease of use and comfort. One of our least experienced users found the session to be stressful and uncomfortable in the beginning of the test but rated her level of comfort high at the end of the session. We can conclude that all participants felt at ease and communication through remote means was successful.

3. DISCUSSION

Our test design, although successful in its purpose, did not thoroughly address all proposed protocols.

- Training was conducted during the test in terms of user comprehension. Hardware/software installation and training were not required.
- The physical presence of a greeter may have heightened user comfort with the test.
- Although attention was given to cultural background, it was confined to the factor of user bilingualism.
- Test materials were distributed in paper form, leveraging the presence of the greeter, as well as rendering the test a one-time stop instead of a more complex process. Digital test material distribution such as online questionnaires was not factored in.

Although our pilot study may have tested the success of a few protocols, it should be noted that each test design addresses a different set of needs. The use of all protocols combined during a single test is unlikely.

Taking a closer look at our assumptions and the results yielded in the remote setting, we observed reduced total tests timings and think aloud timings, results other than the ones anticipated. Several reasons may have contributed to this deviation. The moderator was not physically present, which may have reduced remote users' inclination to speak at greater length. Another contributing factor may have been skills acquired by the moderator. The local test session was conducted on a separate occasion prior to the human-moderated remote test. As a result, during remote testing the moderator was aware of most usability problems and may have been more efficient in accommodating users when these occurred.

In addition to the above, we were surprised to find proportionately lower qualitative data timings. In contrast to that, the usability problems discovered in both tests were almost identical, potentially signifying that the objectivity we attempted to attain when assessing qualitative data may depend on the breadth of information rather than on time measurements. Further investigation would be needed to ascertain this point.

4. FUTURE WORK

Future test design refinements would include counterbalancing the two sessions to account for skills acquired by the test moderator. In addition to that, a larger pool of users would be needed to establish the success of human-moderated remote testing with greater certainty. We would also like to obtain better insight into the efficiency of human-moderated remote testing when working with users of variable computer experience and cultural backgrounds. During our pilot test we found communication via the network to be hampered by the inexperience of our least experienced user. The authors expect that users' behavior will be affected by experience, age and culture. Qualitative data and general test times are another topic we would like to investigate further, to discover whether there are correlations between the amounts of qualitative data provided by users and the usability problems uncovered during the test session.

In conclusion, we found that our suggested protocols for human-moderated remote testing contribute to a close simulation of local user testing. Supporting this statement are our reported findings. The number of usability problems discovered in our remote sessions was comparable to that of local user testing. When considering a low cost alternative for local user testing, conducting human-moderated remote testing is a viable alternative solution, which does produce concrete results leading to usable, efficient and targeted products for a global market.

REFERENCES

[Bartek & Cheatham, 2003] Bartek, V. & Cheatham, D. *Experience remote usability testing, Part 1*. 2003. [On-line]. Available: http://www-106.ibm.com/developerworks/web/library/wa-rmusts1.

[Boren & Ramey, 2000] Boren, M.T. & Ramey, J. Thinking Aloud: Reconciling Theory and Practice. *IEEE Trans. Prof. Comm., 43,* 2000, pp 261-278.

[Burmeister, 2001] Burmeister, O. Usability testing: revisiting informed consent procedures for testing Internet sites. In J. Weckert (Ed), *Volume 1 - Computer Ethics 2000*. Canberra, AU, 2001, pp. 3-10.

[Cleary, 2000] Cleary, Y. An Examination of the Impact of Subjective Cultural Issues on the Usability of a Localized We Site – The Louvre Museum Web Site. *Archives & Museums Informatics*, 2000. Available: http://www.archimuse.com/mw2000/abstracts /prg_80000197.html

[Ericsson & Simon, 1984] Ericsson, K. A. & Simon, H.A. *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge, MA, 1984.

[Hammontree et al., 1994] Hammontree, M., Weiler, P. & Nandini, N. Remote Usability Testing. *Interactions, 1(3)*, 1994, pp. 21-25.

[Hartson et al., 1996] Hartson, H. R., Castillo, J., Kelso, J. & Neale, W. Remote Evaluation: The Network as an Extension of the Usability Laboratory. *Proceedings CHI'96 Conference*, Vancouver, Canada, 1996.

[Hartson & Castillo, 1998] Hartson, H. R. & Castillo, J. Remote evaluation for post-deployment usability improvement. *Proceedings of AVI '98 Conference*, L'Aquila, Italy, 1998.

[Kawaguchi, 2000] Kawaguchi, M. The X-Factor. *Newsweek*, 2000.

[Nielsen, 1993] Nielsen, J. *Usability Engineering.* Academic Press, Boston, 1993.

[Nielsen, 2002] Nielsen, J. *Offshore Usability.* [On-line]. Available: http://www.useit.com/ alertbox/20020916.html

[Ramey & Boren, 2001] Ramey, J. & Boren, R. Keep Them Thinking Aloud: Two Ways to Conduct a Verbal Protocol and Why It Matters. *Proceedings 2001 Conference Usability - A Winning Experience*, Las Vegas, NV, 2001.

[Rapanotti et al., 2002] Rapanotti, L., Dunckley, L. & Hall, J. G. Extending Low Cost Remote Evaluation with Synchronous Communication. *Proceedings 16th British HCI Group Annual Conference*, London, UK, 2002.

[Rodríguez & Gutierrez, 2000] Rodríguez, M. & Gutierrez, D. Data Gathering Agents for Remote Navigability Testing. *Proceedings SCI'2000 Conference (Systemics, Cybernetics and Informatics)*, Orlando, FL, 2000.

[Rubin, 1994] Rubin, J. *Handbook of Usability Testing*, John Wiley & Sons, Canada, 1994, pp.30-46.

[Scholtz, 2001] Scholtz, J. Adaptation of Traditional Usability Testing Methods for Remote Testing. *Proceedings of the 34th Hawaii International Conference on System Sciences,* Hawaii, 2001.

[Trillo, 1999] Trillo, N.G. The Cultural Component of Designing and Evaluating International User Interfaces. *Proceedings 32nd Hawaii International Conference on System Sciences*, Hawaii, 1999.

[Westat, 2002] Westat, J. F. *The 2002 User-Friendly Handbook for Project Evaluation.* NSF, Arlington, VA, 2002.