Synchronous Remote Usability Testing - A New Approach Facilitated By Virtual Worlds

Kapil Chalil Madathil
Clemson University
110 Freeman Hall Clemson SC 29634
mchalil@clemson.edu

Joel S. Greenstein
Clemson University
110 Freeman Hall Clemson SC 29634
iejsg@clemson.edu

ABSTRACT
This study proposes a new methodology for conducting synchronous remote usability studies using a three-dimensional virtual usability testing laboratory built using the Open Wonderland toolkit. This virtual laboratory method is then compared with two other commonly used synchronous usability test methods: the traditional lab approach and WebEx, a web-based conferencing and screen sharing approach. A study was conducted with 48 participants in total, 36 test participants and 12 test facilitators. The test participants completed 5 tasks on a simulated e-commerce website. The three methodologies were compared with respect to the following dependent variables: the time taken to complete the tasks; the usability defects identified; the severity of these usability defects; and the subjective ratings from NASA-TLX, presence and post-test subjective questionnaires. The three methodologies agreed closely in terms of the total number of defects identified, number of high severity defects identified and the time taken to complete the tasks. However, there was a significant difference in the workload experienced by the test participants and facilitators, with the traditional lab condition imposing the least and the virtual lab and the WebEx conditions imposing similar levels. It was also found that the test participants experienced greater involvement and a more immersive experience in the virtual world condition than the WebEx condition. These ratings were not significantly different from those in the traditional lab condition. The results of this study suggest that participants were productive and enjoyed the virtual lab condition, indicating the potential of a virtual world based approach as an alternative to the conventional approaches for synchronous usability testing.

Author Keywords
Synchronous remote usability testing, remote usability testing, virtual worlds.

ACM Classification Keywords
User interfaces, human factors, usability testing

General Terms
Human factors, experimentation, performance

INTRODUCTION
Usability studies on software interfaces analyzing how users interact with computer applications began in the early 1980’s [5, 22]. At this time, several usability evaluation methodologies evolved, the most common one being laboratory-based testing. This methodology, usually conducted in a lab equipped with audio and video recording capabilities, involves a test facilitator and participant in front of a one-way mirror with the application developers watching and recording the participant’s completion of the tasks assigned. User performance is then evaluated based on parameters such as speed, accuracy and types of errors. These quantitative data are combined with subjective information obtained through verbal protocols [10], critical incident reporting [7], and user satisfaction surveys [6]. Traditionally, usability evaluation has been conducted during the final stage of the design process, the cost and time requirements associated with it being significant. To address this issue, the early 1990’s witnessed research developing alternative cost-effective usability evaluation methods and the inclusion of usability as a product attribute early in the design process. These results led to the development of such methodologies as heuristic evaluation [19], discount usability testing [17], cognitive and usability walk-throughs [15, 3], formal usability inspection [17] and heuristic walk-throughs [21].

The emergence of high speed internet technologies has resulted in the concept of the global village and next generation products addressing its needs. In such a scenario where usability evaluators, developers and prospective users are located in different countries and time zones, conducting a traditional lab usability evaluation creates challenges both from the cost and logistical perspectives. These concerns led to research on remote usability evaluation, with the user and the evaluators separated over space and time. The development of the Internet has enabled usability testing to be conducted remotely, with significant cost savings [13]. However, remote testing may lack the immediacy and sense of “presence” desired to
support a collaborative testing process. Moreover, managing inter-personal dynamics across cultural and linguistic barriers may require approaches sensitive to the cultures involved [8]. Other disadvantages include having reduced control over the testing environment and the distractions and interruptions experienced by the participants' in their native environment [8].

The use of three-dimensional (3D) virtual world applications may address some of these concerns. Collaborative engineering was redefined when these worlds integrated high fidelity voice-based communication, immersive audio and data-sharing tools [9]. In addition, such 3D virtual worlds mirror the collaboration among participants and experts when all are physically present, potentially enabling usability tests to be conducted more effectively when they are located in different places. Virtual world applications are relatively new and as a result have been the focus of limited research. To address this need, this study compared the effectiveness of synchronous usability testing in a 3D virtual usability testing lab built using the Open Wonderland toolkit with traditional lab usability testing and WebEx, one of the commonly used tools for remote usability testing. The results of usability tests employing the three methodologies were compared based on qualitative and quantitative measurement of the work performed and feedback from the participants forming each team.

RELATED WORK
The research conducted by Hammontree et al. [11] on interactive prototypes at Sun Microsystems and Hewlett Packard is one of earliest studies to analyze the potential of remote usability testing. They used window/application sharing, an electronic white board, a computer-based video conferencing tool and a telephone to support the remote usability test. The shared windowing tools and telephone supported the remote think-aloud evaluations. The shared window facilitated the observation of the user interactions remotely. The researchers anticipated an improvement in the quality of tools designed to support remote collaborative work.

Remote testing, which facilitates evaluations being done in the context of the user’s other tasks and technology can be either synchronous or asynchronous [20]. The former involves real time one-on-one communication between the evaluator and the user, while the latter involves the evaluator and user working separately. Synchronous usability testing methodologies involve video conferencing or employ remote application sharing tools such as WebEx. Asynchronous methodologies include automatic collection of user’s click streams, user logs of critical incidents that occur while interacting with the application and subjective feedback on the interface by users. Numerous tools are available to address the needs of both these approaches. For example Microsoft NetMeeting, WebEx, WebQuilt and IBM Lotus Sametime support online screen sharing and collaborative capabilities for synchronous remote usability evaluation.

Numerous studies have been conducted comparing the effectiveness of synchronous remote usability testing methodologies. In an early such study, Hartson et al. [13] compared traditional lab usability testing to desktop video conferencing using Kodak's server pages as the interface to be tested. In the remote approach, a telephone connection was used to facilitate voice communication and subjective questionnaires were emailed to the remote users. The results of their study suggested that remote evaluation using video conferencing is feasible, producing similar results to the traditional lab approach.

Tullis et al. [24] conducted a study in which remote tests were conducted at the participant’s work location without real-time observation. In the remote approach, an email was sent, including a link to the website, explaining the purpose of the study. The participants used two windows, the first one presenting the task to be performed and the second the prototype application to be tested. User actions were automatically recorded and analyzed. Users were also provided with a subjective questionnaire to rate the difficulty of each task. The data analyzed included successful task completion rates, task completion times, subjective ratings and identified usability issues. The results of the study indicated that the task completion time and task completion rate from the remote and traditional approaches were similar.

To identify the differences in the qualitative experience from the participant’s and facilitator's perspective, Brush et al. [4] compared synchronous remote usability testing with conventional in-lab testing using a plug-in for the integrated software development environment Eclipse as the interface to be tested. The remote method was facilitated using a Virtual Network Computing based screen sharing program with the audio communication being established through a phone connection. The study revealed that there were no significant differences in the number of usability issues identified, their types and their severities. The participants felt that their contributions to the redesign of the interface were approximately the same in both conditions.

Thompson et al. [23] compared a traditional and remote approach to identify appropriate tools and methodologies for efficient and effective remote testing environments. In the remote approach, Microsoft NetMeeting and Snag-it were used, with the former providing the screen sharing capability and the latter the screen capture capability. A speakerphone was used to communicate with the remote participants.Both the remote and the traditional lab participants were asked to perform the same five search and shopping functions. The results suggested that there were no significant differences for time on task and number of errors. More recently, Andreasen et al. [1] compared the synchronous remote testing approach, with the traditional
lab approach and concluded that remote synchronous method is as effective as the traditional lab method.

Although the results for the two approaches appear to be similar, the disadvantages of remote studies include loss of control over the participant’s test environment, limited visual feedback, session security issues, ease-of-use issues and connection and system performance issues [2]. Dray et al. [8], suggest that “building trust with remote evaluations can be a real challenge”, especially in international remote testing, where the interpersonal dynamics of the evaluation must be managed across cultural and linguistic barriers.

HYPOTHESES
To compare the effectiveness of the online usability testing technologies WebEx and a virtual laboratory with traditional lab usability testing, the following null research hypotheses were tested.

Hypothesis 1: There will be no significant differences in the number and severity of usability defects identified in the three environments.

Hypothesis 2: There will be no significant differences in the time taken to complete usability test tasks in the three environments.

Hypothesis 3: There will be no significant differences in the participants’ subjective experience in the three environments.

Hypothesis 4: There will be no significant differences in the preference of facilitators for the three usability test methodologies.

The synchronous usability testing process involves extensive interaction between the test facilitator and the test participant, as the participant performs the test tasks and thinks aloud. In the virtual lab, the facilitator and participant can see one another’s avatars as they interact with the in-world applications, perhaps enhancing their sense of interpersonal interaction.

EXPERIMENTAL SETUP
Participants
Forty-eight college students familiar with Internet applications were recruited. The 12 test facilitators, 10 males and 2 females between the ages of 24 and 40, were required to have taken at least one course in usability while the remaining 36, consisting of 22 males and 14 females between the ages of 23 and 35, served as usability test participants. These 36 were equally divided among the three environments, 12 in a traditional lab usability test, 12 in a remote usability test using WebEx and the remaining 12 in a remote usability test using a virtual lab. The 12 test facilitators were paired with one participant in each of the environments. Thus, each test facilitator monitored three sessions, one in a traditional lab, one in WebEx and one in a virtual lab.

Testing environments
The independent variable of this study was the usability test methodology, examined at three levels: the traditional usability laboratory, the web-based meeting tool WebEx, and the virtual usability testing lab. The traditional lab usability environment consisted of a participant and a test facilitator physically located together in a lab to perform the usability test. The traditional lab usability test environment included a table, six chairs, one computer, and other supplemental materials, such as pens and paper, as shown in Figure 1. The WebEx environment provided a web browser for the participant and facilitator to share. The WebEx setup is shown in Figure 2. Two computers were provided, one for the usability test participant and the other for the test facilitator. The participant and facilitator were physically separated in different rooms. The virtual lab environment, which had a similar setup as of WebEx, consisted of a virtual usability testing laboratory equipped with an integrated web browser that can be shared, as shown in Figure 3. Using these tools, the participants and facilitators can interact with a web application and record their concerns with the design of its interface. In addition, the team members can use text and audio chat tools to communicate through their avatars. Two computers were provided, one for the usability test participant and the other for the test facilitator. The participants and facilitators were physically separated in different rooms.

Tasks
An E-commerce web application modeled after Amazon.com was developed with usability flaws deliberately embedded. This application was developed using php and deployed on an Apache Tomcat web server running the Windows XP operating system with a MySQL database providing the data tier. All test participants, regardless of the test environment, performed the following tasks on the website:

1) Your watch is not working, and you want to buy a new Swiss watch. After checking the price, add the watch to your cart.

2) Winter is over, and you heard that there are good deals on North Face jackets. Look for a North Face jacket and add two to the cart.

3) One of your friends is a fan of Dan Brown’s novels. Find Dan Brown’s latest novel “The Lost Symbol” and add it to the cart.

4) Look in the shopping cart and change the quantity of Swiss watches to two.

5) Check out.

EXPERIMENTAL DESIGN
The study used a mixed experimental design, with the test facilitators observing the test participants’ interactions with the web interface in a within-subjects design and the test participants experiencing the test environment in a between-subjects design. The experiment was counter-
balanced, such that two test facilitators conducted the usability test session first with the traditional lab method, then with WebEx and finally with the virtual lab environment and two test facilitators conducted the usability test sessions in each of five remaining possible orders.

Figure 1. Traditional lab setup

Figure 2. WebEx setup

Procedure

The facilitators and test participants employed the following procedure in all test environments. Initially the usability test facilitators were trained to conduct a usability test. Steve Krug’s usability test demonstration video [14] was used for this purpose, as well as to refresh the facilitators’ memories of the material in the usability testing class that they had taken.

At the beginning of each test session, the researcher greeted the test facilitator and the participant in a classroom and gave them a brief overview of the study. Then, the test facilitator was asked to read and sign a consent form and to complete a pre-test questionnaire, asking for their basic demographic information.

The test participant was also asked to read and sign a consent form and to complete a pre-test questionnaire to obtain their basic demographic information and their experience with relevant Internet technologies. Next, the facilitator was provided a list of tasks to be performed, including instructions on the tasks the participant would be asked to complete while using the web interface.

The facilitator and participant were then taken to their respective usability testing rooms and given a brief training session of approximately ten minutes to acquaint them with the environment. The test facilitator used Steve Krug’s test script [14] and gave the participant a sample task to familiarize him/her with the nature of the web application to be used during the test session. Next, the facilitator interacted with the participant as in a typical usability test session, asking him/her to complete the individual tasks. The researcher was co-located with the facilitator and recorded the time taken for each task using a stop watch. During each task, the test participant was asked to detail his/her concerns while interacting with the interface in a prospective think-aloud session. The researcher recorded the concerns raised and Camtasia, the screen capture software, was used to record all screen and audio activity.

Upon completing the final task, the participant and test facilitator completed the NASA-TLX workload assessment instrument [12] and the presence questionnaire [26]. The test participants also completed a post-test subjective questionnaire composed of three sections concerning their satisfaction with the usability testing methodology. The first section assessed the quality of the test environment. The user satisfaction portion evaluated perceived ease-of-use while performing the tasks, including how comfortable and confident participants felt in conducting the usability task and detecting the usability defects. The section on the
quality of the collaborative usability test methodology assessed the perceived level of presence and co-presence in the test environment. The participants ranked each metric using a 7-point Likert scale, ranging from 1 (Strongly Disagree) to 7 (Strongly Agree). The questionnaire also contained a section for written comments. Upon completion of this questionnaire, the participants were de-briefed by the researcher. The time taken for each session was approximately one hour. Once the test facilitators completed the three sessions in the three environments, they completed a post-test questionnaire assessing their satisfaction with the three usability testing methodologies and they were de-briefed.

Separately, a heuristic evaluation was individually conducted by three people who had experience conducting usability evaluations. During this analysis, the severities of the problems were also rated. Nielsen’s severity rating scale [16] was used as the basis for this rating. This scale ranges from 0 to 4, with 4 indicating a catastrophic defect.

**Dependent variables**

The three usability test methodologies were compared using objective and subjective measures. The objective measures consisted of the task completion time, the number of defects identified and the defects’ severity. The subjective measures consisted of the subjective data from the NASA-TLX, presence and the post-test questionnaires completed by both the test participants and test facilitators. The data for the number of defects identified were obtained from the observations of the usability test facilitator and analysis of the video recording of the test session. The severity of each defect was obtained from the heuristic evaluation data provided by the usability experts.

The data collected were classified into two sets:

1. Dataset of test participants, which consisted of 36 datasets, 12 for each condition.
2. Dataset of test facilitators, which consisted of 12 datasets, each across all the three conditions.

The evaluation of each dataset was conducted by performing a thorough walkthrough of the videos and analyzing the NASA-TLX, the presence and the post-test subjective questionnaires. The usability test facilitator datasets were analyzed based on the data from the post-test questionnaires.

SPSS 17.0 was used to analyze the data. Initially, a normality test was conducted to determine whether the data followed a normal distribution. All of the data satisfied this test. The data were then analyzed using a one-way ANOVA with a 95% confidence interval to determine the presence of significant differences, if any, among the test environments. If the null hypothesis of an ANOVA was rejected, the results were then subjected to a post-hoc least significant difference (LSD) test to determine the locus of the significant differences.

**RESULTS**

In this section, the three usability testing environments are compared with respect to the time taken to complete the tasks; the usability issues identified; the severity of these usability issues; and the subjective ratings from the NASA-TLX, the presence and post-test subjective questionnaires.

**Objective Measures**

**Time taken to complete the task.** The time taken to complete the task was measured from the time the task was given to the participants to the time they completed it.

The effect of the usability test environment on the task completion time was not significant. The mean task completion times for the five tasks are presented in Figure 4.

**Figure 4. Mean time taken to complete the tasks (in minutes)**

The results were then subjected to a post-hoc least significant difference (LSD) test to determine the locus of the significant differences.
Number of usability problems identified. The mean numbers of usability problems identified are presented in Figure 5. The effect of usability test environment on the total number of usability problems identified was not significant, F (2, 33) = 1.406, p=0.260. The effects of test environment on the number of Severity 1, 3 and 4 problems identified were also not significant.

Severity 2 defects identified. The effect of test environment on the number of Severity 2 problems identified was significant, F (2, 33) = 3.222, p=0.050. Subsequent post-hoc analysis reveals that there is a significant difference in the number of Severity 2 problems identified for the traditional lab and virtual lab conditions (p = 0.018). A larger number of severity 2 problems were identified in the virtual lab condition than in the traditional lab condition.

Test participants’ experience
Total Workload. The effect of test environment was significant, F (2, 33) = 4.00, p=0.028. Subsequent post-hoc analysis reveals that the total workload experienced in the traditional lab testing environment is lower than that experienced in the WebEx (p = 0.010) and virtual lab (p = 0.055) conditions, as shown in Figure 6.

Performance. The effect of test environment on the performance component of workload was significant, F (2,33) = 4.317, p=0.022. Subsequent post-hoc analysis reveals that the workload associated with the performance of the task was higher in the WebEx test environment than in either the traditional lab (p=0.010) or the virtual lab (p=0.028) test environments.

Frustration: The effect of test environment on frustration level approached significance, F (2,33) = 2.557, p=0.093. Subsequent post-hoc analysis suggests that this effect is due to differences between the traditional lab testing and virtual lab testing environments (p = 0.035). These results suggest that frustration was lower for the traditional lab condition than for the virtual lab condition.

No significant differences were observed for the NASA-TLX subscales of mental demand, physical demand, temporal demand and effort.

Presence Questionnaire:
The effectiveness of a virtual environment is to some extent dependent on the sense of presence experienced by its users [26]. The presence questionnaire categorized the overall usability testing experience into subscales of involvement, sensory fidelity, adaption/immersion and interface quality.

Figure 6. NASA-TLX workload indices for the test participants

Involvement. The effect of test environment on involvement was significant, F (2, 33) = 4.529, p = 0.018. Subsequent post-hoc analysis reveals that there is a significant difference between the WebEx and virtual lab testing conditions (p = 0.005). Test participants experienced a lower level of involvement in the WebEx condition than in the virtual lab condition, as shown in Figure 7.

Adaption/Immersion. The effect of test environment on adaption/immersion was significant, F (2,33) = 4.145 p=0.025. Subsequent post-hoc analysis reveals that there is a significant difference in adaption/immersion for the WebEx and virtual lab (p=0.007) testing conditions. Participants achieved a higher level of immersion in the virtual lab environment than in the WebEx environment.

The subscales of sensory fidelity and interface quality did not show any significant differences.

Subjective rating.
The subjective rating questionnaire totaled 15 questions, 4 asking about the naturalness of the environment, 5 asking about the satisfaction with and ease-of-use of the usability testing methodology and 6 questions on the quality of the usability test methodology. The mean value was calculated for each of these categories.

User satisfaction and ease-of-use: The effect of the test environment on user satisfaction and ease-of-use was
significant, \( F (2, 33) = 3.715 \) \( p=0.035 \), as shown in Figure 8. Subsequent post-hoc analysis revealed that the traditional lab test environment scored higher in user satisfaction and ease-of-use than the WebEx (\( p=0.044 \)) and virtual lab (\( p=0.015 \)) test environments.

The scales of naturalness of environment and quality of usability testing methodology did not show any significant differences.

**Test facilitator’s experience**

**Workload:** The effect of test environment on the total workload experienced by the test facilitator was significant, Wilks’ Lambda =0.461, \( F (2, 10) = 5.843 \) , \( p=0.021 \), as shown in Figure 9. Subsequent post-hoc analysis reveals that the total workload experienced in the traditional lab condition is lower than that experienced in the WebEx (\( p=0.015 \)) and virtual lab conditions (\( p=0.007 \)).

**Physical Demand:** The effect of test environment on physical demand placed on the test facilitator was significant, Wilks’ Lambda =0.485, \( F (2, 10) = 5.306 \) , \( p=0.027 \). Subsequent post-hoc analysis reveals that the physical demand experienced in the traditional lab testing environment is lower than that experienced in the WebEx (\( p=0.008 \)) and virtual lab (\( p = 0.023 \)) conditions.

**Effort:** The effect of test environment on the effort exerted by the test facilitator was significant, Wilks’ Lambda =0.445, \( F (2, 10) = 6.241 \) , \( p=0.017 \). Subsequent post-hoc analysis reveals that facilitators exerted less effort in traditional lab testing environment than in the WebEx (\( p=0.026 \)) and virtual lab conditions (\( p=0.050 \)).

**Frustration:** The effect of test environment on the frustration experienced by the test facilitator was significant, Wilks’ Lambda =0.429, \( F (2, 10) = 6.6 \) , \( p=0.014 \). Subsequent post-hoc analysis reveals that facilitators experienced less frustration in the traditional lab testing environment than in the WebEx (\( p=0.035 \)) and virtual lab (\( p=0.003 \)) conditions.

The NASA-TLX subscales of mental demand, temporal demand and performance did not show any significant effects for the test facilitators.

**Presence Questionnaire**

**Involvement:** The effect of test environment on involvement of the test facilitators was significant, Wilks’ Lambda =0.213, \( F (2, 10) = 18.468 \) , \( p=0.000 \), as shown in Figure 10. Subsequent post-hoc analysis reveals that facilitators experienced a lower level of involvement in the WebEx condition than they did in the traditional lab (\( p=0.004 \)) and the virtual lab (\( p=0.000 \)) testing environments.

**Sensory Fidelity.** There was a significant effect of test environment on the sensory fidelity experienced by the test facilitators, Wilks’ Lambda =0.155, \( F (2, 10) = 27.194 \) , \( p=0.000 \) as shown in Figure 10. Subsequent post-hoc analysis suggests that the experience of sensory fidelity was lower for the WebEx condition than it was for the virtual lab (\( p=0.000 \)) condition.

The subscales of immersion and interface quality did not show any significant effects for the test facilitators.
**Subjective Questionnaire**

**Ease of use:** The effect of usability test environment on ease of use for the test facilitators was significant, Wilks’ Lambda =0.162, $F(2, 10) = 25.78$, $p=0.000$, as shown in Figure 11. Subsequent post-hoc analysis reveals that facilitators found the traditional lab test environment easier to use than the WebEx ($p =0.006$) and the virtual lab ($p=0.001$) environments.

![Figure 11. Subjective metrics for the usability test facilitators.](image)

- **Ability to analyze user interaction with the web interface.** The effect of usability test environment on the facilitators’ ability to analyze user interaction with the web interface was significant, Wilks’ Lambda =0.178, $F(2, 10) = 23.10$, $p=0.000$. Subsequent post-hoc analysis reveals that the facilitators’ sense of presence with the test participant was higher in the traditional lab than in the WebEx ($p=0.000$) environment.

**DISCUSSION**

One of the initial research questions was to identify whether there were any objective differences in the effectiveness of the three evaluation approaches, the traditional lab approach, the WebEx approach, and the virtual lab approach, in collecting usability data. Though no significant differences were identified for the total number of defects identified and the number of Severity 3 and Severity 4 defects identified across the three environments. The results are consistent with those of Hartson et al. [13] for the conduct of a synchronous usability test in different settings. Similarly, the studies conducted by Brush et al. [4] found no significant differences between the traditional lab condition and the remote synchronous testing condition in terms of the number, types and severities of usability problems.

The effect of test environment on the number of Severity 2 defects was significant. More Severity 2 defects were identified in the virtual lab condition than in the traditional lab condition. It is not clear why the participants in the virtual lab condition identified more minor usability defects than the participants in the other conditions. Perhaps the novelty of the virtual lab condition increased the motivation of the participants to detect and mention even minor usability issues.

Significant differences in total perceived workload were observed for the three conditions. The NASA-TLX scales, used to determine the total perceived workload, indicate that the participants experienced the least workload in the traditional lab condition. This result could have been due to the co-location of the test facilitator with the test participant during the preparatory stages as well as during the test. In the WebEx-based approach, test participants experienced some difficulty during the preparatory session figuring out how to operate the system. In addition, there was a time delay while the remote session loaded. Though this delay was short, participants complained that they did not know what was happening other than that there was a message displayed explaining that the session was loading. For the virtual lab-based testing, participants clicked on a button to launch the session and soon were transitioned to the virtual world. For both of the remote testing environments, the participants were required to perform a series of button clicks on an interface supporting the remote infrastructure to begin the session.

Significant differences were also found for the performance and frustration subscales of the NASA-TLX. The test participants felt their performance was poorer in the WebEx environment than in the traditional lab and virtual lab environments. They experienced more frustration with the virtual lab environment than with the traditional lab environment. The participants using the virtual lab environment appeared to be frustrated primarily by its slow response to the inputs of the test participants and test facilitators. Nonetheless, a number of test participants using the virtual lab environment commented that they enjoyed moving around inside the virtual usability testing laboratory and interacting with the shared web browser using their avatars.

The NASA-TLX workload indices suggest that the total workload experienced by test facilitators was also lower in
the traditional lab environment than in the two remote testing environments. Significant differences were also observed for physical demand, effort and frustration subscales. Test facilitators felt that physical demands, effort, and frustration were higher for the two remote testing environments than for the traditional lab environment. This may be due to the lower initial setup required for the traditional lab condition. It took time for the test facilitators to become acquainted with the remote testing environments. In addition, the high level of physical demand in the WebEx and virtual lab conditions might also be due to the higher level of interaction with the computer required in these environments during the study. The relatively slow response times of the web browser to user inputs in the remote testing environments may also have led to increased levels of frustration for the test facilitators.

For the test participants, significant differences were observed for involvement and immersion on the presence questionnaire. Involvement was higher in the virtual lab environment than in the WebEx environment. For the test facilitators, involvement was higher in the virtual lab and traditional lab conditions than in the WebEx condition. The level of immersion experienced by the participants was also higher in the virtual lab condition than in the WebEx condition. These results may be due to the multisensory immersive experience provided by the three-dimensional virtual lab as well as its use of avatars to represent the test participant and facilitator.

The subjective ratings provided by the test participants in the final subjective rating questionnaire revealed significant differences in terms of user satisfaction and ease of use. User satisfaction and ease of use were higher for the traditional lab methodology than for the remote testing environments. The test facilitators also rated the ease of use of the traditional lab environment higher than that of the remote test environments. Interestingly, however, some of the test participants in the traditional lab environment commented that they felt some pressure to avoid making mistakes while being observed by a test facilitator. None of the participants in the remote testing environments expressed this concern.

Test facilitators felt that they were best able to analyze the user’s interaction with the web interface in the traditional lab environment. They felt they were least able to analyze the user’s interaction in the virtual lab environment. This may be the result of a deficiency in our implementation of the virtual lab using the Open Wonderland toolkit. Mouse movements made by the test participant within a browser were not visible to the test facilitator in the virtual lab environment.

One of the challenges of remote usability testing is the recruitment of security-conscious participants. These participants, or the organizations employing them, may consider allowing others to access their computers to be a security risk [25]. Remote testing using a virtual lab requires only that participants interact with their web browser. The test facilitator cannot view any information on their computers that is not displayed within the browser. WebEx addresses this concern by allowing its users to selectively share applications on their computers. Another difficulty encountered in remote usability testing is the need to install a client application to enable screen sharing and chat functionalities [25]. WebEx requires that users install either an ActiveX control or a Java applet on the computer at each end of the conference to enable screen sharing. The virtual lab methodology relies on Java applets to achieve these functionalities. As a result, there is no need for the users to install a program to enable these functionalities. In addition, remote usability testing with the virtual lab and WebEx retains the inherent advantages of synchronous remote usability testing, including significant savings in travel time and cost, reduced turnaround time for user-centered iterative product development, recruitment of geographically dispersed participants and the ability for geographically distributed product development teams to participate in a real-time study.

CONCLUSION

This study proposed a new methodology for conducting a synchronous remote usability test using a three-dimensional virtual world, and empirically compared it with WebEx, a web-based two-dimensional screen sharing and conferencing tool, and the traditional lab method. One of the most important findings of this study is that the virtual lab method is as effective as the traditional lab and WebEx-based methods in terms of the time taken by the test participants to complete the tasks and the number of higher severity defects identified. Interestingly, participants appeared to identify a slightly larger number of lower severity defects in the virtual lab environment than in the traditional lab and WebEx environments.

Test participants and facilitators alike experienced lower overall workload in the traditional lab environment than in either of the remote testing environments. Both the test participants and test facilitators experienced a higher level of involvement in the virtual lab condition than in WebEx condition. Moreover, this method supports recruitment of geographically distributed, diverse participants who can remain in their native work environments. Given that it generates usability test results comparable to those of traditional lab testing, remote usability testing in virtual world appears to be a viable alternative to the conventional lab testing approach. The primary disadvantage of testing in the virtual lab environment was the delay the participants experienced while interacting with the interface. The significance of this study is that the synchronous remote usability testing using virtual lab provides similar results as of the traditional lab method and in some respects it works better than the conventional approaches.

ACKNOWLEDGEMENT

Thanks to Nicole Yankelovich and Jordan M. Slott for the advice and helpful comments.
REFERENCES
24. Tullis, T., Fleischman, S., McNulty, M., Cianchette, C. and Bergel, M. An Empirical Comparison of Lab and Remote Usability Testing of Web Sites. In Usability Professionals Association Conference. (Florida, USA),2002