Comparative Study of Synchronous Remote and Traditional In-Lab Usability Evaluation Methods.

Prakaash Selvaraj

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Industrial and Systems Engineering

Tonya L. Smith-Jackson, Ph.D., Chair
Brian Kleiner, Ph.D.
Yee-Yin Choong, Ph.D.

March 24, 2004
Blacksburg, Virginia

Keywords: remote usability evaluation, critical incidents, usability method, synchronous remote usability testing

Comparative Study of Synchronous Remote and Traditional In-Lab Usability Evaluation Methods.

Prakaash V. Selvaraj

ABSTRACT

Traditional in lab usability evaluation has been used as the 'standard' evaluation method for evaluating and improving usability of software user interfaces (Andre, Williges, & Hartson, 2000). However, traditional in lab evaluation has its drawbacks such as availability of representative end users, high cost of testing and lack of true representation of a user's actual work environment. To counteract these issues various alternative and less expensive usability evaluation methods (UEMs) have been developed over the past decade. One such UEM is the Remote Usability Evaluation method.

Remote evaluation is a relatively new area and lacks empirical data to support the approach. The need for empirical support was addressed in this study. The overall purpose of this study was to determine the differences in the effectiveness of the two evaluation types, the remote evaluation approach (SREM) and the traditional evaluation approach, in collecting usability data. This study also compared the effectiveness between the two methods based on user type, usability novice users and usability experienced users. Finally, the hypothesis that users, in general, will prefer the remote evaluation approach of reporting to the traditional in-lab evaluation approach was also tested. Results indicated that, in general, the synchronous remote approach is at least as effective as the traditional in lab usability evaluation approach in collecting usability data across all user types. However, when user type was taken into consideration, it was found that there was a significant difference in the high severity negative critical incident data collected between the two approaches for the novice user group. The traditional approach collected significantly more high severity negative critical incident data than the remote approach. Additionally, results indicate that users tend to be more willing to participate in the same approach as the one they participated previously. Recommendations for usability evaluators for conducting the SREM approach and areas for future research are identified in the study.

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank all of my committee members for their advice and constant encouragement. I would like to thank Dr. Yee-yin Choong for providing me with opportunities to work in various facets of usability engineering and choosing an interesting topic for my research. I would also like to thank Dr. Brian Kleiner for his flexibility, valuable suggestions and willingness to help me out with my last minute requests. Additionally, I would like to thank my advisor, Dr. Tonya Smith-Jackson, for her unwavering support and encouragement. If not for her guidance and enthusiasm in my work none of this would have been possible. I could not have asked for a better team of people to advise and guide me through the course of this work.

I would also like to extend my thanks to all my friends here at Virginia Tech who was my emotional support through out this period. Last but not least, I would like to thank my family for believing in me and providing me with unending support and encouragement all through my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1. Project Background

The Usability Engineering (UE) team at Global eXchange Services (GXS) is currently collecting usability data on their software applications (in-development stage) from their remote customers through a synchronous remote usability evaluation method. The GXS UE team uses WebEx, a comprehensive information sharing and collaborative web conferencing tool for conducting their remote evaluations. Using this web conferencing tool and a global network, such as the World Wide Web, a remote participant is able to drive an application on the evaluator's computer. The UE team envisions this approach as intermediate between their other approaches of traditional in-lab usability testing and heuristic evaluation. The heuristic evaluation method conducted by expert usability engineers lacks representative user data and in the former method it is difficult or impossible to reproduce the users' work context in a laboratory setting (Hartson, Castillo, Kelso, and Neale, 1996). The synchronous remote evaluation approach used by the UE team addresses both of the above issues by using real users in an ethnographic method. Also the remote assessment method allows users to work in their natural work environment and hence makes it most compatible with ethnographic methodologies (Smith-Jackson & Williges, 2001).

Usability testing, in the past, was restricted to collecting summative usability data. This restriction was due to the lack of methods to collect formative usability data of real users in their daily work settings. Traditional lab testing, though effective, had its limitations and was not applicable for collecting such user representative data. This led to the development of remote usability evaluation approaches. Apart from formative usability data, usability practioners are currently interested in real usage data that can be collected after an application has been

deployed. This additional information would aid in future usability improvements. A remote approach would be the best method to capture such valuable information.

1.2. Problem Statement

There is no fixed method for conducting all usability evaluations. The choice of method is dependent on more than one factor. It can be dependent on the area of interest, objectives of the project or the availability of resources (Thompson, 1999). Resources can be classified as cost of conducting the evaluation, time, support technology, or even the availability of participants from the representative user class. The UE team at GXS had been conducting usability evaluations either through formal in-lab evaluation or expert heuristic evaluation. Lately, due to limited access to local representative users and budget restrictions, the UE team had to review the literature for other forms of usability evaluation methods other than traditional in-lab usability evaluation. On reviewing the literature, the many benefits of remote evaluation began to emerge. These benefits combined with the literature and the availability of existing technological resources (resources like the GXS intranet, the Internet and web conferencing tools) led to the development of the synchronous remote usability evaluation method.

A review of the literature revealed that that although there are many approaches to remote evaluation, there is a definite lack of empirical studies comparing remote evaluation approaches to the formal in-lab evaluation. Hartson et al. (1996) conducted feasibility studies on a couple of approaches to remote usability evaluation (teleconferencing and semi-instrumented) and Castillo (1997) made a future recommendation for a formal study to compare remote evaluation to formal in-lab evaluation. Hence, the need for a comparative study of the synchronous remote evaluation method with the formal in-lab usability method was recognized. This research would help in providing the empirical data that has been lacking, for the use of remote evaluation approaches as a cost-effective and efficient alternative to formal in-lab evaluation.

1.3. Research Justification

Software companies of the present generation are increasingly catering to varied and distributed international customers. A suitable example would be the report published by Hammontree, Weiler, and Nayak (1994) stating that Hewlett Packard development teams in the U.S. are designing custom software for users in Europe, Australia, and Asia-Pacific. Traditional in-lab usability evaluation of such products with representative users would have encountered many obstacles like availability of representative users and high cost of testing due to transportation of participants. Remote usability evaluation is one of the means of counteracting the above issues. Though there is some research available on the various approaches to remote evaluation, and its necessity acknowledged, there is still a lack of empirical data in support of the method and the need for more research is strongly noted.

1.3.1. Need for Empirical Support

Many approaches were found, on reviewing the literature, for conducting remote usability evaluation but few have any empirical support (Thompson 1999). For example, Hammontree et al. (1994) report successful implementation of a remote version of the collaborative walkthrough. Though the success observed on the field is widely accepted, it lacks the empirical justification that is needed to support the remote evaluation approach as a viable and efficient method. Very few studies have used empirical methods to assess the effectiveness of approaches used for remote evaluation. Hartson et al. (1996) and Tullis, Fleischman, McNulty, Cianchette, and Bergel (2002) conducted empirical studies comparing in-lab and remote usability evaluation methods. These studies provided support for the feasibility of remote approaches as alternate means of conducting usability evaluations. Despite establishing the feasibility and need for remote evaluation approaches additional empirical support is needed to

state that remote usability evaluation yields comparable if not better user data and establishing the remote method as a suitable inexpensive alternative to traditional in-lab evaluation.

1.3.2. Participant availability

One issue that most usability evaluators face while conducting tests is the availability of representative users. Participant availability can be dependent on many factors (i.e. geographical location, skills, time, cost, etc). The process of setting up sessions with potential participants and bringing them on-site can be an expensive and time-consuming process. Development teams work on applications in "Web time" and cannot afford to allocate time in transporting and scheduling representative users. Apart from scheduling issues the frantic pace of society and the subsequent difficulties faced by participants in allotting time to participate needs to be considered. Finally, the user group of interest might be scattered in various extreme geographical locations and local evaluation might not be feasible.

1.3.3. Monetary Benefits

Remote evaluation in any form will reduce the cost of conducting usability tests of software applications. Costs incurred in traditional in-lab testing in terms of facilities, travel, personnel needed for testing and lab time are drastically reduced through a remote evaluation method. The network being an intrinsic part of any organization also plays an important role in providing usability evaluators with a more cost effective way of building distributed usability labs on the network. As mentioned in the earlier sections, more and more companies are developing software for international user groups and evaluation of those applications with representative users would incur additional expenses in the form of travel and living expenses. Therefore, a simple cost-benefit analysis would indicate that the potential savings of a remote method of evaluation is tremendous. Thus, a study that may provide an indication of the validity

of this possibility would be well warranted. If the presumably less costly remote evaluation method is found to be a valid alternative to in-lab testing, it should also help to facilitate more positive reactions to usability from management personnel.

1.4. Document Overview

This section will give a brief overview on the remaining sections of this document. The rest of the document is categorized into six main categories—literature review, research overview, methodology, results, discussion, and conclusion.

2. LITERATURE REVIEW

2.1 Traditional in-lab usability evaluations

The most commonly used and standard form of formative usability evaluation is the traditional in-lab usability evaluation. It has been used as a benchmark method for comparison with various remote evaluation methods (Hartson & Castillo, 1998). It has also been referred to as the 'ultimate' evaluation method to be used as a standard for comparison of usability evaluation methods (Andre, Williges, & Hartson, 1999). This kind of evaluation can be referred to as a local evaluation as the evaluator and the participant are in the same location at the same time. Traditional in-lab evaluations are considered to be very formal and are driven by quantitative usability specifications with predefined tasks (Whiteside, Bennett, and Holtzblatt, 1988). The use of end users is one of the many advantages of this method. Quantitative data (e.g. task completion time and number of errors) and qualitative data (e.g. critical incident description and verbal protocol) are collected in this method. Quantitative data provide information on the level of usability achieved and qualitative data help in identifying potential usability problems and their causes within the interface design (Hartson & Hix, 1993). Although traditional in-lab usability evaluation seems to be the ideal method for conducting evaluations it has its own drawbacks. Drawbacks such as availability of end users at the location evaluation site, lab setting expenses, and the artificial lab environment in the sense that it might not be a true representation of the user's actual work environment (Hartson & Castillo, 1998).

2.2 Critical Incident Technique

One of the most important types of data collected in a usability evaluation session is critical incident data. del Galdo, Williges, Williges, and Wixon (1987) defined a critical incident

as an interaction event with a feature or element of a system that is either very easy or difficult to accomplish and that results in very effective or extremely ineffective performance. The identification and reporting of this event information constitute the critical incident technique. Also, specifically in the domain of formative usability evaluation, a critical incident can be defined as an event or occurrence that indicates a positive or negative aspect of the usability of the system.

2.2.1 Origin

Fitts and Jones first codified the critical incident technique in 1947, but the term "error" was used in place of "critical incident". They applied the technique formally in a study for analyzing and classifying pilot error experiences in reading and interpreting aircraft instruments. Although the technique was first established in 1947, the first formal paper describing the critical incident technique was published by John Flanagan in 1954. He described the critical incident technique based on his analysis of various generic studies on observing and collecting human activity information in the field. These activities were called "critical incidents" due to their crucial role in system behavior, purpose, effects on human performance, and clear identifiable start and end points.

Flanagan defined originally (1954) the critical incident technique as "a set of procedures for collecting direct observations of human behavior in such a way as to facilitate their potential usefulness in solving practical problems and developing broad psychological principles" (pg, 327). In the earlier studies that were conducted by Flanagan the observance of users' activities was performed under uncontrolled conditions mostly by untrained observers. Hence, Flanagan envisioned a more controlled systematic method where trained, domain knowledgeable,

observers need to be used for recording human behavior that met specific predefined criteria and also in order to main objectivity the observer should only make simple type of judgments.

2.2.2 Application of CIT in Usability Evaluations

The most important aspect of the critical incident technique is its flexibility. Flanagan emphasized that this technique is a flexible set of principles that can be adapted to any specific area or situation and not as a rigid set of rules for data collection. Although the technique is a relatively old concept, it has been applied to the area of usability evaluation effectively in some form in recent years (del Galdo et al., 1987). According to a report by Shattuck & Woods (1994), the critical incident technique has been used often in the past, but rarely in its original form. As predicted by Flanagan, it was adapted and modified to meet the needs and limitations of the respective applications or study goals.

In order to be applied to the area of usability evaluation in human computer interaction, the critical incident technique had to undergo a change in concept rather than a minor modification. This was initiated by the work of del Galdo, Williges, Williges, and Wixon (1987) on evaluating the critical incident technique for software design documentation. The reason for selecting the critical incident technique was based upon its focus on user involvement and its ability to be flexible. However, the flexibility had to be more than minor modifications; it had to be a refinement of the concept. Flanagan (1954) defined the critical incident technique based on a behavior-based approach but in the field of usability evaluation it needs to be more of a cognitive-based approach. In the original method human behavior alone was the crux of the technique but in usability evaluation area it is the interaction between the user and a certain aspect of the system that is of importance. Del Galdo et al. (1987) redefined the critical incident as the combined and interdependent behavior of the users and the interface with which they

interact. In their study, they also made additional modifications to the way in which data were collected. Participants, rather than trained observers, were asked to report critical incidents when they happened during their task and the need for a trained observer was eliminated. This improved technique was implemented by del Galdo et al. (1987) in a couple of iterative studies, and was found to be effective in collecting qualitative data on human computer interaction. Also the elimination of the need for a trained observer in this technique led to the adoption of the technique by researchers for remote evaluation approaches (Hartson et al., 1996)

2.3. Remote Evaluation

In-lab usability evaluations are considered the standard evaluation method. However, there are advantages and disadvantages to in-lab usability evaluation. The limitation is that it does not capture representative usage information in an actual true distributed work environment of the user. Also it cannot be applied to situations where in the user population of interest is small or distributed and remote. As more and more companies are widening their reach and cater to a diverse and dispersed clientele, the need for alternative techniques for conducting effective usability evaluations has become apparent. Advances in support tools and technology in the past few decades have made it possible to conduct usability evaluations remotely over the Internet or any form of computer networks.

Remote Evaluation is defined as usability evaluation where the test evaluator and the participant are separated in space and/or time. The term remote is relative to the evaluator and refers to the remote location of users from the evaluator's location (Hartson & Castillo, 1998). Figure 1 is a representation of a remote evaluation session based on the illustration developed by Castillo (1997),

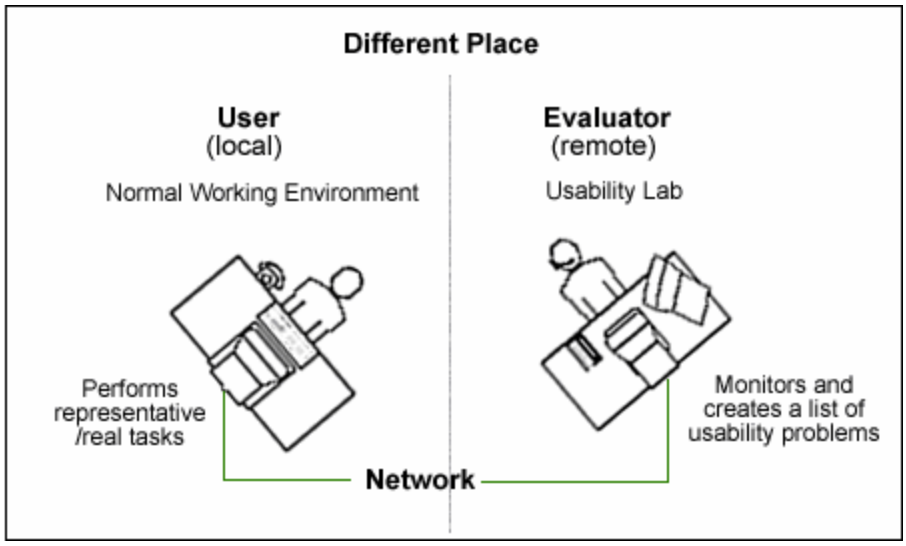

Figure 1: Scenario of a remote usability evaluation session

In a remote evaluation session, the evaluator observes the participant through a computer network using technologies such as computer based video/audio conferencing, window/application sharing tools, and other whiteboard applications.

2.3.1 Remote Evaluation & Ethnography:

Before delving further into intricacies of remote evaluation, it should be acknowledged that the remote evaluation method has roots in established approaches used in other sciences. Ethnography is one of many field methods used to collect user-centered data. Field methods can be broadly defined as a collection of tools and approaches for conducting studies of users, their tasks and their work environments in the context of their everyday work culture or environment (Wixon, Ramey, Holtzblatt, Beyer, Hackos, Rosenbaum, Page, Laakso, and Laakso, 2002). Although the ethnographic approach of studying users performing their day-to-day tasks in their work environment has long been used in other areas like sociology and anthropology, it has recently been adapted to the field of HCI.

Ethnography in HCI is practiced in various forms and methods. Contextual inquiry and observational study are few methods. Each of these approaches is compatible with one or more

principles of ethnography. According to Blomberg et al. (1993) the ethnographic approach is based on the following four main principles

- Study user behavior in a natural setting
- Adopt a holistic view that behavior can only be understood in the context in which it occurs
- Represent members' point-of-view
- Need for descriptive understanding

As the remote evaluation approach is focused on using real users in their natural environments to collect user driven usability data, it also complies with ethnographic approaches (Smith-Jackson & Williges, 2001).

2.4 Approaches to Remote Usability Evaluation

In the past decade, different methods have been developed for conducting usability evaluations remotely. In some of these approaches, the interactions between the evaluator and user are in real time (synchronous) while others are not in real time (asynchronous). Users' level of involvement in the remote approach might be active or passive (i.e. either participate or not participate in the data gathering process). For example, in a technique called Instrumented Remote Evaluation, the application to be tested is coded to automatically collect continuous data and the user is passive in the data gathering process (Hartson et al., 1996). A list of different techniques to remote evaluation identified and defined by Hartson et al. (1996), Hammontree et al (1994), and Castillo (1997) are listed and briefly discussed in the following sections:

- Remote questionnaire/survey.
- Remote think aloud.
- Design storyboard walkthroughs.

- Instrumented remote evaluation.
- User-reported critical incident method (Semi-instrumented remote evaluation)
- Video conferencing.
- Third party services.
  - Remote inspection.
  - Third-party laboratory evaluation
- Remote control evaluation

2.4.1 Remote questionnaire/survey

In this method, the application being evaluated can be augmented to elicit subjective information from the user in the form of an online questionnaire. The launch of the questionnaire can be triggered by either the completion of the task or occurrence of a specific event or series of events. An example of this technique is the User Partnering (UP) module from UP technology (Abelow, 1993). It uses event driven triggers to 'awaken' dialogues that ask users for subjective information on their usage. Responses are batched and sent via the Internet to the developers. The collected data are limited to fresh subjective data, but lacks the qualitative data that are usually acquired in lab sessions.

2.4.2 Remote thinking aloud

The user and the evaluator are connected via the network with video and audio connection. As in the usual think aloud process, the user goes about performing the pre-defined representative tasks provided by the evaluator. The user is encouraged to speak continuously about their actions and their perception of the application. The evaluators play the role of an observer and take notes. To conduct such a method for evaluation, the minimum requirement is

that the prototype be accessible and controllable by the user and all interactions be visible to both the user and the evaluator.

2.4.3 Design storyboard walkthroughs

Using shared window/application tools the evaluator remotely assists the user to walkthrough a predefined series of linear paths through the interface that addresses the task domain of interest. A shared whiteboard surface is used for sharing the storyboard screens and eliciting feedback from the user and another shared word processing window to send descriptions of the task scenario to the user. The shared white board is also used for receiving information from the user about problems, design changes or suggestions.

2.4.4 Instrumented remote evaluation

Instrumented remote evaluation is a completely automated usability evaluation. The application to be evaluated can be instrumented with embedded code for capturing user data related to user actions and storing them in the form of journals or logs. User actions can be any representative actions performed during the task, like keystrokes, mouse clicks, program usage, and project time. Evaluators using pattern recognition techniques (Siochi & Ehrich, 1991) later analyze the data logs to determine when and where usability problems occurred. Automatic and accurate problem detection methods are possible. Swallow, Hameluck, and Carey (1998) have worked on the development of problem indicator criteria to facilitate the automation. The main advantage of this approach is the non-interference with the user's work activities. However, it is not suitable for formative evaluation as it might be difficult to interpret certain usability problems effectively. Instrumented remote evaluation also needs more human resources to review and analyze the large quantities of collected data. However, instrumented remote

evaluation has been successfully used for automatic evaluation of certain kinds of usability problems in summative evaluation.

2.4.5 User-reported critical incident method

Hartson et al. (1996) developed an approach called the User Reported Critical Incident Method. It was initially referred to as the Semi-instrumented Remote Evaluation method. This approach uses selective data collection triggered directly by users while performing tasks in their daily work routine. The user is given minimal training to identify critical incidents as they appear and report specific information relative to the critical incident (ex. problem description, system state and severity of problem) while interacting with the application. Reports are then sent to the developers along with context information about the user task and system state. Additional information in the form of video clips of screen capture is also sent to the developers. This is an asynchronous method where the user and evaluator do not interact in real time and attend to the process at different times and at different locations. Evaluators then use these data to identify and classify usability problems. As this approach does not require the evaluator to participate in the identification of the critical incidents, the onus is completely on the user's capacity to accomplish the same with minimal training. This complete dependency on the user might be the only drawback of this approach.

2.4.6 Third party services

Commercial contract services for conducting usability evaluations are increasingly available and present software developers with an option of "contracting out" their occasional usability evaluation needs. The use of network is used restrictively only for communication and test material exchange, the network is not used for connecting with remote users. The evaluation is 'remote' to the developers and 'local' to the contracted third party.

Such an approach is most suited for software developers who do not have the necessary capabilities or can afford to own such facilities. However, the methods employed and the quality of services provided by such parties may vary and the applicability of the process to specific needs of a development team may not be suitable. Brief descriptions of some of those third party services are provided below.

Remote inspection

Some developers send design documents, software samples, and/or prototypes to remote evaluators for evaluation. These evaluations are performed using ad hoc, intuitive inspection of interface and heuristic evaluation of the application. The use of design guidelines, software guidelines and user profiles may be applied in the setting up of the evaluation. Due to the lack of empirical process and real user involvement, results are dependent on the skills and knowledge of the remote evaluators.

Third-party laboratory evaluation

Unlike the remote inspection approach, this is a formal laboratory-based evaluation with representative users and tasks provided by commercial consulting groups in their usability lab, which is remote to the developers. Results, as in any formal laboratory evaluation, include quantitative data such as performance measures, task completion time, and qualitative information (such as user satisfaction, recommendations for future improvements, and user opinions). In some cases a portable usability evaluation kit is used for the evaluation (Rowley, 1994). In such an evaluation, the usability lab is taken to the user's workplace. Equipment comprised of a laptop computer, video camera, mixer, and other audio/video equipment are transported to the user's work environment and set up to conduct the evaluation. Users participate actively in the data gathering process as in a traditional in-lab evaluation. While this

approach can be considered as an alternative to formal in-lab usability evaluation, it cannot be classified as an approach to remote evaluation due to the fact that the user and the evaluator are in the same location and not separated in space or time.

2.4.7 Live or Collaborative evaluation

Live or collaborative evaluation can be conducted by the following two methods:

Video conferencing

In traditional in-lab usability the video and audio information of the session is highly critical. Advances in technology have given us the opportunity to capture such information in a remote session. Users separated into remote locations, such as their work environments, can be connected to evaluators using the network and established video and audio links. This approach of using video conferencing software over the network as a mechanism to transport video data in real time is one of the closest approaches to usual lab testing. However, limited bandwidth, communication delays and low video frame rates are some of the drawbacks of this method.

Remote control evaluation

This method is very similar to the previous approach. However, in this method the evaluators have control over the remote user's computer using web conferencing software (e.g., GoToMyPC ™, WebEx ™). An audio connection can be established via the computer or by a separate phone line. The user's actions are captured using a screen capture of the evaluator's computer, which is connected to the user's computer. This approach allows the user to participate from their work environment and also has the advantage of being asynchronous or synchronous in real time. When the user and the evaluator are linked via an audio connection, the evaluator can collect data synchronously in real time. Alternatively, data can be collected asynchronously by automating the data collection procedure. When the user and the evaluator are

separated in time, data capture can be a continuous on going process or can be triggered by using a particular application. This added benefit gives the evaluator the flexibility of conducting the evaluation at their convenience. This study, synchronous remote usability evaluation, is based on the real time data capture approach of this remote control evaluation.

2.5 Synchronous Remote Usability Evaluation

The synchronous remote usability evaluation method is a remote evaluation method that is very similar to the traditional in-lab usability evaluation. It involves real users participating in the evaluation process from their own natural work environments using their own computers. In the following Table 1 by Castillo (1997), characterizing different situations possible in remote evaluation in terms of user location (space) and time of evaluation, the synchronous remote evaluation, due to its remote convenience and close similarity to the established in-lab usability evaluation, is located at the top right quadrant.

Table 1: Characterization of user location and time of evaluation

		Time of evaluation	
		Different (Asynchronous)	Same (**Synchronous**)
USER LOCATION	User's Working Environment	User-reported critical incident method. Instrumented data collection Remote questionnaire Live or collaborative evaluation.	**Synchronous remote evaluation**
	Controlled Environme	N/A	Third party laboratory evaluation *Traditional in-lab usability evaluation*

In this collaborative approach the remote user's computer (remote location) is connected to the evaluator's computer (usability lab) through the Internet using commercially available web conferencing software (e.g., GoToMyPC ™, WebEx ™). As this is a synchronous approach, the user and the evaluator are connected in real time by an established audio connection using a telephone line. Using the share screen feature of the web conferencing software the remote user will be able to view and use the website or application loaded in the evaluator's computer. The evaluator monitors the user's actions via the computer in the usability lab and video records the user's screen actions using a scan converter.

As shown in Table 1, the synchronous remote evaluation approach mimics the established formal lab evaluation as close as possible, in a remote setting. This approach also provides the best of both worlds, a list of possible benefits/advantages of the approach are listed below:

- Employs real users of interest.
- As users participate from their work environment and not a sterile usability lab this approach captures usage patterns pertinent to their actual work environment.
- This approach gives the experimenter full control over the session, but in the meanwhile does not hinder usage patterns pertaining to the user's work environment. The evaluator's role is well within the boundaries of the ideal trained expert as defined by Flanagan.
- Controlled tasks that are developed from real world scenarios.
- Addresses circumstances where users of interest are distributed and unavailable locally.
- As this approach eliminates the need for participants to travel for on-site testing, thousands of dollars can be saved on travel and facilities costs (Hartson et al., 1996).

The following section, research overview, will encompass information on the sub sections – research purpose, research questions and hypotheses of the proposed study

3. RESEARCH OVERVIEW:

3.1. Research Purpose

The main purposes of this research were to:

1) Compare the synchronous remote usability method with the traditional in-lab usability evaluation method and determine its effectiveness in collecting usability data.
2) Compare the subjective feedback data on evaluation type preference collected from users of both the remote evaluation approach and the traditional evaluation approach.
3) And secondarily, investigate the effectiveness of the approaches based on the type of users, novice users and experienced users.

3.2 Research Questions and Hypotheses

The study was designed to answer the following questions:

1) What are the differences in the effectiveness of the two evaluation types, the remote evaluation approach and the traditional evaluation approach, in collecting usability data?

To answer this question, the following hypothesis was tested. The remote approach will be at least as effective as the traditional approach in collecting usability data. Results supporting this hypothesis would indicate that the remote evaluation approach is as good and effective as the traditional approach in collecting usability data.

2) Which evaluation approach is preferred by different kinds of users (usability novice, usability experienced)?

To answer this question, the hypothesis that both novice and experienced users will prefer the remote evaluation approach of reporting to the traditional in-lab evaluation approach was tested. Support for this would imply that both kinds of users would opt for the remote approach in the future.

3) What are the differences in the effectiveness of the two evaluation types, the remote evaluation approach and the traditional evaluation approach, for the novice user group?

To answer this question, the hypothesis was tested that the remote approach will be more effective in collecting usability data with the novice user group compared to the traditional approach with the same user group.

4. METHODOLOGY

4.1. Experimental Design

A 2x2 between subject design was used for this study. The independent variables were the usability evaluation type and the user type. Table 2 below illustrates the various factors and their levels.

Table 2: Data Matrix

User Type (U)	Usability Evaluation Type (E)	
	Traditional	*Remote*
Usability Experienced	S_1 - S_8	S_9 – S_{16}
Usability Novice	S_{17} - S_{24}	S_{25} - S_{32}

4.1.1 Usability Evaluation Type (E)

The usability evaluation type factor consisted of two levels, the remote approach and the traditional approach. Each approach had equal numbers of unique participants and both approaches were used in evaluating a common website. A common set of predefined meaningful tasks was provided to the users of each approach.

4.1.2 User Type (U)

A pre-test background information questionnaire (Appendix B) was used to classify users into the two different user profiles, experienced users and novice users.

Experienced users were people who had prior exposure to usability engineering activities (prior experience in conducting usability testing and/or evaluation studies, user interface design, or at

least educated in usability principles as a student in the following fields of computer science, psychology or human factors.) and novice users—people who had no experience or knowledge of usability and were new to the field.

4.2. Participants

Thirty-two participants were recruited for the purposes of this study from the student population at Virginia Tech. Each approach was allocated 16 participants. Participants were classified into one of the two earlier mentioned user profiles. The sixteen participants for each approach consisted of an equal number of participants from both the user profiles. Users were randomly assigned to one of the evaluation approaches, after determining their user profile. Table 3 provides the detailed demographic distribution of the participants across the approaches.

Table 3: Demographic information of participants

User Type (U)	**Usability Evaluation Type (E)**	
	Traditional	*Remote*
Usability Experienced	Male: 5 Female: 3 Age Mean(SD): 30.38 (9.1)	Male: 5 Female: 3 Age Mean(SD): 26.25 (3.4)
Usability Novice	Male: 5 Female: 3 Age Mean(SD): 25.25 (3.2)	Male: 3 Female: 5 Age Mean (SD): 29.63 (9.4)

4.3. Materials and Equipment

4.3.1 Traditional in lab usability evaluation

The schematic of the HCI lab where the testing was carried out is shown below.

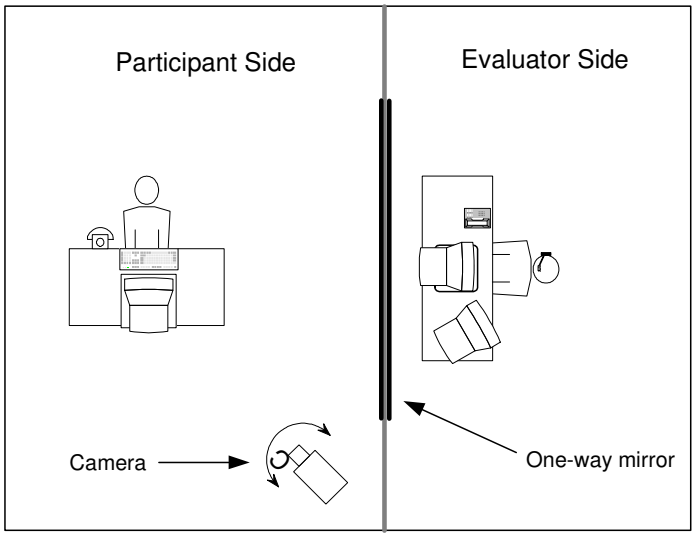

Figure 2: Schematic of a usability lab at Virginia Tech

The participant's area was equipped with a Pentium based, Windows 2000 workstation, video camera and audio recording devices. The computer used for the study was connected to the Internet and equipped with a browser (version 4 or higher). The evaluator's room was separated from the participant side by a wall and was equipped with a screen viewing and recording facility. Screen capture software was used for capturing the participant's screen footage, which was combined with the audio data and recorded along with the video stream.

4.3.2 Synchronous remote usability evaluation

The remote participants were not provided with any equipment. However, participants were requested to abide to certain requirements, such as a computer with Internet access and a browser (version 4 or higher), and also simultaneous access to a telephone line throughout the session.

The evaluator was located at a usability lab similar to the one at Virginia Tech that was described earlier and was connected to the participant telephonically and through the Internet. WebEx ^{TM}, a web conferencing software was used for connecting the participant and evaluator through the Internet.

4.3.3 Performance test

A performance test was administered to determine the effects of the independent variables. The performance test consisted of a list of tasks the user had to perform in the website. These tasks were representative of scenarios that a regular user would experience when using the website. These tasks provided an opportunity to gather information on variables such as time spent on completing a task (task-on-time), and the number of total negative and positive critical incidents that occurred and subsequently identified in a session. In order to maintain control and match both the synchronous remote evaluation approach and the traditional evaluation approach closely, the tests were conducted in one single sitting. Also, data collected during the test session was restricted to the observation of participants' screen actions and think aloud (verbal comments) in both the approaches. This was done to restrict the traditional evaluation approach from collecting data based on participant's facial expressions, which is not possible in the remote evaluation approach as the evaluator will not be able to see the participants face.

In addition to the number of critical incidents another measure that was collected was the severity of the critical incidents. This severity is a representative measure of the quality of the usability problem found by the Usability Evaluation Method (UEM) (Andre, Williges, & Hartson, 1999). According to Andre et al. (1999), to evaluate the effectiveness of a usability evaluation method, specifically for comparison of evaluation methods, evaluators need to establish a comparison criterion. Also, the realness can be determined by comparing with a standard list. In this study, apart from the number of critical incidents, the severity of critical incidents was one of the criterions. However, measuring the quality or severity of a critical incident can be difficult. In this study a standard list of rated critical incidents were compiled by expert review of the critical incidents of the de facto traditional in-lab method. Then the

severities of critical incidents found through the remote approach, were measured by comparing and matching them against the standard expert's list of rated critical incidents.

Finally, it was important to collect subjective data on user satisfaction with respect to the usability of the evaluation method. These subjective data provide important information on ease of use and user comfort in participating in the usability evaluation session. The participants were asked to rate the usability of the method using subjective ratings (Appendix C). Also towards the end of the session, in order to measure the participants' preference, they were provided with a description of the other approach (the one in which they did not participate) and asked for their preference for usability evaluations in the future.

4.4. Procedure

All participants underwent the same procedure, irrespective of their location (whether remote or lab-based). In the case of the lab-based approach, the evaluator was physically present whereas in the remote session the evaluator interacted through a telephone. Participants were greeted by the evaluator and given a brief introduction about the study. On completion of the introduction, participants were asked to read and sign an Informed Consent Form (Appendix A). On signing the Informed Consent Form participants were asked to complete a pre-test questionnaire. The purpose of this questionnaire was to gather basic background information about the participant.

Participants were then provided with a task list with detailed instructions, containing scenarios that a typical user would experience when using the web application. In this study the IMDB.com website was used as the test application. IMDB.com is a comprehensive internet movie database with wide range of information pertaining to films. A sample task employed in the evaluation was to find the top ranked movie in the website's top movies list. Upon

completion of the tasks, a retrospective think aloud session was scheduled where participants were asked to recall their thoughts while being shown a recording of their task performance (Ericsson & Simon, 1993). Finally participants were asked to complete a subjective questionnaire (Appendix C) and were then debriefed.

4.4.1 Evaluator's role

The evaluator was, at all times, located in the evaluator area of the usability lab during the test session and was recording the whole evaluation session. In both the approaches, the participant's screen activities were captured using a screen capture software and combined with the audio data. The evaluator also constantly observed and took notes on participant behavior (on-screen) and other significant instances that might hinder participant performance. These observational data comprised of critical incidents experienced by users, verbal comments by participants upon completion of the test session, and also the data logs of completed sessions.

The evaluator provided assistance, in both approaches, only if the participant was unable to complete the task and requested for assistance. At the end of the performance test, the evaluator collected the subjective feedback and then debriefed the participants and thanked them for their participation.

5. RESULTS

This section consists of the analyses of data collected from the performance test and subjective data collected at the end of the session. All analyses were conducted using the Statistical Analysis System (SAS). Alpha was set at 0.05. The dependent measures tested for significance are listed in Table 4.

Table 4: Dependent Measures and Descriptions

Dependent Variable	Description
1. Participation time	The length of time taken on the performance test measured from the start of the first task to the completion of the last task.
2. Critical Incident data	
a. Total critical incidents	The total count of all critical incidents (positive and negative) that were identified in the test session.
b. Positive critical incidents	The total count of positive critical incidents that were identified in the test session.
c. Negative critical incidents	The total count of negative critical incidents that were identified in the test session.
d. High severity negative critical incidents	The total count of negative critical incidents that critically impacted user performance and usability of the website.
e. Medium severity negative critical incidents	The total count of negative critical incidents that moderately impacted user performance and usability of the website.
f. Low severity negative critical incidents	The total count of negative critical incidents that were minor issues and did not affect user performance and were more or less just an irritant.

Table 4 (continued): Dependent Measures and Descriptions

Dependent Variable	Description
3. Subjective feedback data	
a. Overall rating	The rating provided by the participant on overall experience in participating in the test session on a scale of 1 to 10 with one being poor and ten excellent.
b. Comfort	The rating provided by the participant on how comfortable they felt in participating in the test session on a five point scale with a rating of 1 for strongly disagree and 5 for strongly disagree.
c. Ease	The rating provided by the participant on how easy it was for them to participate in the test session on a five point scale with a rating of 1 for strongly disagree and 5 for strongly disagree.
d. Realism	The rating provided by the participant on how realistic it was for them to participate in the test environment on a five point scale with a rating of 1 for strongly disagree and 5 for strongly disagree.
e. Convenience	The rating provided by the participant on how convenient it was for them to schedule and participate in the test session on a five point scale with a rating of 1 for strongly disagree and 5 for strongly disagree.
f. Future participation	The rating provided by the participant on how willing they would be in participating in the future in a similar evaluation session on a five point scale with a rating of 1 for strongly disagree and 5 for strongly disagree.
g. Preference for future participation	Participant preference of evaluation type, either traditional or remote, for future participation.

5.1 Hypothesis 1

Hypothesis-1 was that the remote approach would be at least as effective as the traditional approach in collecting usability data. Two dependent measures were used to assess this hypothesis

- Participation time
- Critical incident data (total, positive, negative, high severity, medium severity, and low severity).

5.1.1 Participation time

Participation time values were subjected to a two-way analysis of variance (as shown in Table 5), having two levels of evaluation type (remote, traditional) and two levels of user type (novice, experienced).

Table 5: ANOVA for Evaluation type (E) x User type (U)

Source	**df**
Evaluation Type (E)	1
User Type (U)	1
U x E	1
Subjects (S/UE)	28
Total	**31**

No significant order effects were found, F (1, 28) = 0.76, p = 0.389. Participation time of users in the remote evaluation method (M = 540.19, SD = 239.22) was not significantly different from the participation time of users in the traditional evaluation method (M = 471.69, SD =193.03). Table 6 presents the summarized participation time data for all groups.

Table 6: Participation time (seconds)

Evaluation type	User type	Participation time - Mean (SD)
Remote Evaluation	Novice	573.50 (244.18)
	Experienced	506.87 (245.90)
	Total	540.19 (239.22)
Traditional Evaluation	Novice	508.50 (183.78)
	Experienced	434.87 (207.30)
	Total	471.69 (193.03)

5.1.2 Critical incident data

The critical incident data (frequency count data) were subjected to Chi-Square test for independence between the two evaluation methods. However, as the expected counts for the Chi-square test were low, Fisher's exact test was performed as an alternative test on the same data. The results of both the tests for all the critical incident data are summarized in Tables 7.

Table 7: Test Results for Critical Incident Data (Evaluation Method).

Critical Incident Data – Number of	Results of tests: χ^2 test for Independence	Fishers	Interpretation
Total CIs	χ^2 (8, N=32) = 4.920, p = 0.766	p = 0.893	Not significant
Negative CIs	χ^2 (7, N=32)= 5.819, p = 0.561	p = 0.633	Not significant
Positive CIs	χ^2 (5, N=32)= 5.714, p = 0.335	p = 0.347	Not significant
High severity negative CIs	χ^2 (3, N=32)= 6.756, p = 0.080	p = 0.080	Not significant
Medium severity negative CIs	χ^2 (7, N=32)= 8.071, p = 0.326	p = 0.307	Not significant
Low severity negative CIs	χ^2 (4, N=32)= 4.443, p = 0.349	p = 0.387	Not significant

No significant differences were found between the traditional and remote evaluation approaches in terms of the number of critical incidents collected. However, the results for the high severity data showed that the difference between the traditional approach (M = 2.000, SD = 0.894) and remote approach (M = 1.250, SD = 0.856) may have been approaching significance (p = 0.080, Fishers exact test). Figure 3 provides a graphical representation of the means for both the approaches. A detailed table of mean and standard deviation for the number of critical incidents collected for each group can be found in Appendix E.

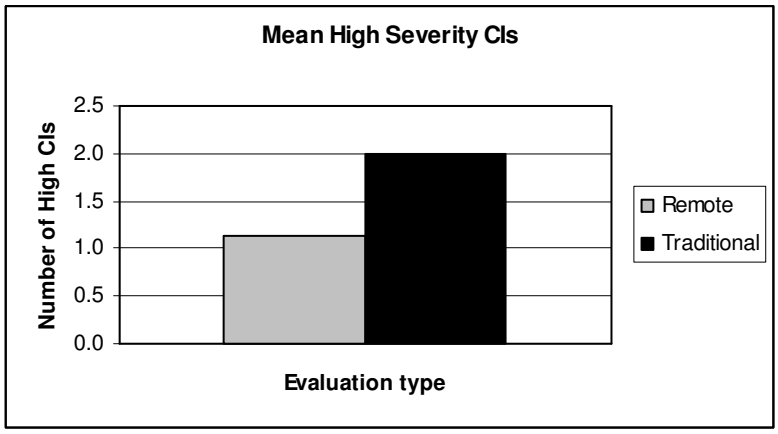

Figure 3: High severity negative CIs results

5.2. Hypothesis 2

Hypothesis-2 was that both novice and experienced users would prefer the remote evaluation approach of reporting to the traditional in-lab evaluation approach. The following dependent measures were used to test this hypothesis

- Preference
- Subjective ratings (overall, comfort, ease, realness, convenience and future participation)

5.2.1 Preference

Two 2x2 Chi-Square tests were performed to determine if future preference was independent of user type and evaluation type. As the expected counts for the Chi-square test

were low, Fisher's exact test was performed as an alternative test on the same data. The results of the tests are shown in Table 8.

Table 8: Test Results for Preference Data

2 x 2 Table	Results of χ^2 test for Independence	Fisher's	Interpretation
User type by Preference	χ^2 (1, N=32) = 0.183, p = 0.669	p = 1.000	No relationship.
Evaluation type by Preference	χ^2 (1, N=32) = 8.960, p = 0.003	p = 0.007 *	Significant difference.

* $p < 0.05$

The results for user type by preference indicate that preference was not dependent on user type. However, results of evaluation type by preference show that user preference was dependent on evaluation type (p = 0.007, Fishers exact test). According to the observed and expected values shown in Table 9, participants preferred to participate in the future, in the same evaluation type as the one they participated previously.

Table 9: Table of Preference vs. Evaluation type

Evaluation type	Preference	Expected	Observed
Remote	Remote	12.5	16
	Traditional	3.5	0
Traditional	Remote	12.5	9
	Traditional	3.5	7
	TOTAL	32	32

Figure 4 provides a graphical representation of the preference data for both the approaches.

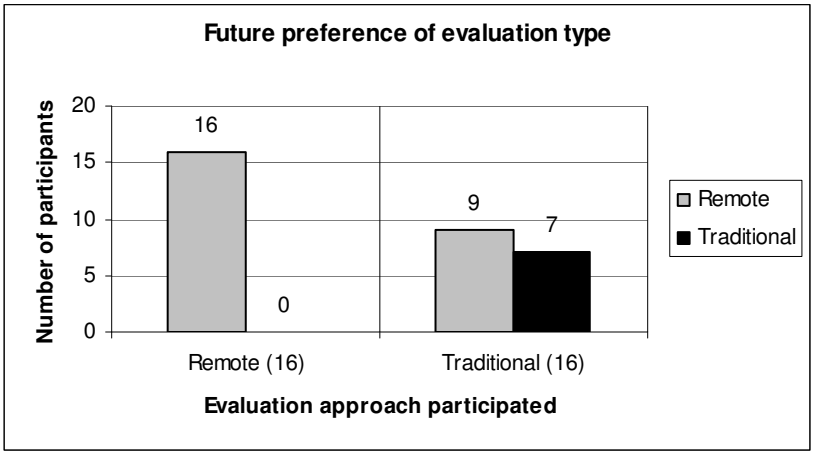

Figure 4: Future preference of evaluation type

5.2.2 Subjective ratings

Overall Rating

A Wilcoxon-Mann-Whitney test was conducted to test for significant differences between the ratings received by each evaluation method. The ordinal data obtained were the overall rating provided by the participants on their session experience. The participants where asked to rate the evaluation method they participated in on a scale of 1 to 10 with one being poor and ten excellent. The results indicated that there was no significant difference between the two evaluation approaches ($Z = 0.775$, $p = 0.44$).

Additional subjective ratings

Beyond preference, additional subjective feedback were collected and analyzed. Data included ratings on five variables - comfort, ease, realness, convenience and future participation. The participants where asked to rate each variable on a five point scale – with a rating of one for strongly disagree and a rating of 5 for strongly agree. Each point or potential response where

assigned a number, starting from 1 for the lowest and 5 for the highest. A sample item with the rating item is shown in Figure 5.

2. Overall, I felt comfortable participating in the session.

Strongly Disagree Disagree Neither Disagree nor Agree Agree StronglyAgree

3. Participating in the session was easy.

Strongly Disagree Disagree Neither Disagree nor Agree Agree Strongly Agree

Figure 5: Sample items from subjective questionnaire

Additionally, the rating questions of the questionnaire were tested for internal consistency reliability by computing the Cronbach's coefficient alpha for a multiple-item rating scale. The calculated Cronbach's standardized alpha was found to be 0.803. An alpha coefficient of 0.70 or greater is considered to be an acceptable reliability coefficient (Nunnally, 1978). The Cronbach's coefficient alpha computed for the questionnaire used in this study is 0.803, indicating that it exceeds the threshold needed. In brief, the rating scale used in the study was reliable. A complete copy of the questionnaire can be found in Appendix C.

The rating scores from the subjective feedback were analyzed in a repeated measure ANOVA with the rating variables (comfort vs. ease vs. realness vs. convenience vs. future participation) as a within-subject factor. The main effect of rating variables was significant, F (4, 112) = 4.96, p = 0.001. Post-hoc comparisons were performed using the Bonferroni adjustment for multiple comparisons. The Bonferroni adjustment was computed to set an appropriate alpha level for the multiple comparisons. According to the Bonferroni adjustment, in order to keep the experiment wise error rate to a specified level (0.05), we need to divide the acceptable alpha level by the number of comparisons. As the number of post-hoc pairwise comparisons was ten

the new alpha level was set at 0.005 (original alpha level/number of comparisons). The significant main effect was analyzed by single degree of freedom, "repeated" contrasts. The results of contrasts indicate that there was a significant difference between the rating scores of *ease* and *realness* in the remote approach, F (1, 28) = 11.67, p = 0.004. Participants in the remote group gave significantly higher ratings for *ease* (M = 4.562, SD = 0.51) than for *realness* (M = 4.125, SD = 0.50). Figure 6 provides a graphical representation of the means for both the variables.

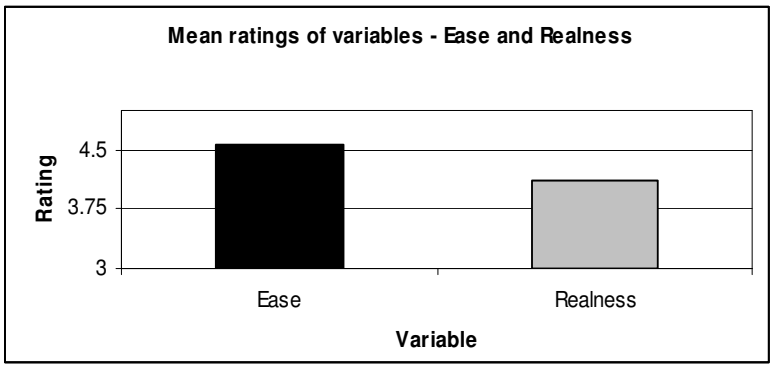

Figure 6: User ratings for variables ease and realness – remote approach

5.3. Hypothesis 3

Hypothesis-3 was that the remote approach will be more effective in collecting usability data with the novice user group compared to the traditional approach with the same user group. Two dependent measures were used to assess this hypothesis

- Participation time
- Critical incident data (total, positive, negative, high severity, medium severity, and low severity).

5.3.1 Participation time

Participation time values of the novice group were subjected to a one-way analysis of variance, as shown in Table 10.

Table 10: ANOVA for Evaluation Type (E) x User type (U) – Novice Users

Source	df
Evaluation Type (E)	1
Subjects (S)	14
Total	**15**

It was found that participation time of novice users in the remote evaluation method (M = 573.50, SD = 244.18) and the participation time of novice users in the traditional evaluation method (M = 508.50, SD = 183.78) were not significantly different, F (1, 14) = 0.36, p = 0.557.

5.3.2 Critical incident data

The critical incident data (frequency count data) of novice users were subjected to the Chi-Square test for independence between the two evaluation methods. However, as the expected counts for the Chi-square test were low, Fisher's exact test was performed as an alternative test on the same data. The results of both the tests for all the critical incident data for novice users are summarized in Table 11.

Table 11: Test Results for Critical Incident Data (Novice Users)

Critical Incident Data – Number of	Results of		Interpretation
	χ^2 test for Independence	Fishers	
Total CIs	χ^2 (6, N=16) = 6.200, p = 0.401	P = 0.620	Not significant
Negative CIs	χ^2 (5, N=16) = 1.867, p = 0.867	P = 1.000	Not significant
Positive CIs	χ^2 (5, N=16) = 8.000, p = 0.156	P = 0.095	Not significant
High severity negative CIs	χ^2 (3, N=16) = 7.619, p = 0.055	P = 0.032 *	Significant
Medium severity negative CIs	χ^2 (5, N=16) = 6.667, p = 0.247	P = 0.284	Not significant
Low severity negative CIs	χ^2 (4, N=16) = 3.200, p = 0.525	P = 0.739	Not significant

* $p < 0.05$

As shown in Table 11, the relation between the evaluation type and number of high severity negative critical incidents collected among novice users was significant, $p = 0.032$ (Fishers test). Novice users in the traditional evaluation approach were more likely to provide higher number of high severity negative critical incidents ($M = 2.375$, $SD = 0.916$) than novice users in the remote evaluation approach ($M = 1.125$, $SD = 0.641$). Figure 7 provides a graphical representation of the means for both the approaches. Additionally, a detailed table of mean and standard deviation of the number of critical incidents collected for each group can be found in Appendix E.

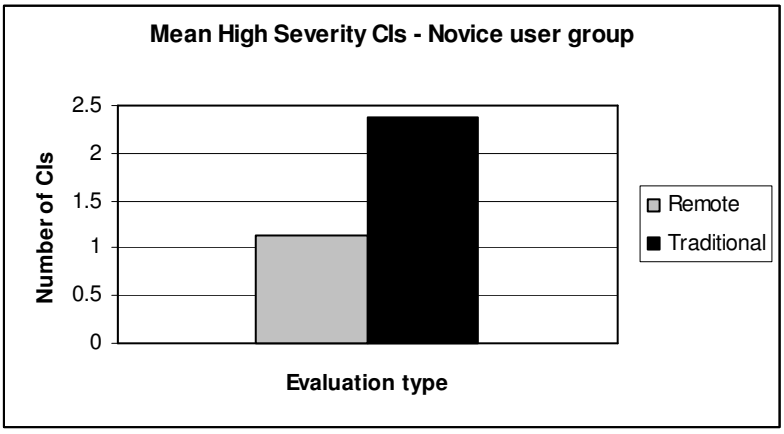

Figure 7: High severity negative CIs results – Novice group.

Examples of frequently occurring critical incidents (pertaining to the web application) for all evaluation type and user groups and the frequency counts of the critical incidents are listed in Table 12. A complete list of all critical incidents can be found in Appendix G.

Table 12: Samples of most Frequently Occurring Critical Incidents

Type of Critical Incidents	Sample Critical Incidents	Evaluation type	User type	Frequency	
High severity negative CIs	Users expect a "film by actors" or actor's section link in the home page and spent a lot of time looking for it.	Remote	E	4	5
			N	1	
		Traditional	E	3	7
			N	4	
	Users complained that the home page of the website was too cluttered and had lot of textual information.	Remote	E	3	6
			N	3	
		Traditional	E	2	6
			N	4	
Medium severity negative CIs	Zip code dialog box on the left in the showtimes page was not functioning, it just repopulated the screen.	Remote	E	3	8
			N	5	
		Traditional	E	2	3
			N	1	
	The Map shown in the showtimes page was unusable.	Remote	E	2	5
			N	3	
		Traditional	E	3	4
			N	1	
Low Severity negative CIs	Actor Al Pacino's filmography page was not well laid out. Difficult to scan and find the required information.	Remote	E	0	3
			N	3	
		Traditional	E	5	5
			N	0	
	Annoying pop-ups.	Remote	E	5	11
			N	6	
		Traditional	E	3	6
			N	3	
Positive CIs	Users found the 'Top movies' link in the homepage to be intuitive.	Remote	E	7	11
			N	4	
		Traditional	E	4	9
			N	5	
	Users thought that the Search in the home page was very good.	Remote	E	2	7
			N	5	
		Traditional	E	3	4
			N	1	

* E = Experienced, N=Novice.

5.4 Participant comments

Samples of selected positive and negative comments provided by the participants, on the evaluation approach they participated, in their post-test questionnaire and verbal think aloud are listed in Table 13 & Table 14. A complete copy of all comments can be found in Appendix F.

Table 13: Positive Comments on the Evaluation Method

Evaluation type	Positive Comments	Participant code
Traditional	Very hands on, very realistic tasks.	N2
	Evaluator not sitting next to you, therefore making you less nervous.	N3
	Very Objective. Evaluation is done in hard numbers.	N5
	Communication was fast and if I could not figure out something I could get answers fast from experimenter.	E2
	Help is available always.	E5
	Like the idea of retrospective think aloud – talking afterwards with visual cues.	E6
	Well set-up lab. Good explanation of tasks in questionnaire.	E7
Remote	Common Tasks. The user could do the evaluation from his/her office.	N1
	Ease and comfort in participating	N2
	Nice cool short experiment. Remote testing is a nice idea & very comfortable.	N4
	Free of evaluator's intrusion. In your own comfortable environment.	N6
	Very convenient & short (time-wise). Instructions were clear.	N7
	Realistic evaluation of the user. Thorough evaluation.	E4
	New & innovative way. No extra setup required.	E5
	Ease of scheduling. Non-intrusive.	E6
	Cost is less, as there is no need for co-location. Comfort of working from my workplace (familiar settings)	E8

* N-Novice, E- Experienced

Table 14: Negative Comments on the Evaluation Method

Evaluation type	Negative Comments	Participant code
Traditional	Not convenient.	N1
	Even though I am aware that evaluator is not evaluating me. I still felt conscious that evaluator is sitting next room and watching. I felt I would something wrong and laugh.	N3
	Making Appointment	N4
	Not my usual OS	E1
	Spending time on commute. It took me more time to come here than do the tasks.	E2
	I have to leave my apartment.	E3
	Requires co-location of evaluator and participant. can be difficult with large distances between them.	E6
	Need to visit a different place.	E7
	Somewhat awkward being in an unfamiliar room.	E8
Remote	Remote usage is a little slow and might frustrate the user.	N1
	Feeling out of contact. Not having someone in the same room.	N3
	Use of a third party website (web-ex). A better way is to use VNC.	N4
	Time delay	N5
	The computer was slow and that irritated me.	E2
	Delay of technology.	E1
	Speed of scrolling.	E5
	Speed of connection. Nee to use phone when help is required	E6

* N-Novice, E- Experienced

6. DISCUSSION

6.1 Answering the Research Questions

6.1.1 Research question 1

The first research question asked whether there was any difference in the effectiveness of the two evaluation types, the remote evaluation approach and the traditional evaluation approach, in collecting usability data. The hypothesis for this question was that the remote approach would be at least as effective as the traditional approach in collecting usability data.

The results showed that although it may have been approaching significance for the high severity data collected, participants in the traditional approach did not perform significantly better than participants in the remote approach at a 0.05 level. Overall, results provided evidence supporting the hypothesis.

6.1.2 Research question 2

The second research question asked, which evaluation approach is preferred by different kinds of users (novice, experienced) for future participation. The hypothesis for this question was that both novice and experienced users would prefer the remote evaluation approach of reporting to the traditional in-lab evaluation approach.

The results indicated that there was no significant relationship between the user types and their preference. However, when the variables, evaluation method and preference, were tested for a significant relationship, the Fishers exact test results revealed that participants tended to incline towards participating in the same evaluation approach as the one they participated previously.

6.1.3 Research question 3

The third question asked, whether there was any difference in the effectiveness of the two evaluation types, the remote evaluation approach and the traditional evaluation approach, for the novice user group. The hypothesis for this question was that the remote approach would be more effective in collecting usability data with the novice user group compared to the traditional approach with the same user group.

The results indicated that there were no significant differences between the evaluation approaches when it came to collecting general critical incidents data. However, in collecting high severity negative critical incident data there was a significant difference between the two evaluation approaches. The novice participants in the traditional evaluation approach tended to provide a higher number ($M = 2.375$, $SD = 0.916$) of high severity negative critical incidents than the novices in the remote evaluation approach ($M = 1.125$, $SD = 0.641$).

6.2 Implications of this Research

Initially, significant results from this study revealed that the remote evaluation approach is as effective in collecting usability data as the traditional lab evaluation approach. Results show that there was no significant difference across the evaluation methods in the number of critical incidents collected and the severity of the critical incidents collected across both, novice and experienced users. However, further examination of the approaches for the novice user group alone produced some interesting results. Although there was no significant difference between the evaluation approaches in the total number of critical incidents collected, there was a significant difference between the methods when it came to the number of high negative severity critical incidents collected. The novice user group in the traditional approach provided more high

severity negative critical incidents than their counterparts in the remote evaluation approach. Apart from the significant results obtained for the novice user group, the results for the experienced user group showed that it was approaching significance in collecting higher negative critical incidents using the traditional approach than the remote approach. Hence based on these results it could be said that the traditional evaluation approach tends to be more suitable for collecting high severity negative critical incidents.

Another interesting piece of information seemed to emerge when the subjective feedback on future preference was analyzed. Results showed that participants preferred to participate in the future, in the same approach as the one they participated in initially. Additionally, though the hypothesis that both types of users would prefer the remote approach in the future was not supported, it would be interesting to note that out of the 16 participants who participated in the traditional approach 7 wanted to migrate to the remote approach in the future while all 16 participants of the remote approach preferred to stay with same approach for future participation.

Additional analyses of subjective ratings feedback on five variables (comfort, ease, realness, convenience, and future) revealed that participants in the remote approach rated the ease of the approach higher than the realness of the approach. Also, though the results of the data analysis on overall rating of evaluation methods showed no significant effects, the data on further scrutiny revealed that the range of the rating scale might not have been sensitive enough to collect user feedback on their overall experience in participating in the evaluation approach. In this study the evaluator used a 10 point numerical scale for rating the overall experience attribute and the subsequent ratings collected from most participants was in the high end between 8 and 10. The author recommends that future studies employ alternative scales, such as a descriptive

rating scale with clear descriptive phrases, rather than a numerical rating scale to see if they yield any significant results.

6.3 Recommendations

Based on the results of this study, user feedback, informal observations by the evaluator and the evaluator's own experience in setting up the study, a list of recommendations have been drafted for the benefit of future usability evaluators considering the SREM approach for their usability evaluations.

1. As with any remote evaluation approach it would be ideal to employ real users of interest in order to collect real world representative data.
2. Instruct participants to take part in this study from their work environment. According to Bloomberg et al. (1993), it is necessary to adopt a holistic view that behavior can only be understood in the context in which it occurs. Also such ethnographic approaches need to be in the natural work setting of the user inorder for it to be a true representation of the users point of view. Additionally, some of the participants in the traditional evlauation approach commented on how conscious they were in participating in the study and were constsantly aware of the evaluator's presence. Hence in order to gather more relevant usability data based on real world usage it is necessary for users to participate from their natural work surroundings.
3. As with any usability study, use tasks that are developed from real world scenarios in order to collect user representative data.
4. As participants are at their workplace, ensure that test sessions are short and do not consume much of their valuable time. Based on the evaluator's conversations with the participants and user feedback in the subjective questionnaire, it was obvious that most

participants appreciated the short time span of the study. Hence it is necessary to keep the number of tasks to a minimal reasonable number within which you would be able to test the various issues of your study or evaluation. Four tasks were used in this study and the range of participation time for a complete session was between 30 to 45 minutes.

5. Though there was no significant difference between the evaluation approaches with respect to participation time, one consistent feedback from most users in the remote approach was the delay in computer response due to network latency. Network connection speed plays a major role in most synchronous remote approaches hence it is important to conduct tests over a high speed Internet connection to minimize delay.
6. Use web conferencing tools that do not require users' to download and install any major files. Users tend to be apprehensive of downloading any software. In this study, when the remote participants were requested to setup their machines for remote participation almost every user had concerns about downloading any software to their machines. As Webex, the web conferencing tool used in this study did not require any major downloads or changes to their computer's registry they were more conducive to the setup process.
7. Also, users in general do not like to share their machine. Hence, use the same procedure as in this study, upload the test application in the evaluator's machine and share the screen with the user by giving them remote access. This also would help companies test products in development stages (beta products) without distributing it.
8. Finally, almost all remote users were conducive to the idea of remote participation and showed lot of interest in the evaluation process and in future participation. Hence it would be ideal to advertise the novelty of the remote approach to entice more users to participate in your evaluation.

7. CONCLUSION

SREM (the Synchronous Remote Evaluation Method) is a remote adaptation of the traditional in-lab usability evaluation method. This method employs real users, participating from their own natural work environment, to participate in a usability evaluation through the Internet. The test evaluator monitors the participants through the shared screen of the web conferencing software (WebEx ^{TM}) and collects the necessary usability data similar to a traditional in-lab evaluation method.

The overall purpose of this study was to compare the SREM to the traditional in-lab evaluation method for its effectiveness in collecting critical incident data. Effectiveness was measured by comparing the frequency and severity of critical incidents data collected by the remote approach to the same data collected using the traditional approach. The same web application was used for testing the two approaches.

Generally, the results showed that participants in the remote approach did not differ significantly from the traditional approach participants in the number of critical incidents collected. Also, no significant results were obtained between the evaluation approaches when it came to collecting negative and positive critical incidents. Based on the results of the critical incident data analyses in this study, remote testing in general was as effective in collecting usability data as the traditional lab evaluation approach. This outcome indicating that remote testing in general yields similar results to that of traditional lab testing supports earlier studies by Hartson et al. (1996) and Tullis et al. (2002). However, when user type was taken into consideration interesting results where obtained. It was the objective of the evaluator to additionally investigate the effectiveness of the approaches based on the type of users, usability novice and usability experienced, participating in the evaluations. The results showed that the

novice user group in the traditional in-lab approach provided more high severity negative critical incidents than the novice group in the remote evaluation approach. This trend could be attributed to the theory that novice users in the traditional approach were more aware of their responsibility as participants in a study and were more critical in their critiquing. According to Martin (2000), participants tend to be concerned with how they will be judged by the experimenter and hence try to perform well in the evaluation and try making themselves look good. Additionally, though this research did not significantly support the idea that users of both the approaches, remote and traditional, would prefer the remote approach to the traditional approach in the future, it did reveal that participants in the remote evaluation approach unanimously preferred to participate in the same approach again while more than half of their counterparts in the traditional approach preferred to migrate to the remote approach in the future.

Although no cost data was collected in this study, the many cost benefits of the SREM approach were apparent. First and foremost the results showed that there was no difference in participation time between the approaches and there was no need for the remote participants to travel. Hence, travel costs usually associated with traditional in-lab evaluation are avoided in the SREM approach. Secondly, as all remote participants participated from their work environment it was not necessary for them to take time off from work. Hence, the user does not incur any monetary loss due to time off from work. Finally, monetary savings due to the use of Webex ™ web conferencing tool for remote participation. Webex ™, unlike other network conferencing software, is an online service and there is no need for each participant to download the software. This potentially saves costs associated with multiple licenses for a remote conferencing tool.

According to Hartson et al. (1996), barriers such as limited access to local representative users and financial costs of transporting users have led to the development of remote approaches

to usability evaluation. As the SREM approach is a remote approach it can be used as a suitable alternative to the traditional in-lab evaluation in such situations. Additionally, software companies of the present generation are increasingly catering to varied and distributed international customers. For example, Hammontree et al. (1994) reported that several Hewlett Packard development teams in the United States produce custom software for customers located in Europe, Australia, and Asia-Pacific. As traditional usability evaluation with representative users would not be feasible and cost effective, it is believed by the evaluator that the SREM approach, due to its remote setting, would be an ideal choice for conducting such usability evaluations.

7.1 Future research

Compare SREM against other remote evaluation approaches

In this study the SREM approach was compared with the established traditional in-lab evaluation approach. In the future, SREM approach can be compared with various other remote usability approaches. Though there have been comparisons of remote approaches individually against the traditional approach, there are not enough studies comparing the various remote approaches against each other. Studies where measures like cost effectiveness, development needs, data type collected and other facility requirements could be collected and evaluated. Such studies could provide usability evaluators with valuable information based on which they would be able to choose a suitable and financially viable remote approach for their usability needs in the future.

SREM approach for different software applications and usability techniques.

A major area for future research can be the application of the SREM approach for evaluation of different software applications other than websites. In this study, a movie website was selected for evaluation because of its applicability to a wide user group. However, the SREM approach need not be limited to websites and should be tested for evaluating various other applications that are built on different platforms. It can also be used for formative evaluations of software applications that are in various development stages.

In addition, it would also be interesting to explore if the SREM approach can be adopted for other usability techniques apart from usability testing (e.g. cognitive and design walk throughs). Results of these studies would help in tweaking the SREM approach and also marketing it, as a suitable remote approach for all usability needs.

Investigate new technologies in web conferencing and remote access.

The basic technologies needed for remote testing are available in varying degrees in most platforms (Hammontree et al., 1994). In this study the basic equipment used was a personal computer with a high bandwidth internet connection and an IE browser (version 6). The software tool used for web conferencing between the evaluator and the participant computers was Webex ™. This web conferencing tool allows users to participate in the study, without downloading and installing any software to their computers, by using the Webex website for establishing connection between the user and the evaluator's computer. Additionally, TechSmith's Camtasia software was used for screen capture of participants' on screen activities and their verbal comments in the retrospective think aloud session. Presently, there are many similar technologies that are used for various remote approaches. However, new technologies and tools with more

features are added to the existing list from time to time. Hence it is imperative to keep up with the changes in technology and investigate these new technologies in order to enhance remote evaluation methods. Future advancements in development of high network bandwidth and web conferencing techniques will aid in designing a more robust SREM.

REFERENCES

Andre, T. S., Williges, R. C., & Hartson, H. R. (1999). Effectiveness of Usability Evaluation Methods: Determining the Appropriate Criteria. *In Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting, September 27-October 1,* (1090-1094). Santa Monica, CA: Human Factors and Ergonomics Society.

Blomberg, J., Giacomi, J., Mosher, A., and Swenton-Wall, P. Ethnographic field methods and their relation to design. *In D. Schuler and A. Namioka (Eds.) Participatory Design: Principles and Practices*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1993, pp. 123-154.

Castillo, J. C. (1997). *The User-reported Critical Incident Method For Remote Usability Evaluation.*Unpublished thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA.

del Galdo, E. M., Williges, R. C., Williges, B. H., and Wixon, D. R. (1987). A critical incident evaluation tool for software documentation. *In L. S. Mark, J. S. Warm, and R. L. Huston (Eds.) Ergonomics and Human Factors*. New York: Springer-Verlag. 253-258.

Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* Revised edition. Cambridge, MA: The MIT Press.

Fitts, P.M. and Jones, R.E. (1947). Psychological aspects of instrumented display: Analysis of 270 "Pilot-Error" experiences in reading and interpreting aircraft instruments. *In Sinaiko, H.W. (Ed), Selected Papers on human factors in the design and use of control system*. New York: Dover Publications, Inc. 1-38.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin,* 51, 327-358.

Hammontree, M. J., Weiler, P., & Nayak, N. (1994). Remote Usability Testing. *Interactions*, 1(3), 21-25.

Hartson, H.R., Castillo, J.C., Kelso, J., Kamler, J., & Neale, W.C. (1996). Remote Evaluation: The network as an extension of the usability laboratory. *In Proceedings of CHI'96 Human Factors in Computing Systems*. 228-235.

Hartson, H. R., & Castillo, J. C. (1998). Remote evaluation for post-deployment usability improvement. *In Proceedings of the Working Conference on Advanced Visual Interfaces (AVI '98), May,* (22-29). New York: ACM Press.

Hix, D. & Hartson, H. R. (1993). *Developing user interfaces: Ensuring usability through product and process.* New York, NY: John Wiley & Sons.

Martin, D. W. (2000). *Doing Psychology Experiments.* Belmont, CA: Wadsworth/ Thompson learning.

Nunnaly, J. (1978). *Psychometric theory*. New York: McGraw-Hill.

Rowley, D. E. (1994). Usability testing in the field: bringing the laboratory to the user. *In Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence*, 252-257.

Rubin, J. (1994). *Handbook of Usability Testing*. New York: John Wiley & Sons

Shattuck, L. G., & Woods, D. D. (1994). Critical incident technique: 40 years later. *In Proceedings of the Human Factors and Ergonomics Society* 38th Annual Meeting, (1080-1084). Santa Monica, CA: Human Factors and Ergonomics Society.

Siochi, A. C., & Ehrich, R. W. (1991). Computer Analysis of User Interfaces Based on Repetition in Transcripts of User Sessions. *10 ACM Transactions on Information Systems*. 9, 4 (October 1991), 309-335.

Smith-Jackson, T.L., & Williges, R.C. (2001). User-Centered Design for Telesupport Systems, *Assistive Technology*, 13, 144 – 169.

Swallow, J., Hameluck, D., and Carey, T. (1998). User Interface Instrumentation for Usability Analysis: A Case Study. HCI + Learning Laboratory, University of Waterloo (March) *Previously published at Cascon '97* (Toronto, Ontario) Nov, 1997.

Thompson, J. A. (1999). *Investigating the Effectiveness of Applying the Critical Incident Technique to Remote Usability Evaluation*. Unpublished thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA.

Tullis, T., Fleischman, S., McNulty, M., Cianchette, C., & Bergel, M., (2002). An Empirical Comparison of Lab and Remote Usability Testing of Web Sites. *Presented at the Usability Professionals Association Conference,* July 2002, Orlando, FL.

Whiteside, J., Bennett, J., & Holtzblatt, K. (1988). Usability Engineering: Our experience and evolution. *In* Helander*, M. (Ed.), Handbook of human-computer interaction*, Amsterdam, The Netherlands: North-Holland. 791-817.

Wixon, D. R., Ramey, J., Holtzblatt, K., Beyer, H., Hackos, J., Rosenbaum, S., Page, C., Laakso, S. A., Laakso, K.P. (2002). Usability in practice: Field methods evolution and revolution. *In Proceedings of CHI '02 extended abstracts on Human factors in computer systems.*

APPENDIX A: INFORMED CONSENT FORM

Virginia Polytechnic Institute and State University

Informed Consent for Participants of Investigative Projects

Title of Research Project: A Comparative Study of a Synchronous Remote Usability Evaluation Method to a Traditional In-Lab Usability Evaluation

Investigators:

Prakaash Selvaraj, Dr.Tonya Smith Jackson

Phone No. (Office): (540) 231 4119

Location of Investigation: Blacksburg, Virginia

Purpose of Study:

This study is being conducted to compare the synchronous remote usability evaluation method with the traditional in-lab usability evaluation method and measure the effectiveness of the remote method in collecting usability data.

Principal investigators

Prakaash Selvaraj

Dr. Tonya L. Smith Jackson, PhD

Procedure

If you decide to participate in this study, first you will be asked to fill out a questionnaire on your background and experience in using computers. The test administrator will then assign you to either the traditional lab user group or the remote user group.

Remote Users: If you are a part of the remote user group you will be asked to connect to the experimenter's computer through the internet from your computer at your office at a particular time. Instructions will be provided as to how to establish connection with the experimenter's computer. Upon establishing connection you will be asked to perform a set of representative tasks using the website that is being evaluated. Video and audio data of your actions will be collected. You may be encouraged to think out aloud and asked questions to

better understand your evaluation. The session will last for about 30 minutes at the most. At the end of this session you will answer a questionnaire to help evaluate the effectiveness of the evaluation approach. A follow up session (max 30 minutes) for retrospective think aloud will be scheduled in which the investigator sits with you and goes through video data of the previous session and elicits more information to better understand your evaluation.

In-lab Users: If you are a part of the in-lab user group you will be asked to report to the usability lab at Virginia Tech at a scheduled time. You will then be asked to complete a set of representative tasks using the website that is being evaluated. Video and audio data of your actions will be collected. You may be encouraged to think out aloud and asked questions to better understand your evaluation. The session will last for about 30 minutes at the most. At the end of this session you will answer a questionnaire to help evaluate the effectiveness of the evaluation approach. A follow up session (max 30 minutes) for retrospective think aloud will be scheduled in which the investigator sits with you and goes through video data of the previous session and elicits more information to better understand your evaluation.

Risks

This study has no risks associated with it apart from the regular risks associated with daily personal use of computers.

Benefits

Your participation in this project will provide information that may be used to improve the usability and effectiveness of the synchronous remote usability evaluation approach. Apart from the monetary payment for participating in the study there is no other guarantee or promises to encourage you to participate. If you wish to receive a copy of the findings of this research study please leave a self-addressed envelope with the experimenter.

Extent of Anonymity and Confidentiality

Any published data will not reveal your identity. You will be assigned a Participant Identification Number so that we may keep all data collected confidential. If you decide to participate, you are free to withdraw your consent at any time without penalty. If you have any questions, we expect you to ask via email or telephone. The test administrators will be happy to

answer them. Your electronic or real signature indicates that you have read and understand the above information, that your questions have been answered to your satisfaction, and that you have decided to participate based on the information provided on this form.

Approval of Research

This research has been approved, as required, by the Institutional Review Board for projects involving human subjects at Virginia Polytechnic Institute and State University, and by the Department of Industrial and Systems Engineering.

Freedom to Withdraw

You are free to withdraw from this study at any time for any reason without any penalty.

Participants Responsibilities:

I voluntarily agree to participate in this study. I have the following responsibilities:

- I should not volunteer for participation in this study if I now know that I will not be able to complete this experiment.
- I will give honest answers to the questions asked and complete the tasks to the best of my abilities.
- Upon completion of the study I will not discuss with any other individual my experiences with this study.

Participant's Permission

I have read and understand the Informed Consent and conditions of this project. I have had all my questions answered. I hereby acknowledge the above and give my voluntary consent for participation in the project and for the videotaping of my session. If I participate, I may withdraw at any time for any reason without penalty.

_____ _____

Signature Date

Should I have any questions about this research or its conduct, I may contact:

Prakaash Selvaraj, Investigator 951-1023

Dr. Tonya Smith-Jackson, Faculty Advisor	231-6270
Dr. David Moore, Chair, IRB Research Division	231-4991

APPENDIX B: PRE-TEST BACKGROUND QUESTIONNAIRE

GENERAL

Name:

Age:

Gender:

Do you have an office/workplace: □ Yes □ No

If 'yes' answer the following:

Do you have a computer with internet access at work: □ Yes □ No
Do you have a telephone access at work: □ Yes □ No

EDUCATION

1. Please check your academic level below?

□ Undergraduate student
□ Graduate student (Masters or Ph.D.)
□ Other (please specify _____)

2. List your Major area of study: _____

COMPUTER EXPERIENCE

3. How long have you been using computers?

□ Less than a year □ 1- 3 years □ 3- 5 years □ More than 5 years.

4. List the operating systems you are familiar with. (ex: Win (95/98/NT/2000) / Mac (OS 6-9))

5. How would you rate your experience with Windows (95/98/NT/2000) computers?

□ Very experienced □ Moderate experience □ Minimal experience □ None

6. How would you rate your experience with Macintosh computers?

□ Very experienced □ Moderate experience □ Minimal experience □ None

7. Please check mark the Internet browsers you are familiar with

□ MS Internet Explorer □ Netscape Navigator

□ AOL □ Other (please specify _____)

8. How often do you use MS Internet Explorer?

□ Every day □ Sometimes □ Rarely □ Never.

9. How would you rate your level of expertise in using MS Internet Explorer?

□ Very experienced □ Moderate experience □ Minimal experience □ None

USABILITY EXPERIENCE

10. Have you taken any university level courses in the field of Human Computer Interaction, Usability evaluation or Human Factors? If yes, please list them.

11. Have you participated (or been involved) in any usability evaluations in the past? If yes, please list how many usability evaluations and in what roles you participated?

APPENDIX C: SUBJECTIVE QUESTIONNAIRE

Instructions: Please fill in the appropriate information or circle the item that most accurately depicts your opinion.

1. How would you rate your experience in participating in the session?

Poor Excellent

0 1 2 3 4 5 6 7 8 9 10

2. Overall, I felt comfortable participating in the session.

Strongly Disagree Disagree Neither Disagree nor Agree Agree StronglyAgree

3. Participating in the session was easy.

Strongly Disagree Disagree Neither Disagree nor Agree Agree Strongly Agree

4. I found participating in such a study to be very realistic and non- intrusive.

Strongly Disagree Disagree Neither Disagree nor Agree Agree Strongly Agree

5. It was very convenient for me to participate in this session.

Strongly Disagree Disagree Neither Disagree nor Agree Agree Strongly Agree

6. I would be interested in participating in similar evaluation sessions in the future.

Strongly Disagree Disagree Neither Disagree nor Agree Agree Strongly Agree

7. List the most positive aspects of this evaluation method.

8. List the most negative aspects of this evaluation method

Please read the description of the alternative usability evaluation approach (that you did not participate in) below and then answer the questions that follow.

Traditional in-lab evaluation: In this approach the participant and the evaluator are located in the same location, in a usability lab. A usability lab basically consists of two rooms, the participant room and the evaluator room. The evaluator's room is separated from the participant side by a one-way mirror and is equipped with a viewing and recording facility. In a typical usability session, similar to the remote approach, participants would be given a set of tasks to complete using the computer in their room. The evaluator's role is to monitor the participant's actions on screen and in person through the one-way mirror and take down notes. A descriptive layout of a traditional usability lab is shown below for your viewing.

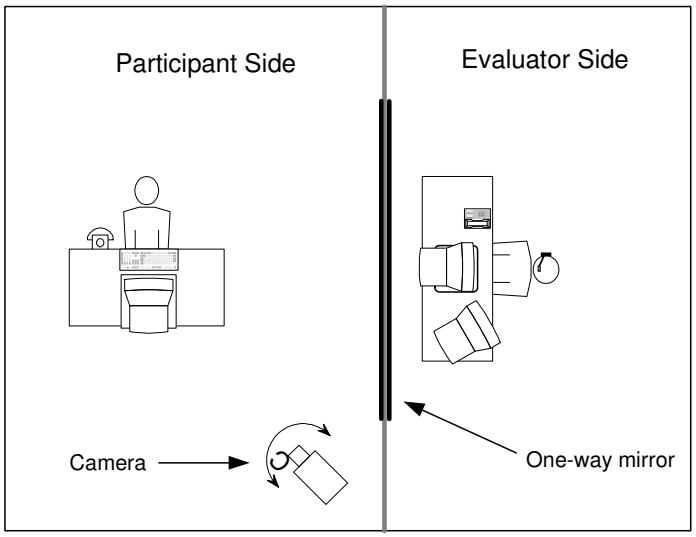

Figure 8: Schematic of a traditional usability lab

Synchronous Remote Evaluation: is a collaborative approach where the test evaluator and the participant are separated in space and/or time. The term remote is relative to the evaluator and refers to the remote location of users from the evaluator's location. The remote user's computer (remote location) is connected to the evaluator's computer (usability lab)

through the Internet using commercially available web conferencing software (e.g., GoToMyPC ™, WebEx ™). As this is a synchronous approach, the user and the evaluator are connected in real time by an established audio connection using a telephone line. Using the share screen feature of the web conferencing software the remote user will be able to view and use the website in the evaluator's computer. In a typical usability session, similar to the traditional lab approach, participants would be given a set of tasks to complete. The evaluator monitors the user's actions via the computer in the usability lab and video records the screen actions using a scan converter. A descriptive layout of a remote evaluation set up is shown below for your viewing.

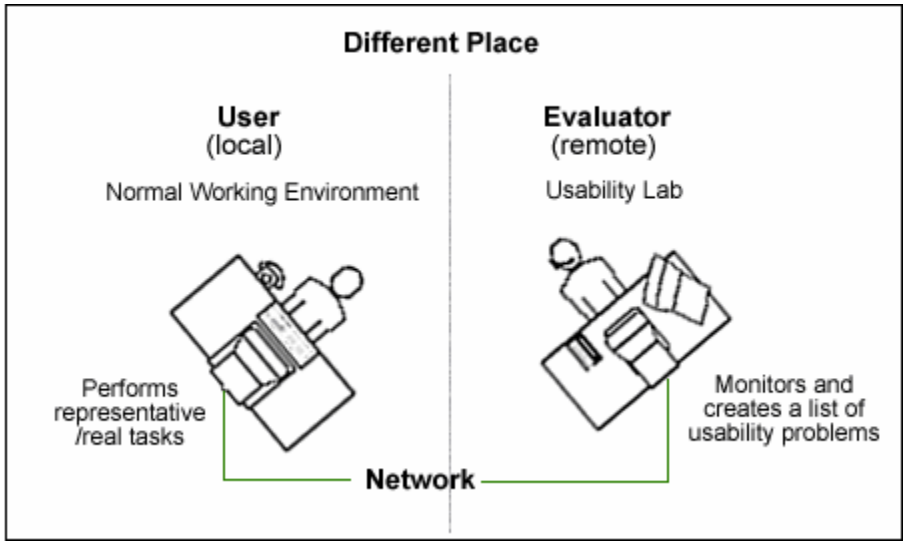

Figure 9: Schematic of a remote evaluation set up

Please answer the following question:

Please check the evaluation method you would prefer to participate in the future?

□ Remote Evaluation Method

□ Traditional in-lab Evaluation Method

APPENDIX D: PERFORMANCE TEST

Instructions:

Please complete the following tasks to the best of your capabilities. Also please remember that the purpose of this evaluation is to test the usability of the IMDB website and in no way is this a test of your skills or knowledge. The main focus of the test is to measure the effectiveness of the evaluation approach used for evaluating the usability of the website.

Feel free to browse around the website to gain familiarity before you begin the tasks.
Please do not start until the Evaluator asks you to.

TASKS:

1. Find out the one movie Al Pacino acted in 1972 and what was the character (name) he played in that movie?

1. Find out which films Christopher Reeve was in the seventies and find the one film in which Gene Hackman co-stared with him.

2. Find the latest independent film featuring Denzel Washington and find out who directed the film.

3. Find the top ranked movie the IMDB top 250 movie list.

APPENDIX E: MEAN & SD FOR ALL CRITICAL INCIDENT FREQUENCY DATA

Evaluation type	User type	No. of CIs - Mean (SD)					
		Total	Positive	Negative	High	Medium	Low
Remote Evaluation	Novice	8.750 (1.832)	2.375 (1.598)	6.375 (1.408)	1.125 (0.641)	2.625 (1.408)	2.625 (1.408)
	Experienced	9.000 (2.777)	2.875 (1.356)	6.125 (2.850)	1.375 (1.061)	3.125 (2.416)	1.625 (1.061)
	Both	8.875 (2.2770	2.625 (1.455)	6.250 (2.176)	1.250 (0.856)	2.875 (1.928)	2.125 (1.204)
Traditional Evaluation	Novice	8.375 (1.923)	1.875 (1.246)	6.500 (1.852)	2.375 (0.916)	2.125 (1.126)	2.000 (1.690)
	Experienced	8.500 (2.507)	2.375 (1.506)	6.125 (1.885)	1.625 (0.744)	2.375 (1.847)	2.125 (0.641)
	Both	8.438 (2.159)	2.125 (1.360)	6.313 (1.815)	2.000 (0.894)	2.250 (1.483)	2.063 (1.237)

APPENDIX F: POSITIVE AND NEGATIVE USER COMMENTS

F.1 Positive Comments

Evaluation type	Positive Comments	Participant code
Traditional	It is easier to communicate problems and ask questions.	N1
	Very hands on, very realistic tasks.	N2
	Evaluator not sitting next to you therefore makes you less nervous.	N3
	Realistic tasks	N4
	Very Objective. Evaluation is done in hard numbers.	N5
	It was fun.	N6
	Quick	E1
	Communication was fast and if I could not figure out something I could get answers fast from experimenter.	E2
	Have instant feedback from experimenter.	E3
	Help is available always.	E5
	Like the idea of retrospective think aloud – talking afterwards with visual cues.	E6
	Well set-up lab. Good explanation of tasks in questionnaire.	E7
	Personal Communication. Knew the evaluator was there in the next room if there was an issue.	E8
Remote	Common Tasks. The user could do the evaluation from his/her office.	N1
	Ease and comfort in participating	N2
	Easy to fit into a schedule	N3
	Nice cool short experiment. Remote testing is a nice idea & very comfortable.	N4
	Flexibility of accessing information from a remote site.	N5

	Free of evaluator's intrusion. In your own comfortable environment.	N6
	Very convenient & short (time-wise). Instructions were clear.	N7
	No access problem had done it from my office. No problem in understanding task.	N8
	Easy, convenient to participate. Didn't take a lot of time.	E1
	I was at my workplace, was pretty comfortable.	E2
	Workplace, Comfort	E3
	Realistic evaluation of the user. Thorough evaluation.	E4
	New & innovative way. No extra setup required.	E5
	Ease of scheduling. Non-intrusive.	E6
	Novelty, Interesting	E7
	Cost is less, as there is no need for co-location. Comfort of working from my workplace (familiar settings)	E8

N-Novice, E- Experienced

F.2 Negative Comments

Evaluation type	Negative Comments	Participant code
Traditional	Not convenient.	N1
	Might not be a good representation of Overall Usability issues.	N2
	Even though I am aware that evaluator is not evaluating me. I still felt conscious that evaluator is sitting next room and watching. I felt I would something wrong and laugh.	N3
	Making Appointment	N4
	How the questions were created may affect the test results.	N5
	Not my usual OS	E1
	Spending time on commute. It took me more time to come here than do the tasks.	E2
	I have to leave my apartment.	E3
	Requires co-location of evaluator and participant. can be difficult with large distances between them.	E6

	Need to visit a different place.	E7
	Somewhat awkward being in an unfamiliar room.	E8
Remote	Remote usage is a little slow and might frustrate the user.	N1
	Feeling out of contact. Not having someone in the same room.	N3
	Use of a third party website (web-ex). A better way is to use VNC.	N4
	Time delay	N5
	Slow to browse around.	N6
	Delay in computer response.	N7
	Delay of technology.	E1
	The computer was slow and that irritated me.	E2
	System lagging.	E3
	Speed of scrolling.	E5
	Speed of connection. Nee to use phone when help is required	E6
	Lag	E7
	Network Latency. Loss of data that might result from traditional think aloud protocols.	E8

N-Novice, E- Experienced

APPENDIX G: LIST OF CRITICAL INCIDENTS

G.1 High severity negative critical incidents collected from this study

UE type / User group	High Severity Negative Critical Incidents
Remote/ Experienced	Home page - Expected a "film by actor" link on the homepage. Searched for it
	Home -> Browse imdb -> Films&more -> Power search by keywords – Wasn't helpful, did not return movie listing of al pacino. Too much text, did not read the instructions.
	Top movies page - Just went for top grossing section (the wrong section) because user couldn't see the top rated section. It was not visible and at the bottom.
	Home page - Expected a "film by actor" link on the homepage. Looked for it
	Home->Browse->Star & More-> Find by People – Already decided to search by people but still had too many options of searching by movie/title etc in that page - was very confusing
	Home – No actors' category/section in this page….looked for it.
	Home – too much content in the home page -
	Home page - Expected a "film by actor" link/section on the homepage. Looked for it. - couldn't find any.
	Lot of text on the home page – difficult to scan
	Requested unwanted registration – because user had come through 'My imdb' page. User wasn't aware and tried registering and was irritated by this procedure.
	Homepage-Too much content
Remote/ novice	Too cluttered home page
	Home – the two top horizontal menus were confusing, too cluttered and did not realize initially the "us showtime" tab
	Home page –Did not realize that the "Tops at box office" link was only for US. They should have a clear distinction like "Tops at box office in US"
	Home – Too many categories/sections page and was difficult to find info.
	Lot of extraneous info on the home page…it is just too much
	Search criteria "people" was misleading thought it was search for members.
	Home - Left nav is too busy/ lot of links – clicked on wrong link
	Top movies page – Too confusing to find Top 250 info on the page as it was at the bottom. - prefer if it was displayed at the top.

	Home page - Expected a "film by actor" link/section on the homepage. Looked for it.- couldn't find any
Traditional/ Experienced	Went to the wrong list by clicking on the Top grossing link – the link was worded similar to top rated.
	The UK option was not obvious and took awhile to find how to get to the UK site. The UK link was hidden at the bottom. Too much text on homepage. Had to Use IE Find to locate the UK link.
	User did not realize that she went out to another website by clicking on the showtimes link in the homepage. It populated in the same page. The Showtimes tab in Imdb was not distinct enough.
	User did not realize that she went out to another website by clicking on the showtimes link in the homepage. It populated in the same page. The Showtimes tab in Imdb was not distinct enough.
	Home page - Expected an actor/ movie link/section on the homepage. Looked for it.- couldn't find any
	System error – website did not update for the current date and kept saying no timing for the previous day. Confused user a lot.
	Top movies page - Too much text and did not scroll down to see top rated section. – lead to wrong answer
	Too much info – they should design and show only info that is relevant to the user.
	Home page - Expected an actor/ movie link/section on the homepage. Looked for it.- couldn't find any
	Upon clicking on the UK link was not sure if I was at the UK site, no change, so had to go down & see if the link was used
	Home Page – Too crowded
	Box office Table – Totals were shown on two sides that led to wrong answer– prefer if they were grouped together on the same side (right).
	Home page - Expected an actors list on the homepage. Looked for it. - couldn't find any so went to search
Traditional/ novice	Did not see the showtimes tab at the top – think it was because the tab and the background ad were same color-blue
	Expected to find earning information on the movie page – but nothing was listed – also looked into all the links in the left and tried few but couldn't find any descriptive links.
	Clicked on the more link under tops at the box office section (UK site) it gave information about US earnings – expect to see UK info when on UK site.
	Tried searching for the movie first using the homepage search but it did not work (My movie parameter was misleading…also other search parameters are not clear)
	Tried Browse IMDB link thinking that it will provide a movie list index from A-Z. However on clicking on A-Z it did not list movies.

	Lower toolbar in yellow background was more prominent than the upper blue tab menu hence I didn't see showtimes tab in the beginning.
	Top menu blended with the background and was not obvious. The lower menu was more obvious because of f/g b/g color. Could have listed both menus together in one single line instead of cluttering it in 2 rows.
	Home page - Expected a "film by actor" link/section on the homepage. Looked for it.- couldn't find any so went to browse imdb page
	Did not see the showtimes tab at the top – think it was because the tab and the background ad were same color-blue
	Top movies page – Couldn't see the info ('top rated' word) in the initial screen, top part. Hence went to top grossing section which looked similar to top rated.
	Home page - Expected a "film by actor" link/section on the homepage. Looked for it.- couldn't find any
	Home page has too much info - prefer simplicity like google.
	Font –times new roman was not clear – didn't read because of that
	Looked for an actor link or section in the homepage couldn't find any so went to search.
	Too much info on the home page – couldn't figure out properly initially.
	Clicked on the map for Virginia as the map was more prominent as user didn't realize that there was zip code search at the bottom. Map was too crowded and did not take the user to Blacksburg listing, had to type in the zip code for bburg
	Looked for an actor link or section in the homepage couldn't find any so went to search.
	Too cluttered and not well organized.
	US showtimes page – searched for Blacksburg – website expected user to register – but on trying again – it worked.

G.2 Medium severity negative critical incidents collected from this study

UE type / User group	Medium Severity Negative Critical Incidents
Remote/ Experienced	Search did not work
	Assumed that Zip code would yield movies for bburg. Had to type the zip code again
	Too much scrolling – 'Jump to' links would have helped –M
	Home page - Too many links on the left navigation
	Home -> Browse imdb - Here expected a "films by actor" link couldn't find it and went into "Films & more"
	Home -> Browse imdb -> Films&more – Too many links and not clear as to how to search by actor
	The IMDB home icon wasn't consistent with the other tabs and wasn't clear if it is a home link

	Browse IMDB page – Looked for a "films by actor" sub menu under "Stars & More" – couldn't find any- they should have that first
	Map page – Map was not good, unusable. Prefer the Zip code on the top of the page
	Used the back button as the Home icon wasn't obivious
	Search isn't great – tried searching Al Pacino + 1983 for closer match– dint work
	Font size is really small
	US showtimes page – After selecting the movie from the drop down – expected show times for that movie – but had to go click on the movie from a list
	Home – Text on the home page saying "type in your zip code for show times" was misleading – there was no place to type in the zip code and actually have to go to the show times page to do it
	US showtimes page – Map was totally unnecessary
	US showtimes page – Zip code box on left did not work – just repopulated the same page – expected movies for that zip code
	Home - Left Nav is very busy/crowded
	Finding Nemo box office page – The display of box office info was confusing – prefer table layout with better description
	Home – Search - Confused between search parameters – People & Character- expected a parameter called Actor
	Home icon was not a hyper link, too small, not intuitive
	Home – Left nav has too many links – didn't see the UK box office link initially.
	Box office table page – Lot of info and too confusing.
	US showtimes page – Zip code box on left did not work – just repopulated the same page
	US showtimes page – expected showtimes for the movie Radio on selecting it but it asked for zip code again.
	Top Movies Page – reduce scrolling & scanning for various tops by providing overview tabs or jump tos at the top of the screen.
Remote/ novice	Al Pacino page – didn't realize that info was divided into various roles in the same page – prefer links to his various roles (as actor, producer etc) – or prefer separating the info into diff pages – too much content
	Showtimes page – upon entering zip code expected a list of movies for bburg but had to enter zip code again in another page – very redundant
	Top movies Page – Too many tops in this page – prefer different and better description
	Showtimes page – expected movies to show up on entering zip code…had to search for the movie
	Top movies Page – Too many tops in this page
	US showtimes page – Zip code box on left did not work – just repopulated the same page – expected movies for that zip code
	US showtimes page – Map was totally unnecessary- confusing and had to

	type zip again.
	Left nav is too busy/ lot of links – had to scroll a lot to read
	US showtimes page – Two zip codes in the same page – confusing
	US showtimes page – redundant typing of zip code.
	Movie Page – The business & box office link did not have any earning information- it was very irritating- u expect to find box office earnings info here
	US showtimes page – Zip code box on left did not work – just repopulated the same page
	US showtimes page – Map was totally unnecessary- confusing and had to type zip again.
	Home Icon looked like logo, didn't realize it was a link.
	Top movies page - Did not realize that the top movies page had other top sections below. Prefer links to the bottom sections at the top part of the page.
	US showtimes page – Zip code box on left did not work – just repopulated the same page – expected movies for that zip code
	US showtimes page – expected showtimes for the movie on selecting it but it asked for zip code again.
	US showtimes page – Map was totally unnecessary- confusing
	Two searches next to each other – went to google because of familiarity
	Box office page – Table was confusing.
	Thought 'My movies' search parameter was intuitive but it wasn't
Traditional/ Experienced	Search was disappointing - tried searching Al Pacino + 1983 for closer match– dint work.
	Search did not help – gave movie names instead of info on box office earnings.
	User did not see the Zip dialog page in the now playing page. Not enough visibility.
	US showtimes page – Zip code box on left did not work – just repopulated the same page – expected movies for that zip code
	US showtimes page – expected showtimes for the movie on selecting it but it asked for zip code again. Redundant, it should have remembered my choice.
	Again used IE find feature to locate box office link because there were too many links
	Top movie page – Too many sections worded Top this, Top that etc - very confusing.
	Search - tried searching Al Pacino + 1983 for closer match– dint work.
	US showtimes page – Zip code box on left did not work – just repopulated the same page – expected movies for that zip code
	US showtimes page – terrible Map – did not even try because it had small dots.
	Finding Nemo page – had too many links on the left.

	Map was bad, cant click on it. Typed Zip code.
	Had to select Radio again from a list though had pre-selected it earlier.
	By chance saw the top 250 link as it was not visible in the initial screen space.
	US showtimes map page – expected theaters for the movie radio upon entering the zip code in the map page but it showed me list of movies and their showtimes
	Surprised by the large number of movies that were listed for the zip code – expected just a list of showtimes for the movie Radio.
	US showtimes page –Map – did not even try because it had small dots.
	User was surprised that the font size was smaller than text on screen – expected it to be more prominent. – the content is more prominent than the menu tabs – menu tabs were like ads.
	Home icon wasn't intuitive…not consistent with other tabs.
Traditional/ novice	Search isn't great – tried searching Al Pacino + 1972 for closer match– didn't work.
	Zip code Area on the right was misleading. Thought that the right one with the movie selection option would yield showtimes for a particular selected movie. It did not do so.
	Search did not help. No option for searching box office info.
	A-Z page – Here tried the Search index link thinking it is some kind of search but it was a page with instructions on how to search-misleading
	Expected a search parameter like "actor". Prefer a more general search.
	In the browse imdb page – expected to search by layers – Prefer search filters to be layered – Search for year first then search for that actor's movies for that year.
	Expected to see showtimes for Radio after putting in the zip code as user had pre-selected Radio before - but it didn't help.
	Top movies page – too many similar words (All-time etc)
	Search box looked like an ad & it was next to google search.
	US showtimes page – Zip code box on left did not work – just repopulated the same page – expected theaters for that zip code
	Map page – the zip code at the bottom was not visible in the initial screen. Saw it the next time.
	Had to use back to go to home, couldn't find a Home link or icon.
	Had to use the bottom menu again but would prefer if the same links were provided on top.
	Again search drop down was useless just searched directly.
	Again search drop down was useless so went to top movies page
	"All time" links under Top grossing section misled the user to a diff page.
	Too many tops in the top movie page – titling should be different.

G.3 Low severity negative critical incidents collected from this study

UE type / User group	Low Severity Negative Critical Incidents
Remote/ Experienced	Irritating pop-up
	Another pop-up
	Pop-up
	Top menu bar is too crowded and caused me to click on the wrong link
	Animation ad was confusing
	Box office Table page- prefer if the Title of movie was aligned to the left
	Showtimes page – Not sure as to what way the movies were listed (alphabetic or not
	Top Movies Page – Table layout was confusing, prefer one or two column layout
	Showtimes page – Not sure as to what way the movies were listed (alphabetic or not)
	Expected a call out when placing a mouse on a link
	Annoying pop ups
	Annoying pop ups
	Expected ascending order filmography
	Out of time movie page - Expected the US movie showtime tab to map to the current movie and show the times for that movie
	Al Pacino page – Filmography term was not familiar.
Remote/ novice	Al Pacino Filmography page -No clear instruction/indication that the name next to the movie name was the character name – had to confirm by looking at other movies
	Annoying pop-ups
	No “back to top” link after scrolling so much.
	Unable to copy text from the box office table cell.
	Home - Looked for a show times link on the left nav but couldn’t find any
	Annoying Animation ads
	Menu Tabs not descriptive enough
	Home – No option for entering zip code and searching directly – looked for it
	Animation was distracting
	Al Pacino page – Though the movies were listed by year it wasn’t clear, not intuitive and hard to figure out the layout

	The box office link in the left nav of the Scarface movie page was misleading – took you to that particular movie's box office info – expected a general box office page
	Annoying pop-ups
	Annoying pop-ups
	Box office final page – showing unnecessary information about Argentina when user is in UK website.
	Too much scrolling in this page
	No clear indication that the name next to the movie name was the character name – had to go into the movie to verify.
	Pop-up
	Hate pop-ups
	Did not understand what 5m,10m meant, had to click and check.
	Pop up – tried ignoring
Traditional/ Experienced	Pop-up annoying
	Al Pacino page – Cast wasn't clear
	Al Pacino page – Did not realize that cast overview had the character name. It was not obvious. Had to browse & read a bit to figure that out.
	Hate pop ups
	Top rated page – Heading was top voted, so was little confused if top rated was the same as top voted and took awhile to figure out.
	Browse imdb page – the cast/crew field was not clear - but guessed that it must be the place to search.
	Home – too much text/too busy
	Al Pacino page – Did not realize that cast overview had the character name. It was not obvious. Had to browse & read a bit to figure that out.
	Al Pacino page – Did not realize that cast overview had the character name. It was not obvious. Thought that was of the reviewer. The way that information was displayed/designed was not clear.
	Expected to see earning info in the top movie (finding nemo) page. Did not find any so looked for other links.
	Browse Imdb page -> Search Page-> Find by People page -> did not understand why they had to have search filters like actor, actresses (they are all the same- people)
	Hate pop-ups.
	Expected to see earnings/numbers in the top movie's (finding nemo) page but couldn't find any.
	Al Pacino page – Did not realize that cast overview had the character name. It was not obvious. Had to browse & read a bit to figure that out.
	Home Page - The UK US icons were more prominent and emphasized and took the user's attention. User did not see the top 250 link the first time.
	Al Pacino page – Did not realize that cast overview had the character

	name. It was not obvious. Had to browse & read a bit to figure that out.
	Home - Did not see the showtimes tab at the top – think it was because the tab was on the far right – actually went for Now playing which was the second tab from the left.
Traditional/ novice	Showtimes text in the homepage was bold and thought it was a link-but it wasn't -weird
	Stupid pop-up
	UK Top 15 page – the dates were in the English format and was confusing initially.
	Did not see the menu tabs at the top – think it was because the tab and the background ad were same color-blue
	Two Zip code searches on the same page doing the same function was not understandable.
	Constant moving animation ad in the homepage was distracting.
	The box office link in the left nav of the Scarface movie page was misleading – took you to that particular movie's (scarface) box office info – expected a general box office page.
	Lots of stuff in this website compared to others.
	Little confused with the Box office earnings table – prefer a colored column approach to distinguish between various information.
	Now playing page – Us movie showtimes dialog box on the right was accidentally seen and not very obivious.
	Home page - Tried searching for the showtimes page using the general search but it returned movies named showtime or something similar
	Pop up threw me off
	Was looking for info on how it was rated and that was not provided in the Top rated 250 page but in the previous page (top movies)
	Box office table page - Unwanted info at the bottom about "highest top box office" movie – was very confusing and unwanted.
	Annoying pop-up
	Search drop down list was useless - not enough parameters to search by. Expected the result to show Top ratings page but showed movie names similar to the words 'top rated'

G.4 Positive critical incidents collected from this study

UE type / User group	Positive Critical Incidents
Remote/ Experienced	'Top Movie' link was useful
	'Showtime' tab was easy to find.
	Visual clue – 'Top rated' text at the top helped.
	Liked the layout of showtimes listing

	Home page - “Tops at the Box office” in the left nav was pretty useful.
	US movie show times tab on the top menu was helpful
	“Top Movies” link in the top menu was very clear
	Left menu navigation was very clear
	No hesitation in clicking on “Top movies link” due to prior use
	Home Page -Liked it, liked the menu bar on top
	Home - “Top movies” link helped me, drove me.
	Top movies link was intuitive
	Now playing page – liked the zip code search dialog
	Home – search was very comprehensive, liked the search parameters
	Home page - Liked the left nav – pretty detailed
	Now playing page – the US movie show times search dialog was helpful.
	Liked the “top 250” link in the home page
	Top movies link was helpful
	Home page icon was intuitive
	“Top Movies” link in the top menu was a good feature
	Lots of info- that’s good
	General Search was good. Liked the search parameters
	They maintain a consistent layout that’s good.
	Liked the Top 250 link at the bottom.
Remote/ novice	Search is pretty good
	Clear distinction between current & archives helped
	Map page – the Map was a nice option additional to the zip code.
	Top movies link was helpful- I new that’s where I should go
	Search was very helpful – made it a lot easier
	Looked for link “Top 250” and then it was easy.
	Liked the top movies page
	Well organized
	It was good that they (pop ups) went to the background.
	Search was good – most logical. Liked Search parameters
	Top movies link was helpful- just went for it.
	Jump tos (1970,1980) in the AL Pacino page were helpful.
	Search was good – best way.
	Remembered seeing UK in the top movies page – was helpful and went to it.
	The top 250 link in the top grossing table helped in directing to the correct information.
	Search was very helpful
	Remembered seeing Top UK in the top movies page – was helpful and went to it.
	Was quite comfortable with the zip code entry
	Top movies link was helpful

Traditional/ Experienced	Color layout was good.
	Box office Table page – Very clear distinction of earnings
	Saw the top 250 link in the top grossing page it helped in leading to the correct list.
	Remembered 'Top movies' link from my previous task and expected to find the necessary info for this task here and went to it.
	'Top movies' link was intuitive and helpful.
	Search was good, pretty easy
	Box office Table page – Very straight fwd and good table layout.
	Now playing Tab was intuitive
	Search - tried searching Al Pacino + 1983 for closer match– dint work. But gave link to Al Pacino page.
	Home page – the links to UK were good.
	Home page icon was intuitive
	The top 250 link was easily visible in the home page and just went for it.
	Al Pacino page – Liked that they were listed by year
	Business data page was good
	Home – US movie showtimes tab was very visible and easy.
	Clicked on "Top movies" link – that made sense.
	Al Pacino page – Liked that they were listed by year – chronological.
	Search was convenient
	Remembered 'Top movies' page had a top UK from my previous task and expected to find the necessary info for this task here and went to it.
	Top movies link was obvious
Traditional/ novice	User used the showtimes link for another movie to get to the showtimes page
	The "top movies" was descriptive and remembered to go there.
	Thought Search was the best way to do it and it was
	Looked for a link for UK and found it easily in the homepage
	The lower horizontal menu (colors) was more obvious then the upper blue menu tabs
	Remembered that there was a link "Top 250" and then it was easy.
	Thought it should be under the Top movies link. Hence went for it and found the info
	Figured anything to do with top ranking, earning, etc would be under "top movies" so went to that page
	Pretty simple once u get to top movies page.
	The website was very comprehensive
	Top movies link was intuitive
	Home page -US showtimes tab was right there - was helpful
	Again went to the Top movies page directly

	Liked top movies page with good links.
	Good was – mo moving images or graphics