

2004

## The Effect of Testing Location and Task Complexity on Usability Testing Performance and User-Reported Stress Levels

Chris Andrzejczak

*Embry-Riddle Aeronautical University - Daytona Beach*

Follow this and additional works at: <https://commons.erau.edu/db-theses>



Part of the [Applied Behavior Analysis Commons](#)

---

### Scholarly Commons Citation

Andrzejczak, Chris, "The Effect of Testing Location and Task Complexity on Usability Testing Performance and User-Reported Stress Levels" (2004). *Theses - Daytona Beach*. 13.

<https://commons.erau.edu/db-theses/13>

This thesis is brought to you for free and open access by Embry-Riddle Aeronautical University – Daytona Beach at ERAU Scholarly Commons. It has been accepted for inclusion in the Theses - Daytona Beach collection by an authorized administrator of ERAU Scholarly Commons. For more information, please contact [commons@erau.edu](mailto:commons@erau.edu).

**The Effect of Testing Location and Task Complexity on Usability Testing  
Performance and User-Reported Stress Levels**

By

Chris Andrzejczak

A THESIS PROPOSAL SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE  
DEGREE OF MASTER'S OF HUMAN FACTORS & SYSTEMS

Embry-Riddle Aeronautical University

Human Factors Department

© Chris Andrzejczak 2004

Embry-Riddle Aeronautical University

All rights reserved. This work may not be reproduced in whole or in  
part, by physical or electronic means, without express permission from  
the author.

UMI Number: EP32065

### INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI<sup>®</sup>

---

UMI Microform EP32065  
Copyright 2011 by ProQuest LLC  
All rights reserved. This microform edition is protected against  
unauthorized copying under Title 17, United States Code.

---

ProQuest LLC  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106-1346

THE EFFECT OF TESTING LOCATION AND TASK COMPLEXITY ON USABILITY TESTING  
PERFORMANCE AND USER-REPORTED STRESS LEVELS

By

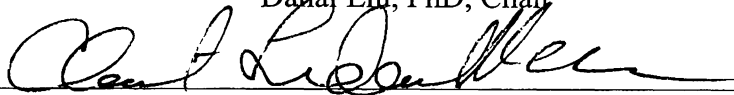
Chris Andrzejczak

This thesis was prepared under the direction of the candidate's thesis committee chair, Dahai Liu, PhD, Department of Human Factors and Systems, and has been approved by the members of the thesis committee. It was submitted to the Department of Human Factors and Systems and has been accepted in partial fulfillment of the requirements for the degree of Master of Science in Human Factors & Systems.

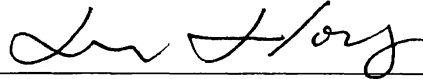
THESIS COMMITTEE:



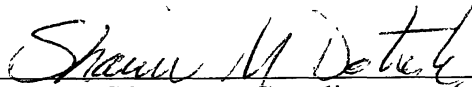
Dahai Liu, PhD, Chair



Christina Frederick-Recascino, PhD, Member



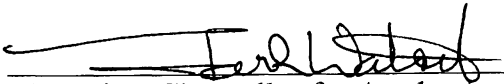
Hong Liu, PhD, Member



MS HFS Program Coordinator



Department Chair, Department of Human Factors and Systems



Associate Chancellor for Academic Affairs

## ABSTRACT

Usability testing is becoming a more important part of the software design process. New methods allow remote usability testing to occur. Remote testing can be less costly and allow more data to be collected in less time in many cases, provided the user can still provide meaningful data. However, little is known about differences in the user experience between the two testing methods.

In an effort to find differences in user experience between remote and traditional website usability testing, this study randomly assigned participants into two groups, one completing a usability test in a traditional lab setting, while the other group utilizing a remote testing location. Both groups completed two tasks, one simple, one complex, using Amazon.com as a test interface.

Task time and number of critical incidents reported were the dependent measures. Significant differences were found for task times both in the between and within-subjects conditions for task times. Task times differed significantly between task types; the complex task took generally twice as long as the simple task. No significant differences were found for critical incident reports for both the between and within-subjects conditions. Participants seemed hesitant to report interface problems, preferring to struggle through the task until they satisfied task requirements. Subjective user assessments of the task and website were similar across both conditions. User behavior navigating the site was remarkably similar in both test conditions. Results suggest a similar user testing experience for remote and traditional laboratory usability testing.

## ACKNOWLEDGEMENTS

This work first started occupying a little space at the back of this student's mind in a class called HFS 510, Dr. Shawn Doherty presiding. Dr. Doherty spoke of an academic adventure in which we would all have to partake to complete our respective degrees. This adventure, complete with all of its upsets, downfalls, blind wanderings through academia, late nights kept awake, and finally serendipitous discovery and appreciation for the whole process draws to a close here in these pages.

Although most adventures seem like they are a solo affair, this adventurer had much guidance, assistance, and support through his foray into academia. The list is long, however not trivial. Gratitude is most expressed to Mom & Dad, whose constant reminders about why I am in Daytona kept me on track. My sister, whose own adventures in academia, exceeded only in their difficulty by their variety, drove me to try to shine as brightly as she does. Professor Dahai Liu, who provided the time, patient guidance, papers, and most needed lunch breaks required in this line of work. My other committee members, Professors Christina Frederick-Rescasino, Ramzi Sekhar, and Hong Liu deserve recognition for their advice and inevitable corrections, foresight, and aura of scholarly presence. Experimental research is only as good as the participants it is performed on, so heartfelt thanks goes out to the 61 people who participated in my half-hour of experimental torture. Thank you, my friends in Daytona and beyond, for putting up with my whining, moaning, and groaning about this process. Thank you to all my inspiring follows, past, present, and future, let's dance! Thank you too, the reader, who is the driving force and end purpose of academia; the same reader who will of course overlook the remaining glaring typos and grammar mistakes; I assure you I have read and reread this work to try and keep it error-free. Finally, thanks to the underappreciated, overworked, and most certainly underpaid (on many levels) ILL system that provided astounding service throughout, a service not just extended to this lone adventurer.

## TABLE OF CONTENTS

APPROVAL .....	II
ABSTRACT .....	III
ACKNOWLEDGEMENTS.....	IV
LIST OF FIGURES .....	VIII
LIST OF TABLES .....	IX
INTRODUCTION .....	1
PURPOSE FOR RESEARCH .....	2
REVIEW OF THE LITERATURE .....	4
Usability.....	4
Software Usability .....	6
Usability Testing.....	8
Usability Testing Techniques .....	9
Discount Usability Methods .....	10
Remote and Traditional Usability Testing.....	13
Critical Incident Technique.....	17
User Stress and Task Performance .....	19
Definition of Stress .....	19
Arousal Levels .....	20
Evaluation Arousal.....	20
Stress and Cognitive Performance .....	22
Task Complexity.....	22

Objectives .....	24
Hypothesis.....	25
<b>METHOD .....</b>	<b>26</b>
Participants.....	26
Apparatus .....	26
Design .....	26
Procedure .....	28
Rationale for Method .....	30
<b>RESULTS .....</b>	<b>32</b>
Participant data.....	32
Task times .....	36
Critical incidents .....	39
Stress levels.....	42
Subjective data .....	44
Summary .....	45
<b>DISCUSSION .....</b>	<b>46</b>
Major Findings.....	46
Internal Validity .....	47
Participant behavior .....	48
Potential difficulties .....	50
Study limitations .....	51
Practical Implications.....	52



Future research.....	53
CONCLUSION.....	54
REFERENCES .....	56
APPENDIX .....	59

## LIST OF FIGURES

Figure 1 Yerkes-Dodson Law .....	3
Figure 2. Remote testing illustration.....	14
Figure 3. Number of participants reporting number of hours spent weekly in labs .....	32
Figure 4. Years of experience with computers .....	33
Figure 5. Number of users reporting prior experience types .....	34
Figure 6. Mean task times between test conditions .....	37
Figure 7. Estimated marginal means for task time .....	38
Figure 8. Number of critical incidents reported per task .....	40
Figure 9. Number of participants reporting critical incidents per condition.....	41
Figure 10. State and trait anxiety scores for both test conditions .....	43
Figure 11. Remote and traditional subjective appraisals (higher scores are better) .....	44

## LIST OF TABLES

Table 1. Variables and levels .....	27
Table 2. Variables and assessments .....	27
Table 3. Descriptive statistics for remote condition .....	35
Table 4. Descriptive statistics for traditional condition .....	35
Table 5. ANOVA source table for within subjects task times .....	36
Table 6. ANOVA source table for between subjects task times.....	36
Table 7. ANOVA source table for within subjects critical incidents reported .....	39
Table 8. ANOVA source table for between subjects critical incidents reported .....	39
Table 9. Nonparametric test results .....	45

## INTRODUCTION

As software increases in complexity, the interface employed to afford interaction with said software increases in complexity as well. Emphasis on a good interface design has prompted usability testing to be included within the software design regimen.

Usability testing typically takes the form of user recruitment, task assignment, and observation of user behavior within a usability lab. The drawback to this process is that sample sizes are generally small due to the expense of recruiting and physically transporting test participants to a testing location. Within the past decade, new networking technologies emerged that enable test participants and evaluators to be physically apart in space and/or time.

However, little is known whether the user performs similarly in a usability lab setting as in a remote testing environment. In particular, no studies have been performed that assess user's stress levels in either testing scenario: remote or traditional on-site.

## PURPOSE FOR RESEARCH

The literature review turned up only a few studies that compare participant performance in a formal testing lab and in a remote setting. Furthermore, no studies have thus far compared participant stress levels between the two conditions. It is possible that the effect of the observer's presence or the user testing within a formal testing environment affects task performance. In addition, number of critical incidents reported may vary with stress level.

A successful website or software interface is intuitive and meets standards for usability. Usability testing continues to evolve; new methods such as remote testing allow evaluators to employ larger participant groups. These groups are potentially more diverse, and this diversity contributes value and quality to the data collected. As an added bonus, administering the test in a user's natural work environment enhances external validity.

The lack of information on participant stress levels between traditional and remote usability test settings inspired the creation of this study. Stress level affects user task performance. Yerkes and Dodson (1908) realized a non-linear relationship exists between stress levels and performance. Performance at low stress levels can be just as low as when stress levels are high; this suggests an "optimum" stress level where performance is greatest may exist. The study aims to quantify these stress level differences.

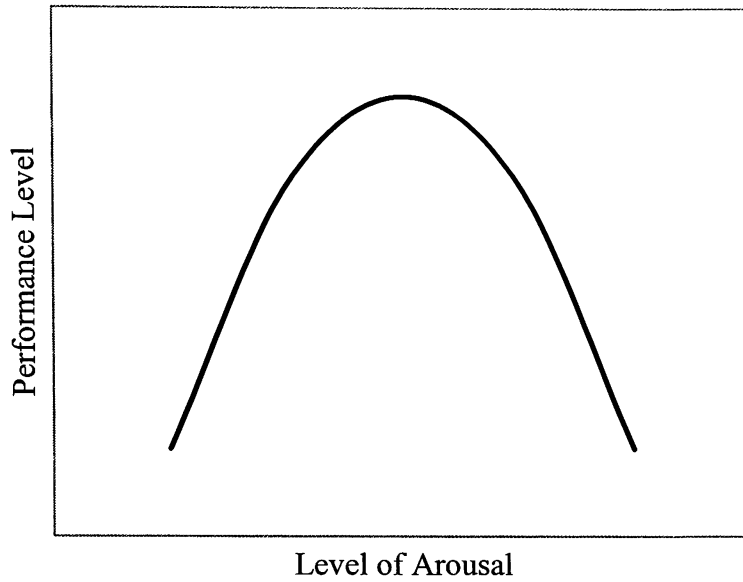


Figure 1 Yerkes-Dodson Law

Ideally, the testing platform should elicit an optimum level of arousal in users. The unfamiliarity and formal nature of a traditional testing lab may arouse users past the “optimum level” and yield undesirable testing results. Conversely, remote testing may not arouse the user to perform the tasks as they would during a normal workday or traditional lab test.

## REVIEW OF THE LITERATURE

### *Usability*

Usability is a term that refers to the degree of how easy, intuitive, and efficient an entity is. The entity in question may comprise a software interface, a website, a control panel, or anything that requires user input. The Institute of Electrical and Electronics Engineers (IEEE) defines usability as “the ease with which a user can learn to operate, prepare inputs for, and interpret outputs of a system or component.”

Usability is not a simple single-dimensional aspect of an interface; rather it comprises a range of components. Nielsen (1993) defines five usability attributes.

- Learnability – the degree to which the system is intuitive to the user and thus requires little or no training for the user to start productive work
- Efficiency – a measure of how productive a user is once training on the interface is complete
- Memorability – a facet of the interface that allows the casual user to be away from using the system and yet retain a high level of proficiency upon their return to using the system
- Errors – the system should afford a low error rate, and when errors do occur due to user input or system faults, they should be recoverable; the system should not contain catastrophic errors
- Satisfaction – the system should allow the user to have a pleasant experience

Nielsen’s goal for usability is one in where the discipline is: “approached systematically and eventually measured according to the previously listed criteria” (1993). Usability measurement involves testing a representative sample of real users

performing prescribed tasks. Usability tests can also take the form of observation of users in their natural work environments, going about their usual activities. Usability measurements are taken relative to users and their respective tasks; measurements cannot be equated across tasks that vary completely in their outputs. For example, a user who requires indexing software for his/her personal photographs will prefer a different interface than a user indexing data files for backup.

When designing for usability, the best guess of the designer is not good enough (Nielsen 1993). Instead, an attempt at understanding the users and their tasks drives the design followed by verification against accepted standards and models. Usability engineers often face contradictory statements and requirements made by users, and even the methods for designing a “usable” interface are riddled with inconsistencies. Nielsen states that changing an interface based on user testing is the mark of a mature usability engineer.

It is important to remember however, that the user is not always right when it comes to interface design (Nielsen, 1993). There are instances in which the user does not know what is best for them, particularly when introducing a completely novel interface or design (Gray et al, 1990).

It is important to remember that users are not designers and that designers are not users (Nielsen, 1993). The mental model that designers assume users possess rarely matches the actual user mental model. The classic example of this is the difficulty involved in programming a VCR to display the correct time; typical users find the task frustrating. Designing an intuitive way to do this appears to have fallen out of favor. Modern VCRs set themselves according to a signal embedded in radio transmissions.



Finally, it is important to note that designing for usability involves trade-offs. A system that appeals to the novice user may not appeal to an expert user of that system. For example, the novice-tailored system may contain simplistic features and guided menus that appeal to novice users. An experienced user of the system would prefer shortcuts and advanced features to automate and customize operation, but including such features may overwhelm said novice user. Nielsen (1993) proposes “accelerators” that gradually introduce advanced functions and shortcuts to those users who desire such features.

To summarize, usability is a difficult concept to measure and design for with any interface. It encompasses multiple dimensions. It is measured by the observation of users performing specific, relevant tasks. Usability standards, models, and heuristics do exist for a variety of design scenarios, but no one approach or method will solve all usability problems. Taking into account users and their task goals will go a long way towards designing a satisfactory and desirable interface.

### *Software Usability*

Software is the code that determines how hardware will operate. Hardware has been dramatically changing in the past 50 years; it has been almost exponentially gaining complexity and becoming more capable and feature-ridden. In 1965, Gordon Moore, the co-founder of Intel, posited the observation that the number of transistors per square inch on integrated circuits doubles every year since their invention. In loose terms, this means that integrated circuit data capacity doubles every calendar year. Due to these rapid changes in hardware complexity, software must also become more complex and more capable.

To keep up with the demands that these dramatic changes in hardware have placed on software, the software development process has altered to follow suit. As hardware complexity and capability increases, the software that runs on it will eventually increase in complexity as well, affording the software to exploit more features of the hardware.

Software acts as the liaison between the hardware and the user. The part of the software that the user manipulates is the user interface. The user interface allows direct manipulation of the hardware and its features by the user. User interfaces have undergone drastic changes in the past few decades, ranging from punch cards to parser interfaces to modern WIMP (windows, icons, menus, pull-downs) graphical user interfaces. As user interfaces invariably gain complexity, it becomes increasingly apparent that not all users of those interfaces are the same, or interact with the interface in the same fashion. A way of testing these interfaces was deemed necessary, as the software engineers who designed the user interface tended to design the interface for themselves, rather for the actual end user. In other words, the designer's mental model of the interface did not coincide with the user's mental model.

Due to these discrepancies in mental models, system engineers added a new segment to the software development process called usability testing. Usability testing is akin to a good chef tasting his culinary creation prior to the final serving of the dish (Armstrong et al, 2002). Like this taste performed by the chef, it is generally quick and cheap but still likely to reveal much about the design. Armstrong, Brewer, and Steinberg (2002) define usability as "the degree to which the design of a device or system may be used effectively and efficiently by a human." Various definitions of usability exist; these

definitions may include concepts such as user satisfaction. What matters most in usability testing is quantifying the measure of usability the system affords the user, and utilizing these metrics to improve on or drive the design process. When the design process undergoes improvements, the entire software design process becomes more streamlined and efficient, thus generating software up to the task of driving increasingly complex hardware.

### *Usability Testing*

Nielsen (1997) states that actual user testing provides direct information about how a typical person may use a computer, and what usability problems may exist with the interface that is undergoing testing. When conducting a usability test, it is important to pay attention to problems concerning reliability and validity. Reliability is defined as the probability of achieving similar results with repetition of the test. Validity concerns itself with whether the test is an accurate assay of the issue undergoing scrutiny.

The issue of reliability is a real concern for usability testing, concludes Nielsen (1997). His rationale for this claim is that there are great magnitudes of individual differences between test participants. It is conceivable that one user will finish a given task many times faster than another user, perhaps using an inferior interface. Many times usability engineers must work with unreliable, yet externally valid data of this sort. Despite this unreliable data, Nielsen concludes, "some data is better than no data." He concludes that even though a 95% confidence interval is sought after in research, a more practical interval for development purposes is 80% (Nielsen, 1997).

Validity is especially important in a development environment, as it is a measure of whether the test is actually an accurate criterion for an interface's performance in the

real world. Validity is much more difficult to measure than reliability. Statistical methods exist to provide measures of reliability. A more basic understanding of methods used must be present to create an account of validity. Nielsen (1997) claims that most problems involving validity result when improper users are recruited and assigned inappropriate tasks as far as the usability test is concerned. It is also important not to confound the test itself. Evaluators do this by assuring the test administered is actually measuring its assigned variable.

### *Usability Testing Techniques*

Usability testing encompasses three main types of procedures and protocols that aim to provide a measure of system usability. The types range from simple self-reported data collection methods, to usability inspections by experts, to experimental testing. Armstrong, Brewer, and Steinberg (2002) state that a combination of these three procedures make up what is commonly called usability testing.

Surveying techniques used in usability testing typically involve exposing the intended users to the design. These basic ideas of exposing users to the design in question and learning their preferences are the basic underpinnings of surveying techniques. They may seem uncontrolled but can potentially illuminate many shortcomings in the intended design. Armstrong, Brewer, and Steinberg (2002) describe the following forms of surveying techniques (this is by no means an exhaustive list):

- Questionnaires – comprise any form of written feedback obtained from the user's experience of the design in question. These questionnaires are administered by paper or electronic means. They are usually less expensive than other information gathering methods and may not require evaluator supervision. Questionnaires

require meaningful responses from the participants in the usability study for the data to be effective. This places the burden of creating a meaningful and appropriate questionnaire on the usability study administrator. Pilot tests are typically conducted to evaluate a questionnaire's effectiveness.

- Direct Observation – is another method the usability study evaluator may choose to employ. Direct observation involves the administrator monitoring the participant's interactions with a given interface and noting these observations. There can be a large degree of variability in the structuring of these studies. The researcher's bias also becomes a factor in this type of data gathering. However, direct observation removes the burden of participants reporting their experiences and may even reveal problems that the participants themselves did not catch, especially when employing an experienced observer.

#### *Discount Usability Methods*

Usability testing can take many forms. These forms range from paper and pencil mockup to fully featured working prototypes and formal evaluation methods. Discount methods can use the aforementioned paper prototypes as well as employ heuristic evaluations, scenario evaluations, or "thinking out loud" testing. These methods are employable throughout the interface design process, and serve to quickly evaluate or suggest improvements to a design.

A heuristic evaluation is one in which a proposed design is compared to a list of established "good design" principles. Nielsen & Mack (1994) propose a comprehensive list as to what a good design entails. Expert designers usually conduct a heuristic evaluation, performing a critical inspection against the aforementioned rules. Main

purposes of these inspections include adherence to these design principles and provide overall consistency for the design. A timesaving and often attractive feature of this usability testing method is that it does not require participants to be present. Systems engineers often conduct heuristic evaluations prior to actual user testing, particularly during early design stages. The evaluators may conduct heuristic evaluations after user testing as well, in an effort to see if the user testing did not cover a specific aspect of the design.

Heuristic evaluations can involve one or more expert evaluators. If a single evaluator tests the interface, he/she usually has experience with usability testing. When multiple evaluators test an interface, they usually examine the interface separately and report their results upon completion of their respective analyses. When multiple evaluators work on an interface, both usability experts and domain-specific experts work together so that the interface undergoes evaluation from both the usability and subject matter standpoints. Nielsen (1994) reports that one evaluator will usually find 35% of usability problems during his analysis of the design, whereas 3 to 5 evaluators will find 75% of usability design problems.

Some disadvantages of the heuristic evaluation include the fact that heuristics do not provide design solutions to usability problems. The guidelines provided by the heuristics do not identify discrepancies between the designer's mental model and that of the user. In addition, heuristics are only as effective as the expert evaluator who is conducting them.

Another form of discount usability testing that Nielsen & Mack (1994) mention is that of "Paper Prototypes." Paper in this sense is a generic term for inexpensive and

quickly created mock-ups of the final design. Paper prototypes can be physical sketches on paper, or they can be simplified screen shots of interfaces coded so just the visual elements are intact. They can have rudimentary images, but usually are text only, with symbolic representations of links or actions.

Paper prototypes may also be used throughout the development cycle, especially when large changes are introduced to the design. Users perform tasks using paper prototypes by pointing to elements where they would ordinarily click with a mouse. Evaluators substitute interactive elements of the interface by employing folded paper and attached adhesive notes.

Paper prototypes offer an extremely flexible, inexpensive but low fidelity method for users to interact with a preliminary interface. The test platform is portable and not intimidating to even the most inexperienced of users. Potential drawbacks to the paper prototype method include the idea that complex elements of an interface are difficult to represent. Such complex elements would have to be tested later on in the design process.

The last major discount usability tool conjured by Nielsen & Mack (1994) is that of scenario-based testing. Scenarios aim to evaluate a limited portion of the interface. A scenario presents a user with scaled-down version of the final interface, or restricts the number of features available to the user. Evaluators gather user feedback through direct observation and thinking aloud protocols.

Scenario testing occurs often in website development. The data gathered simplifies design choices between major design schemes or page layouts. Scenarios allow quick testing of a specific function or element of a design interface. Problems may arise when two separately tested features must be used in conjunction with one another later

on. The interaction between these two features may have escaped scrutiny by specific scenario testing.

The discount usability testing methods, although not an all-encompassing evaluation of an interface, promote regular testing of an interface (Nielsen & Mack, 1994). Their low time commitment and low cost make them less of a chore for evaluators than full formal user testing. In addition, the data gathered from discount testing may not only prevent major design issues, but also allow more focused formal testing later on in the design cycle.

#### *Remote and Traditional Usability Testing*

Usability testing has existed in the software development process for quite some time now. Traditional usability testing involves a process of user recruitment and task assignment. The goal of selecting users is to have the users be as “typical” as possible, that is, representative of actual users in the field. These users find themselves in a lab with workstations. The evaluators typically hide themselves in an adjacent room with a half-silvered one-way mirror, observing the users and recording their task performance, noting items such as time taken, errors made, and subjective level of frustration. Upon completing the task(s), the users fill out a questionnaire describing their experience. This information allows evaluators to compose a list of suggested changes to the software interface.

Predictably, this type of testing can incur large costs, as users must be taken away from their workplaces, and be physically transported to the location of the testing lab. In addition, the evaluators must monitor the users in a seemingly artificial environment setup. Costs further increase if these users come from out-of-house locations.



New methods have come about that allow testing to occur without the physical presence of users in the traditional usability lab. These new methods allow evaluators to sidestep the problems of traditional usability testing, namely the costs incurred with lab testing and the few participants garnered due to these costs. In addition, the “false atmosphere” that the lab may create for some users no longer becomes a factor. These new methods comprise a new type of usability testing called remote evaluation. Hartson et al. (1996) define remote evaluation as “usability evaluation wherein the moderator, performing observation and analysis, is separated in time and/or space from the user.”

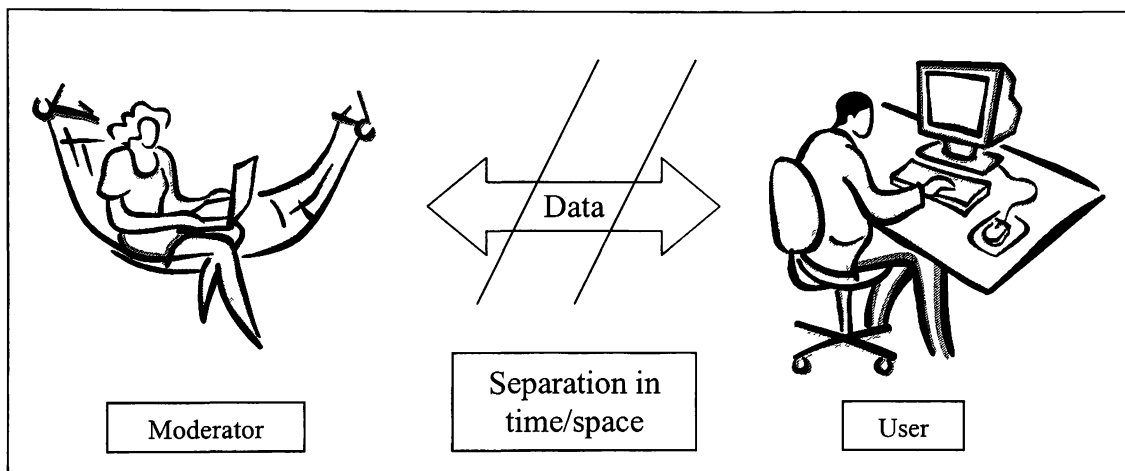


Figure 2. Remote testing illustration

Remote testing implies a discrete separation between the evaluator and the user in space, time, or both. Remote evaluation is not the same as the evaluator traveling to the user’s location and conducting a usability study there. Testing where the evaluator and user share the same environment is called field-testing, and is not akin to remote testing. One of the main ideas of remote testing is that by eliminating the physical presence of an evaluator during the test, a user is more likely to perform in the same fashion as he or she

would while in his or her natural working environment. This is an effort to increase external validity of the test.

Initial remote testing involved the use of videoconferencing equipment, freeing the users and evaluators from having to occupy the same building. Users may still have to go to a designated area away from their usual workplaces that affords the use of the videoconferencing equipment. The videoconferencing technique still requires that evaluators be in real-time observation with the users. Participant sizes are still limited by the number of users evaluators are able to commit their time to as well as the setup of the videoconference itself.

The later, more modern types of remote usability testing remove evaluators from the direct-observation real-time role. Hammontree, Wiener, & Nayak (1994) describe a few of the new tools that have arisen to promote remote usability testing.

For example, a remote tool may involve specialized software that collects and archives data automatically. This process involves downloading data capture software to the user test machine, or the use of a specialized website that captures data using its own devices. Using remote methods that do not require the ever-present supervision of the evaluator allows for the recruitment of a great deal many more participants. All other things equal, more participants tested means more potential usability problems found.

Remote evaluations usually comprise one or more of the following forms: a questionnaire or survey, live evaluation by humans, automated data collection, or user-reported critical incident method. All tests have levels of complexity and cost associated with them. Evaluators must choose the appropriate test and complexity level to meet budget, time, and user patience requirements.

Tullis et al (2002) conducted a study where task times and task performance were under observation for a remote and traditional usability-testing scenario. The research group wrote a JavaScript webpage that would assign tasks and ask for feedback in the form of interval measurements and a textbox for user comment entry. The webpage would appear in a browser window above a main window that presented the site under scrutiny. There was no communication between the two windows, and no software downloaded itself to the test user machine. This limited the amount and type of data available for collection. They found that task performance and times did not vary significantly between the remote and traditional scenarios. In addition, users provided verbose comments within the textboxes supplied by the JavaScript window. This bodes well for remote testing, as it appears users treat it “with the same respect” as traditional supervised training.

Technologies that enable remote usability testing bring users and developers together under the software development umbrella (Hammontree et al, 1994). Remote usability testing tools have therefore been garnered from tools used for Computer-Supported Collaborative Work (CSCW). Such applications provide a means for collecting usability data. Most CSCW tools incorporate elements from the following three activity types.

- Window/Application Sharing – These tools allow two or more users to share the same windowed workspace. One user may monitor what the other is doing, or one user may take control of the other remote workstation to instruct or coach that user on application use. Today’s multi-monitor hardware allows evaluators to

view their own workstation on one screen while viewing a remote test machine on another.

- Shared Whiteboard – A shared whiteboard is a workspace that multiple users may interact with at once. Users may sketch upon a whiteboard using digitizers, tablets, or mice, and in addition paste screenshots, documents, or in some cases 3D elements.
- Computer-based videoconferencing – Much as its name implies, this videoconference takes place on the computer screen, allowing for gestures and expressions to be transmitted, allowing richer, more natural interaction.

One application of the aforementioned technology involves using the shared application tool and a telephone to conduct a “thinking out loud” session with a user. A user interacts with the interface under scrutiny while telling an evaluator over the telephone what he/she is doing. The evaluator collects screen capture data alongside verbal testimonial data from the user.

Hammontree, Weiler, & Nayak (1994) employed a video link when conducting remote usability evaluations whenever possible. The claim was that video link affords a more personal experience, which builds rapport between the user and the evaluator that is otherwise missing in remote evaluation.

### *Critical Incident Technique*

The critical incident technique is a collection of procedures for observing behaviors that contribute to the success or failure of an individual operating a system (Flanagan, 1954). In the real world, users are in the best place to observe critical incidents

caused by interface flaws (Castillo, 1997), as they are usually aware of their choices that determine success or failure.

A critical incident is an occurrence that determines a performer's success or failure in a given task. If a user performs an action whose actual consequences do not match preconceived consequences, a critical incident has occurred. Critical incidents are context-specific. For usability study purposes, a critical incident is an indicator of a usability issue that is either positive or negative. For example, if a user cannot find the switch or knob that turns on the headlights in a car, a critical incident has occurred. Designers can use this critical incident to determine a proper course of action to minimize such occurrences in the future.

Origins of The Critical Incident Technique date back to the work of Fitts and Jones in 1947. They obtained information regarding "pilot-error" from untrained observers regarding critical incidents. Flanagan (1954) developed a practical guideline for using the critical incident technique, and employed trained observers to identify critical incidents as they arose. Flanagan identified a set of procedures and techniques for direct observation of human behavior in real-time, a departure from the earlier work by Fitts and Jones. The purpose for this observation of human behavior was to collect meaningful data to solve problems or investigate new psychological concepts. Critical Incident Technique outlines what constitutes an incident and sets up a systematic standard for its observation.

Flanagan (1954) defines an incident as "any directly observable human action that can be used to draw inferences or predictions about that individual and his/her behavior." A critical incident occurs when the observer has a clear conception of the reasons for

performing an act and its immediate consequences and the actual outcome does not match these preconceptions.

Usability professionals currently use and expand upon Flanagan's concept of critical incidents during their testing. To find usability problems reliably and accurately, critical incidents must be recorded in real time. Typically, this type of data results from a formal usability study conducted in a usability lab.

Castillo et al (1997) attempted to apply the critical incident method to remote usability testing. Their method involved keeping users in their own work environment to report critical incidents without direct interaction with evaluators. A user would click a button on a screen interface when he or she encountered a potential usability problem. This would cause a textbox to pop up asking the user to provide a textual account of the usability problem, and send a video clip of previous screen activity to evaluators. Castillo et al trained their users in critical incident observation prior to participating in the study. Evaluators compiled the information gathered from the user-reported critical incidents into a usability problem description for use in correcting the negative usability problem.

### *User Stress and Task Performance*

#### *Definition of Stress*

Stress occurs when an individual must adapt to a threat or challenge (Friedman, 1992). Stress is a set of neurological and physiological changes within the body that serve to help it adapt to a given situation. Stress can either be positive (eustress) or negative (distress).

### *Arousal Levels*

Miller (1963) found that stress could produce physical tension and anxiety, both factors that increase the onset of fatigue. Fatigue impairs performance. A stressed individual may not be able to focus his or her attention equally among separate tasks. Fixation may result where critical aspects of a task may be ignored, thus increasing the possibility for error.

### *Evaluation Arousal*

Task performance is closely related to evaluation arousal. Evaluation arousal varies between individuals, but it occurs when an individual's performance is tested. Many individuals have an intrinsic desire to perform or excel, especially when they are being watched. Therefore, it is a common occurrence that when people undergo evaluation, their performance may deteriorate due to this extra stress coming from being observed or scrutinized (Sarason, 1984). Sarason states that this deterioration results from disruptive thinking about deficits that may or may not exist within the individual. The best way to combat self-depreciating thoughts is to have the individual undergoing evaluation to focus on the current task. Teaching testers to focus on the current task goes long ways to reducing disruptive levels of test arousal.

Different coping mechanisms come into play depending upon how the individual perceives the current task. Lazarus (1981) developed a paradigm for coping mechanisms. When a stress occurs, a two-stage appraisal mechanism occurs:

- Primary appraisal – the stressor is determined to be a harm/loss, a threat, or a challenge. Harm/loss involves an injury or failure that has already occurred. A

threat is a possibility of something becoming a loss or failure. A challenge affords the potential for growth, learning, mastery, or other positive acquisition.

- Secondary appraisal – resources and options are evaluated at this stage. Physical, social, psychological, and material resources are inventoried by the individual.

Lazarus then defines two coping mechanisms that follow to attempt to reduce the stress introduced. Problem focused coping engages the individual in some sort of constructive problem-solving behavior, whereby planning and utilization of resources by the individual reduce the stress at hand. The other type of coping mechanism is emotion-focused coping, which posits the individual focusing on reducing the symptoms of the stress, rather than the stress itself.

When individuals face an examination or test of their abilities, they tend to appraise that situation as both a challenge and a threat. When individual perceive an examination as a challenge, they engage in problem-solving behavior. This is usually followed by a positive emotion to act as a motivator. However, those individuals who perceive an examination as a threat will employ emotion-based behavior, and in this case create a negative emotion. The individual may focus on handling the emotion and therefore be distracted from the important task. Individuals usually perceive situations they deem as uncontrollable as threats, where more controllable situations become challenges.

By testing users remotely, evaluators hope that one element that makes the process feel like a threatening examination no longer becomes an issue. Users may associate traveling to a formal testing location and being observed as an assessment of their abilities, and a threatening one that may affect their job, despite claims to the



contrary by the evaluator. By allowing the user to participate in the usability study in their natural work environment, evaluators hope to lessen unnecessary arousal level that a formal evaluation may elicit.

### *Stress and Cognitive Performance*

Stress affects cognitive performance. Shaab (1997) has found that stress can affect cognitive behavior negatively. Information processing capacity degrades and error rates increase. Guest (2001) has shown that stress can increase the chances of perceptual narrowing (fixating on one cognitive aspect or feature). In addition, working memory efficiency decreases. Working memory is important in neural processing of information; its degradation is highly undesirable when cognition-intensive tasks are performed.

### *Task Complexity*

Campbell (1988) defined guidelines for task complexity. Task complexity is a function of the psychological state of the individual performing the task. User perceptions of the task directly influence perceived task complexity. Differing elements of the task may deem the task more or less complex for the task performer. Campbell lists the following four elements that increase the complexity of a task:

- Multiple paths to the goal
- Multiple desired goals
- Conflicting interdependent attributes between paths
- Uncertainty or ambiguity of paths

Ideally, none of these elements would be present in a perfectly usable interface, but this would effectively cripple the usefulness or effectiveness of the interface. Such an

interface would be too simplistic and powerless to the user. Providing the user with a mix of task types (Campbell defines 16, ranging from simple to exceedingly complex) and method of completing objectives using more but simpler tasks when applicable make for a good design. A good designer understands and properly considers the tradeoff between complex single tasks versus smaller, simpler, yet more manageable tasks.

Byström & Järvelin (1995) define five task types that derive from real world information search scenarios. These five designations also lie within the perceptions of the task performer. They define *a priori determinability* of a task. This determinability is based on elements such as what background information the user has on the search topic, the method of finding more, and finally the outcomes of those searches. It is important to note that these are all preconceptions the searcher forms prior to the actual search task, and these preconceptions determine the subjective task complexity. More complex search tasks are those whose search terms, method of searching, and desired outcomes are uncertain to the searcher.

Much research has also been completed in the area of menu design. Research has shown that breadth is preferable to depth when designing menu systems (Miller, 1981). By increasing menu depth, task completion times and error rates increase regularly. Ideal menus should be broad, employing tabs to organize tasks in logical groups. Interface designers must strive to avoid designs where sub menus exist within nested submenus. Users may forget which level of a menu has a certain function and how to navigate submenus to reach that function. Spreading functions across readily accessible menus alleviates this potential design issue. Following Campbell's (1988) task complexity

theories, submenus display both the “multiple paths to goal” and “uncertainty of paths” pitfalls.

### *Objectives*

This study aims to find differences in perceptions of user stress levels between traditional laboratory-based usability testing and remote testing situations. The emphasis of the study focuses on user-reported stress levels between the two conditions. In addition, the experiment will employ two tasks of varying complexity.

The users will identify critical incidents as they occur, for both task conditions. Users will be trained in recognizing critical incidents according to Castillo’s (1997) method. Castillo defines a negative critical incident as “an event or occurrence observed within task performance that is likely to be an indicator of one or more usability problems.” Anytime a user has trouble in completing a task that is due to the nature of the interface, a critical incident occurs.

Upon completion of the two tasks, users will fill out a subjective stress questionnaire that will assess their stress level associated with participation in the study. In addition to assessing subjective stress levels, a subjective satisfaction survey that attempts to determine the user’s perception of quality and of pleasurable experience with the interface will be administered. These questionnaires will provide data for stress level associated with test location and satisfaction with both the test location and the interface.

## *Hypothesis*

This study posits the following three hypotheses:

- Hypothesis I: No differences exist between remote and traditional usability test performance as measured by time to complete both tasks.
- Hypothesis II: No differences exist between remote and traditional number of critical incidents reported
- Hypothesis III: No differences exist between remote and traditional user-reported stress levels as reported by the State Trait Anxiety Inventory (Spielberger, 1968)

## METHOD

### *Participants*

Participants were recruited from Embry-Riddle Aeronautical University's Daytona campus. Participants ranged in age approximately from 18-30. Participants were traditional college students ranged 18-25. Participants were recruited from varying majors. Majors included but were not limited to psychology, aeronautical science, human factors, and computer science. In order to be considered for either test condition, participants must report the use the computer labs on a regular basis, about 2 hours per academic week at a minimum.

### *Apparatus*

The experiment required the use of two computers for the remote condition and one test computer for the traditional lab setting. Computers used were a Dell Optiplex GX260 machine and a Dimension XPS R450. The test machines all had the full VNC server/client installed. Both machines had the AOL Instant Messenger Client installed.

### *Design*

The experiment consisted of a 2 X 2 mixed design. The independent variables were test location and task complexity, while the dependent measures were subjective stress score and number of critical incidents reported. The review of the literature found that with minimal training, users are able to identify their own critical incidents as well as rate the severity of those incidents (Castillo et al, 1998) reliably. Additional subjective assessment scores concerning website design and instruction clarity were collected.

The independent variables both had two levels. Participants were divided into either the remote or traditional testing scenario. This was the between-subjects aspect. Participants were exposed to both the simple and complex task. This was the within-subjects element. All participants recorded a subjective stress score upon completion of the usability tasks, followed by subjective appraisal scores.

Table 1.

*Variables and levels*

---

Independent Variable	Levels	Type
Testing Location	Traditional Remote	Between Subjects
Task Complexity	Simple Complex	Within Subjects

Dependent measures were subjective user-reported stress scores, time to complete each task, and number of critical incidents reported. The State Trait Anxiety Inventory (Speilberger, 1968) was the instrument used to assess stress level elicited by completing the tasks. The evaluator kept track of time for each task. In addition, participants filled out subjective questionnaires expressing satisfaction level with the website design, content, and clarity of the evaluator's instructions.

Table 2.

*Variables and assessments*

---

Dependent Variable	Assessment Method
Stress score	State Trait Anxiety Inventory
Time taken to complete task	Evaluator timing
Number of critical incidents reported	Critical incident reporting
Subjective satisfaction	Post-evaluation questionnaire

## *Procedure*

The experiment required about 15-20 minutes of each participant's time. Participants were randomly assigned to the remote and traditional usability testing groups. The use of a random number generator allowed participant assignment into either condition. The evaluator briefed participants that they will be evaluating a website on its content, ease of use, and aesthetic satisfaction level by performing two tasks on this website. Participants were instructed in recognizing a critical incident and reporting its severity. A user profile questionnaire was administered. After completion of the profile questionnaire, participants were given the two tasks to complete. Upon completion of the second task, participants completed the STAI. After completion of the STAI, participants filled out the subjective appraisal questionnaire, then were debriefed and allowed time to ask questions of the experimenter.

Remote Condition: Participants were given the pre-experiment questionnaire and told to proceed to the computer lab. The remote computer had a VNC server application running that afforded the moderator vision of the user's on-screen actions. Participants were encouraged to ask for help if they encountered difficulty with any of the scenarios using AOL Instant Messenger. Specifically, participants were told to report difficulty with the interface if it responded in a fashion inconsistent with their expectations. The remote machine had a dedicated account used for this purpose, cleverly called InterfaceTest. The tasks were administered in the form of printed instructions given to the participant. Critical incidents were reported in real time, as they occurred, thus allowing the experimenter to log critical incident quantities and offer assistance to allow the experiment to continue. Upon completion of the two tasks, the users completed the STAI.

The participant then returned to the evaluator's location to complete the subjective appraisal survey.

Traditional Condition: This condition placed the evaluator and participant within the same room, where the moderator directly observed the user's progress and recording critical incident data just as in the remote condition. Again, the tasks to be performed are given to the participant via printed material. The evaluator's corporeal presence takes the place of the AOL or MSN chat software used in the remote condition.

Assessing Task Complexity: The two tasks were designed according to task complexity criteria as defined by Campbell (1988). Simple tasks had comparatively fewer steps, few and clear paths to the goal, and little or no interdependencies between these paths. Complex tasks, on the other hand, required multiple steps to be taken by the user. They also required manipulations of web pages whose target links and functions are not immediately clear. In addition, the complex task required the users to have some background knowledge of the problem or subject material, thus adding to their complexity (Campbell, 1988). Half of the participants received the tasks in the order of simple first, complex second, and the other half of the participants performed the tasks in the reverse order.

Inspiration for the simple and complex tasks used elements from Liew's (2002) study on website usability testing. Subjective user task complexity was determined *a priori* by exposing participants not used in the study to only one of the two tasks. Users rated subjective task complexity on a Likert scale of 1 to 7 on selected elements of Campbell's criteria for task complexity.



Data Collection: Data collection took the form of time taken to complete each task, critical incidents reported, STAI scores, and questionnaire data. The evaluator timed users during the completion of their respective tasks. When the target certain page was reached, the evaluator stopped the clock and reset the computer for the next task. Profile questionnaires provided background data on users, and the STAI provided stress scores, both current anxiety scores and “general” scores that reflect everyday user anxiety.

Users were trained in recognizing critical incidents and reporting them as they arise. Number of critical incidents reported was recorded for each task. The last bit of data collected was the subjective appraisal questionnaire data, which comprised interval measurements of subjective satisfaction with site function, aesthetics, and clarity of the printed instructions provided by the evaluator.

Website Selection: The website used to conduct usability analysis was Amazon.com. Amazon.com was chosen because the site creators spend more time refining and perfecting their interface than any other consumer website (Flanders, 2004). In addition, Amazon.com provided a familiar shopping cart-style interface used in typical online purchases. Amazon.com employs a rich interface laden with redundant functions, affording users multiple ways to complete each task.

### *Rationale for Method*

This research attempted to mimic both a traditional laboratory setting and a remote setting. Dumas & Redish (1993) make suggestions for a traditional laboratory setup involving cameras pointed at the user. The remote setting was accomplished by having users complete the remote section within one of the on-campus labs, whereby the user was separated in space from the moderator. The rationale for this is based on

research done by Dorazio & Stovall (1997) that reports the environment influences human behavior. It is possible that performing usability evaluations within users' familiar work environments is preferable to testing within a traditional formal environment with cameras and isolated workstations.

Allowing users to remain within their natural work environments introduces confounding variability in the form of non-standardized workstations and software environments. To circumvent this problem, users performed the remote portion of the test within a public on-campus computer lab where hardware/software setups are much less variable.

The actual procedure for this study was a slightly modified version of a method used by Liew (2003) for a comparison between remote and traditional usability testing performed at the University of Nebraska. The method described by Liew (2003) follows suggestions for usability testing configurations suggested by Dumas and Reddish (1993).

## RESULTS

### *Participant data*

Participants chosen were chosen on the basis that they were familiar with the on-campus computer labs, and spent about two hours or more within the labs per week during the academic year. Participants who were not familiar with the labs were not eligible for the study. Most participants reported spending 0 to 2 hours weekly in the computer labs during the course of the academic year. Figure 3 illustrates participant lab usage data.

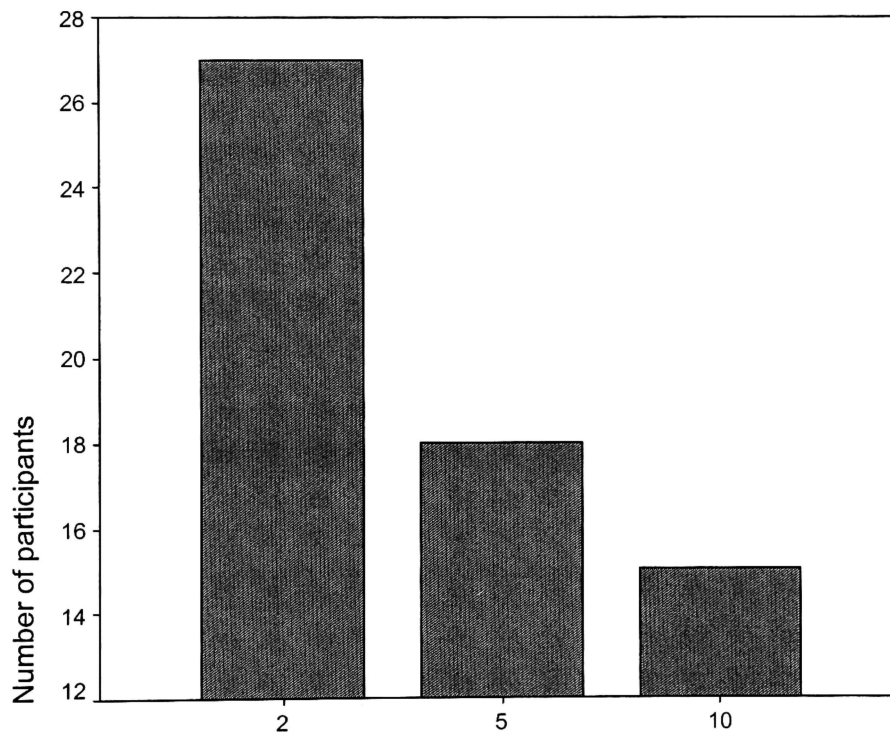


Figure 3. Number of participants reporting number of hours spent weekly in labs

Data were collected regarding number of years of computer experience that participants possessed. Experience was defined as the active manipulation of a software interface running on hardware that can qualify as a personal computer. Participants were given choices ranging from 0-5 years experience to 10 years or more. The bulk of participants reported having 5-10 years of experience with computers. Figure 4 illustrates the collected responses as number of participants reporting each choice.

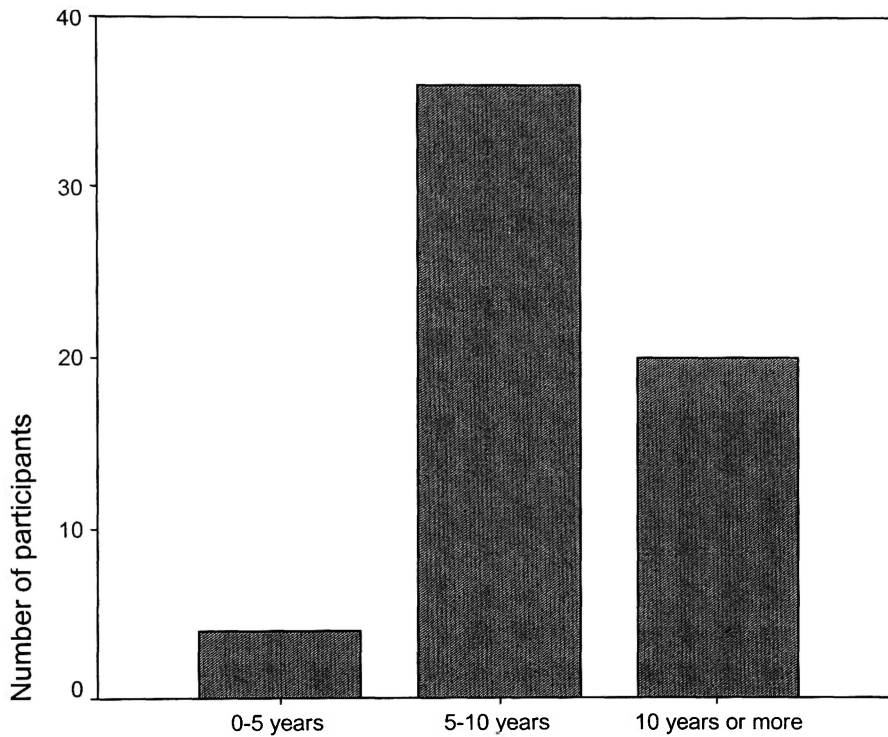


Figure 4. Years of experience with computers

Data about participants' online activities were assessed by the user profile survey. Participants were asked if they beta tested software or hardware, as well as queried whether they shop online and finally, if they create their own web pages. This information was collected to assess if any of these activities affect usability test performance. Users were not selected for or against based on their prior experiences. Figure 5 displays these data.

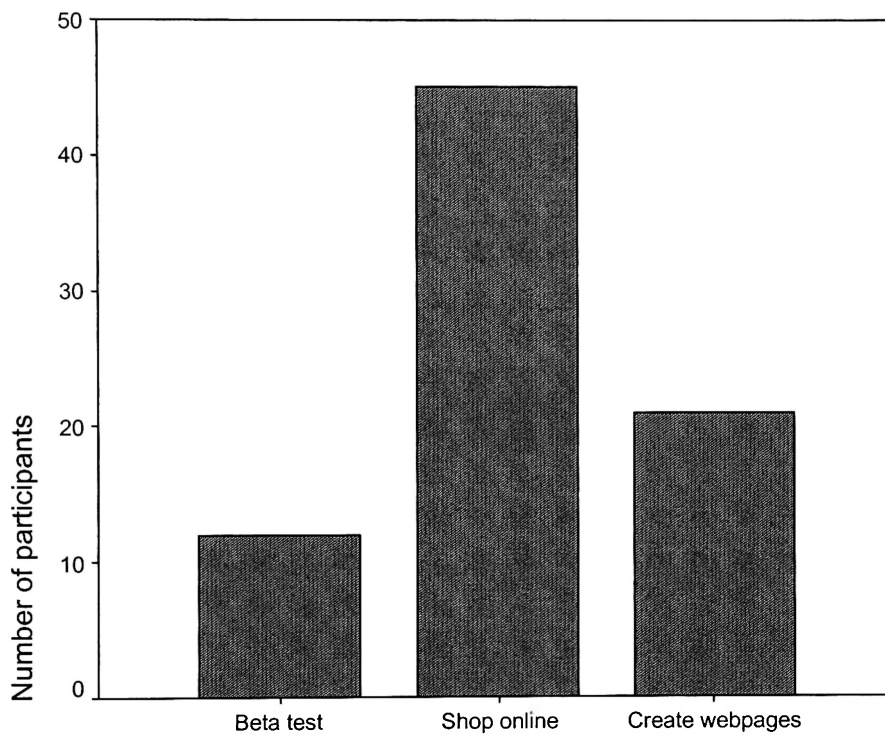


Figure 5. Number of users reporting prior experience types

Table 3.

*Descriptive statistics for remote condition*

Measure	N	Mean	Standard Deviation	Standard Error
Simple Task time (s)	30	130.50	60.27	11.01
Complex Task time (s)	30	356.50	186.09	33.98
Simple Task CI report	30	0.10	0.305	0.06
Complex Task CI report	30	0.37	0.61	0.11
State Anxiety Score	30	32.97	8.04	1.47
Trait Anxiety Score	30	34.77	7.96	1.45

Table 4.

*Descriptive statistics for traditional condition*

Measure	N	Mean	Standard Deviation	Standard Error
Simple Task time (s)	30	109.83	50.42	9.21
Complex Task time (s)	30	284.50	115.68	21.12
Simple Task CI report	30	0.13	0.34	0.06
Complex Task CI report	30	0.40	0.56	0.10
State Anxiety Score	30	33.07	9.67	1.77
Trait Anxiety Score	30	37.57	8.98	1.64

The purpose of this study was to determine if differences exist in user behaviors as measured by task times and critical incident report frequency between remote and traditional software test settings. Usability test performance was quantified by time taken to complete the tasks, as well as number of critical incidents reported. User-reported anxiety was assessed by Spielberger’s State-Trait Anxiety Inventory (STAI).

Descriptive statistics for the remote and traditional conditions are given in Tables 3 and 4. All times reported are measured in seconds. Stress scores fall on a continuum ranging from 20 to 80, where a score of 20 represents a low anxiety level and a score of

80 denotes a high anxiety level. Participants were asked to record subjective appraisal scores that ranged from 1 to 5. A score of 1 meant a low score in terms of subjective user satisfaction or understanding and 5 was a high level of that criterion.

*Task times*

The first hypothesis stated that no differences exist between remote and traditional usability test performance as measured by time to complete both tasks. A univariate analysis of variance was conducted on the first task time data. A significance level of 0.05 was used for all analyses. Within subject task times differed significantly due to task type  $F(1,58) = 103.62, p < 0.05$ . These data suggest that the effect of testing location caused significant differences in between-subject task times,  $F(1,58) = 4.162, p = 0.045$ . Table 5 presents ANOVA results for between subject task times. Table 6 shows the between subjects source table.

Table 5.

*ANOVA source table for within subjects task times*

Source	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>	Part $n^2$	Power
Task type	1	1204003	1204003	103.62	0.000*	0.641	1.000
Test location * Task type	1	19763	19763	1.70	0.197	0.028	0.250
Error	58	673933	11620				

Table 6.

*ANOVA source table for between subjects task times*

Source	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>	Part $n^2$	Power
Location	1	64403	64403	4.162	0.045*	0.067	0.518
Error	58	15475					

In general, mean times to complete Task 1 were lower than that of Task 2. Means did not differ greatly in magnitude between remote and traditional test conditions. Task 2 times were generally twice as long as task 1 times. Figure 6 depicts these data.

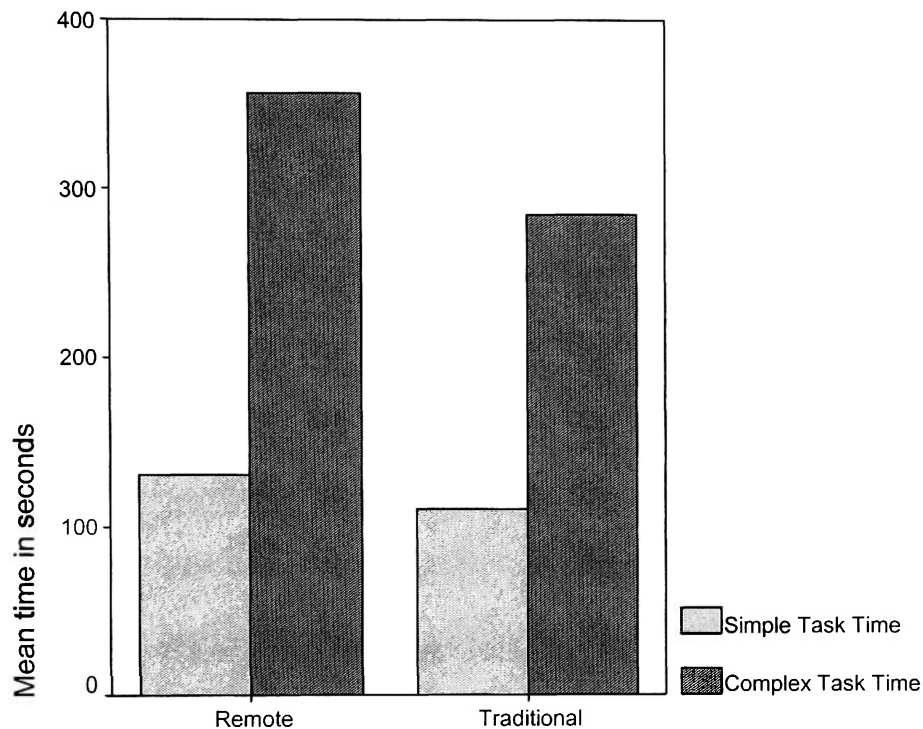


Figure 6. Mean task times between test conditions

Power for the between-subjects element was high due to the small overlap between population means and large sample size of 60 participants. This small overlap is signified by the effect size measure of 0.64. A high *F*-statistic supports the finding that differences exist in the within subjects task complexity variable.

However, for the between-subjects element, power was only computed at about 0.52. This value is the probability that the study will produce a significant result if the proposed hypothesis is true. Between-subjects populations had a large overlap as designated by the effect size measure of 0.067.



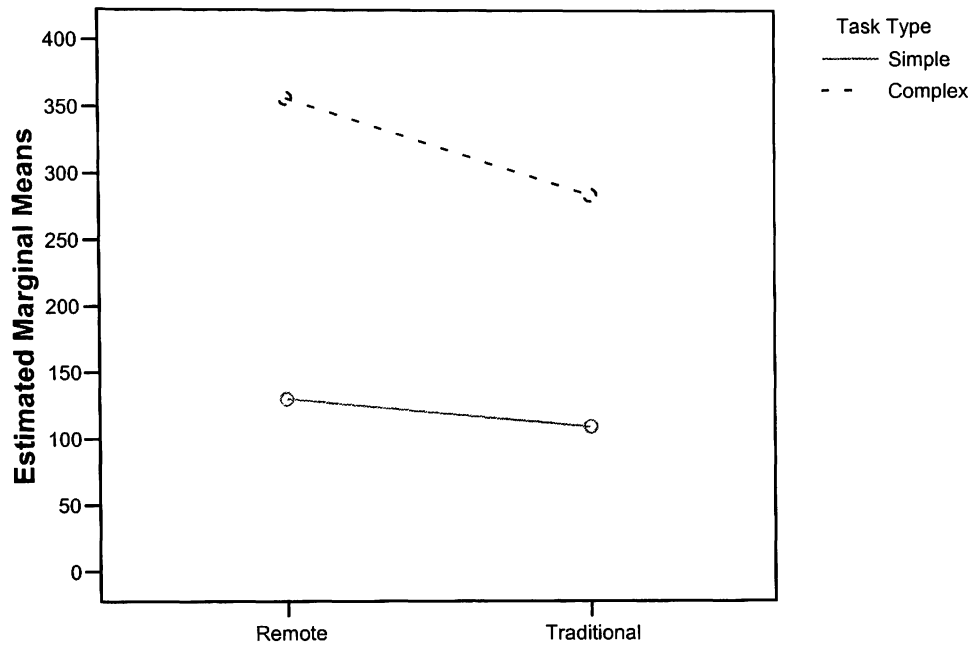


Figure 7. Estimated marginal means for task time

Figure 7 reports estimated marginal means for task times as they relate to each testing location. Traditional task completion times appear slightly shorter than remote task times for both the simple and complex task types. No interaction was found between task type and testing location by the data from this study.

Confidence intervals for task time mean differences were taken at the 95% level.

These values are reported below:

Simple task remote	mean = 130.5 (95% CI	110.2	to	150.8)
Complex task remote	mean = 356.5 (95% CI	299.9	to	413.1)
Simple task traditional	mean = 109.8 (95% CI	89.5	to	130.1)
Complex task traditional	mean = 284.5 (95% CI	227.9	to	341.1)

### *Critical incidents*

Participants were trained in the user-reported critical incident method, and encouraged on its use. Number of critical incidents (CI) reported per task were collected for each task type.

A univariate analysis of variance was performed on the number of critical incidents reported per task type in each testing scenario. A significance level of 0.05 was used. Task complexity had a significant effect on number of critical incidents reported per those tasks,  $F(1,58) = 10.43, p = 0.002$ . Test location did not have a significant effect on number of critical incidents reported per task,  $F(1,58) = 0.134, p = 0.716$ . No significant interactions were found from these results concerning testing location and task type. Table 7 is the ANOVA source table for number of critical incidents reported per task for within subjects. Table 8 shows between subjects comparison data.

Table 7.

*ANOVA source table for within subjects critical incidents reported*

Source	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>	Part $n^2$	Power
Task type	1	2.13	2.13	10.43	0.002*	0.152	0.888
Test location * Task type	1	0	0	0	1.000	0	0
Error	58	11.87	0.21				

Table 8.

*ANOVA source table for between subjects critical incidents reported*

Source	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>	Part $n^2$	Power
Location	1	0.033	0.033	0.134	0.716	0.002	0.065
Error	58	14.47	0.25				

Figure 8 shows the non-significant interaction for the mean number of critical incidents reported per task.

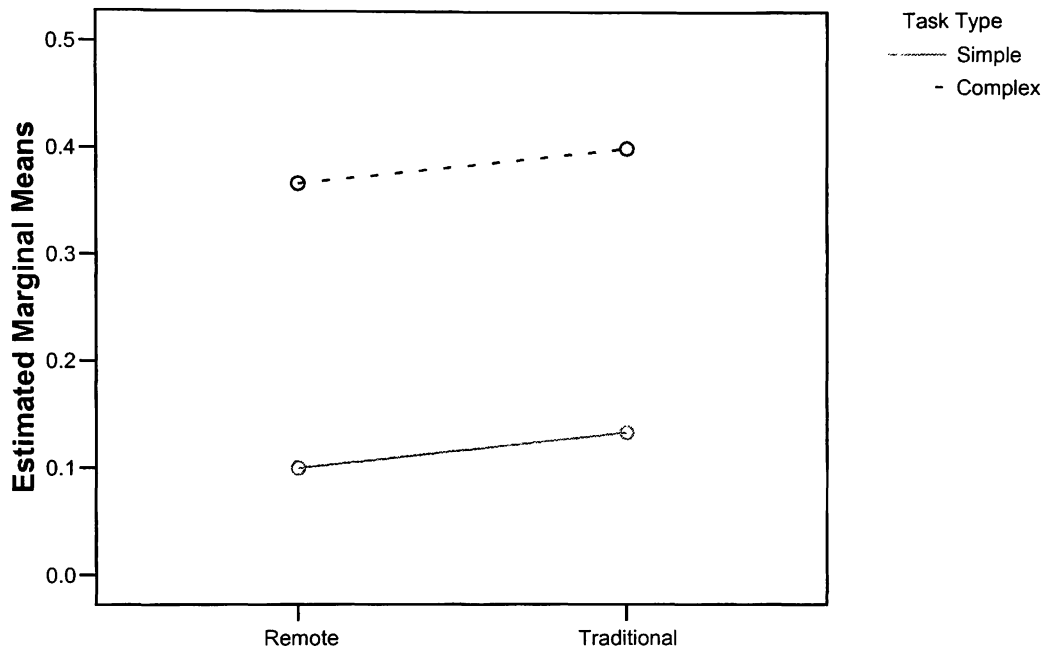


Figure 8. Number of critical incidents reported per task

Power for the within-subjects component was relatively high, measured at 0.88. Effect size was calculated to be 0.152. This value denotes a reasonably small overlap between population means.

The power measure for the between subjects comparisons was found to be much lower than the within subjects power measure. This value was 0.065; meaning that this study has a low probability of finding a statistically significant result if the proposed hypothesis is true. This was due to large overlap between population means, as shown by the low effect size measure of 0.002.

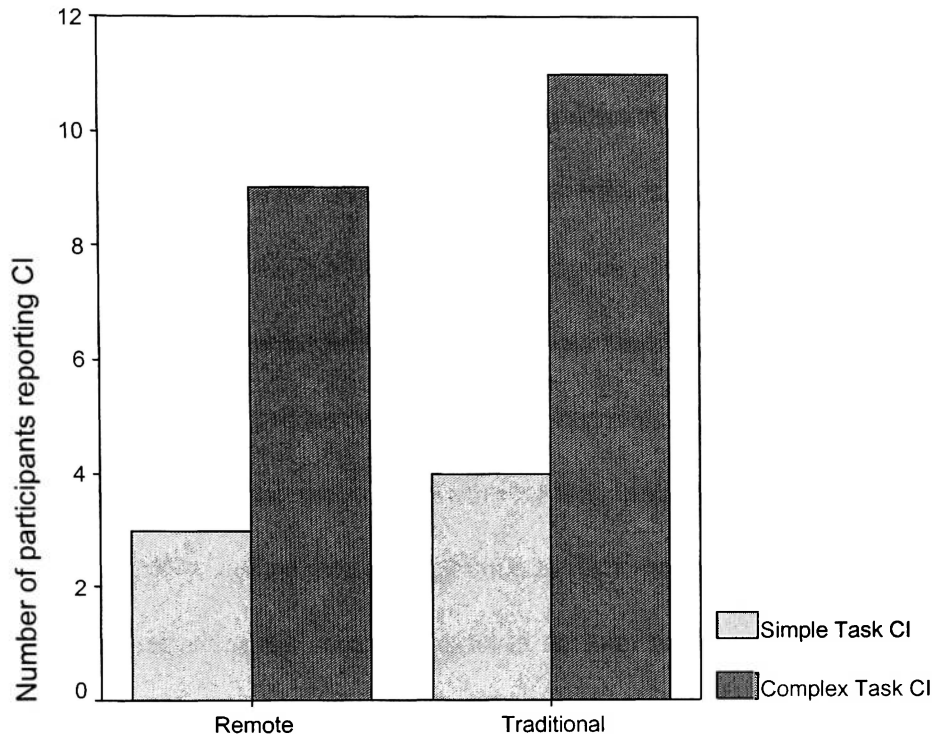


Figure 9. Number of participants reporting critical incidents per condition

Figure 9 shows the total number of participants reporting critical incidents. It also depicts how those numbers differ between the two test conditions.

### *Stress levels*

The third hypothesis stated that no significant differences would be found in user reported stress levels between the two test conditions. Participant stress levels were assessed using the STAI. An independent samples t-test was conducted to determine if differences exist between traditional and remote state anxiety scores. A t-test was used because state anxiety had no within-subjects component; it was strictly a between-subjects variable. Furthermore, state anxiety was interval data collected via survey, and this data type does not meet the requirements for running an ANOVA. The t-test results indicated no significant differences exist in state anxiety scores,  $t(58) = -0.044, p = 0.965$ .

In an effort to determine if the two test populations differed in trait anxiety, a t-test was conducted on these scores. An independent samples t-test found no significant differences in the trait scores attributable to testing location,  $t(58) = -1.277, p = 0.207$ .

User stress levels as measured by the STAI did not differ significantly between the two test conditions.

Confidence intervals for state and trait anxiety means were determined at the 95%

level. These values are reported below:

State Anxiety Remote	mean = 32.9 (95% CI	28.2	to	37.4)
State Anxiety Traditional	mean = 33.1 (95% CI	28.4	to	37.6)
Trait Anxiety Remote	mean = 34.7 (95% CI	27.6	to	36.3)
Trait Anxiety Traditional	mean = 37.6 (95% CI	26.4	to	39.1)

Participants reported both state and trait anxiety levels. State anxiety levels assessed anxiety immediately following the completion of both usability tasks. Trait anxiety scores were used in an effort to provide a baseline for participant anxiety. Trait anxiety scores were generally higher than state anxiety scores in all sixty participants. Figure 10 shows these differences in the scores. Both state and trait anxiety scores were only slightly higher for the traditional setting than the remote setting, however, this difference was non-significant as determined by the preceding t-tests.

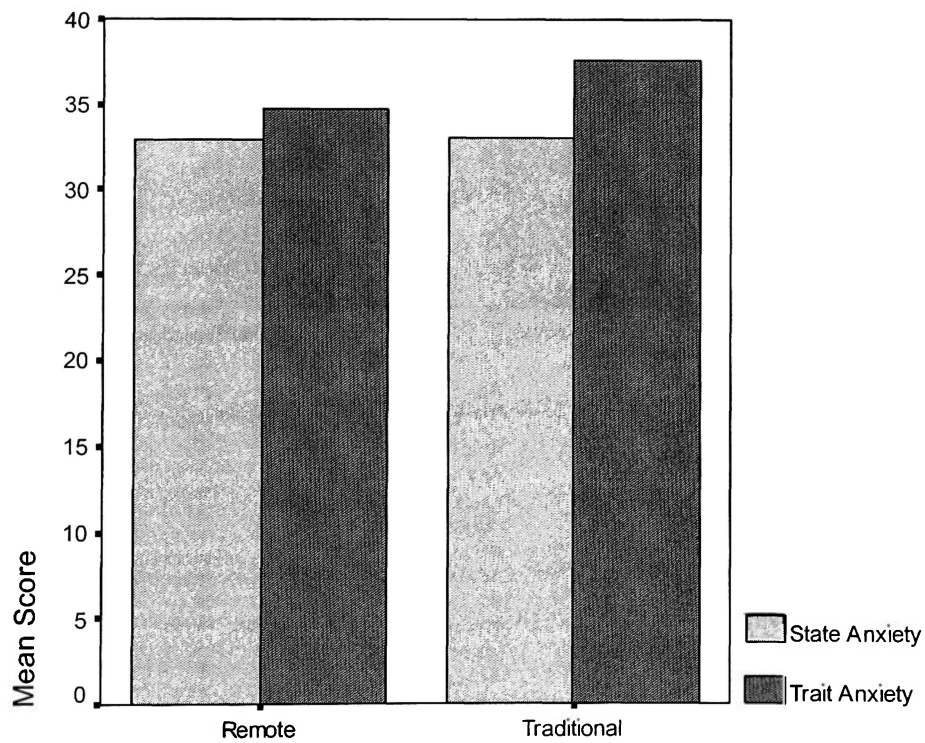


Figure 10. State and trait anxiety scores for both test conditions

*Subjective data*

Upon completion of the STAI, participants filled out a subjective appraisal of the website. Criteria evaluated included: logic in design, website appearance and content, and clarity of the experimenter's written instructions. Scores ranged from 1 to 5, with 1 meaning disagreement and 5 denoting an agreement. Figure 11 depicts these subjective ratings.

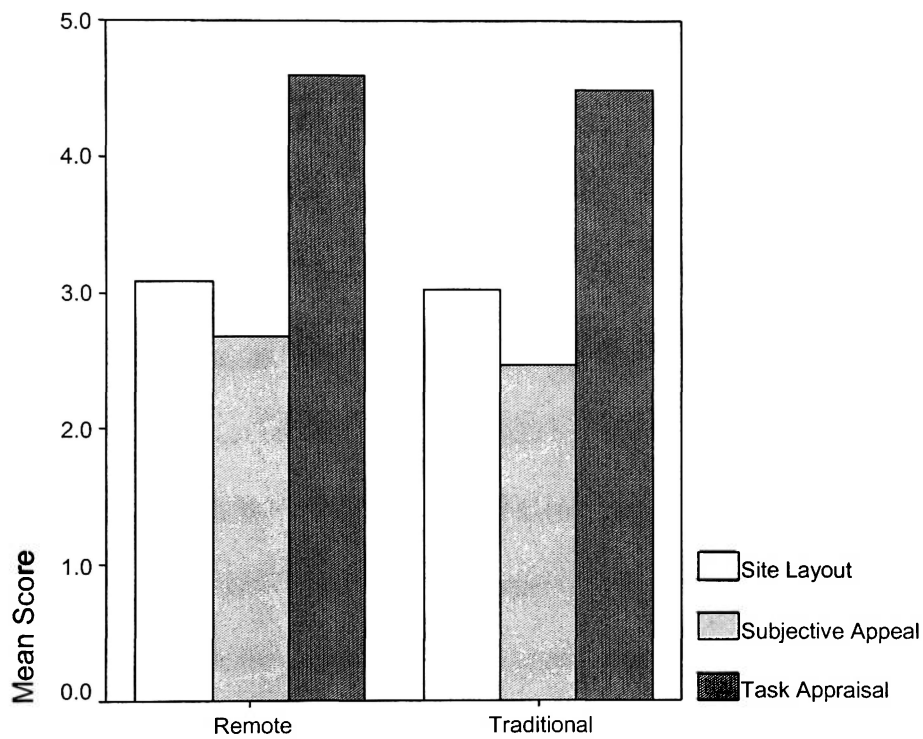


Figure 11. Remote and traditional subjective appraisals (higher scores are better)

Mann-Whitney nonparametric tests were conducted on the three subjective appraisals between the two test conditions. These tests were chosen because they do not require any assumptions about the test data; the distribution is unknown. No significant differences were found, all three appraisal score  $p$ -values were above the significance level of 0.05. Table 7 reports nonparametric test results for user subjective ratings.

Table 9.

*Nonparametric test results*

	Mann-Whitney U	$p$
Site Design	421.0	0.667
Subjective Appeal	399.5	0.452
Task Appraisal	398.0	0.415

*Summary*

From the results found from the univariate ANOVA concerning task times and how they relate to test location, the first null hypothesis has been rejected; testing location appeared to have a significant effect on participant task times. The ANOVA concerning number of critical incidents reported per task in each testing location yields inconclusive results, the second null hypothesis cannot be rejected. The t-test performed on the state anxiety scores for each test condition fails to find significant differences between those scores, and so the third null hypothesis is not rejected. Finally, nonparametric tests do not find any significant differences in subjective appraisals of website design, website appeal, and task clarity between the remote and traditional test conditions. Significant differences were found between task types; complex tasks took longer than simple tasks to complete, and increased numbers of critical incidents were reported during the complex task rather than simple task.



## DISCUSSION

The aim of this study was to find differences that testing location might have on usability test performance and user stress levels. Traditional usability testing is potentially expensive and more importantly may lack external validity due to the fashion in which it is conducted. Remote testing affords the recruitment and evaluation of many more varied participants especially when cost is a factor, and likely can test these participants within their natural work environments.

### *Major Findings*

Results from the study found significant differences in only average task time between remote and traditional laboratory testing. Power level for the between subjects time difference were calculated to be 0.52, rendering this claim on significance debatable. No significant differences were found in number of critical incidents reported. Likewise, no significant differences were found in subjective between-subjects appraisals of the site and task instruction clarity.

Mean times were only about 30-50 seconds longer for the remote task condition. As far as usability study performance is concerned, these additional times are not practically significant. Given proper training, participants perform very similarly in a remote testing condition as they do in a traditional setup. This suggests that usability test experience is remarkably similar between remote and traditional test conditions.

Results in this study were consistent with results reported in Tullis et al (2002) with the exception that user subjective ratings did not differ between remote and traditional settings in this study. Tullis found no significant difference between remote

and traditional task times. Another notable difference is that this study found a 100% completion rate for both tasks.

The study tested 60 Embry-Riddle students, both graduate and undergraduate. Participants came from differing majors, ranging from Aeronautical Science to Human Factors to Computer Engineering. Participants reported having a variety of computer experience, ranging from just under a year to over ten years of work with computers. Some students performing the study were rewarded with an entire letter grade in their class, while others received a few points on an exam, and still others received no academic extra credit at all. No participants received financial compensation of any sort.

Despite having such varied participant pool, times to complete the tasks were all rather similar, with the second task taking roughly twice as much time to complete as the first. The slightly longer task times found for the remote condition may be due to the user having to type out his or her difficulty to the moderator using AIM (AOL Instant Messenger, <http://www.aim.com/>). This act of switching to another application and typing may account for the additional 30-50 second differences in mean task times for the remote condition. However, the moderator could assist the user with the difficulty just as quickly in the remote condition as in the traditional condition

### *Internal Validity*

Participants were all recruited from Embry-Riddle's Daytona Beach campus. The between-subjects portion of the experiment was kept under tight control; participants were read identical instructions and not lead or given hints towards completing the tasks in any way. Sixty participants completed the experiment with similar results. The fact that there was little practical significance due to the effect of the treatment can be

attributed to the fact that there was likely no real effect. The large number of similar participants further supports the idea that no real practical effect existed. Significant results were likely attributable to individual differences in the between-subjects component rather than treatment effect.

### *Participant behavior*

Amazon.com uses a rather consistent interface: for example, purchasing functions are always located on the right side of the screen. However, participants did not catch on to this after the first task, and would look for “Add to shopping cart” buttons or other purchase-related functions near the image of the item being sold rather than the right side of the page.

The first task instructed participants to look for books in the “new jedi order” series. Upon finding such a book, participants had to add, and then remove the book from the shopping cart. Participants had difficulty in locating the “delete” button to remove an item from the shopping cart. This may be due to the experimenter’s instructions stating, “remove the item from your shopping cart;” participants may have expected a “remove” button. Participants were also confused by Amazon.com suggesting other items for purchase instead of taking them directly to the shopping cart when adding an item. The “View Cart” button is small and at the very top of the page, a place that most participants stated they would not think to look for. Many participants opted to click the “Edit your shopping cart” button found on the right side of the page.

The second task instructed participants to find a cell phone with a camera feature. Many participants would use keywords that would result in digital cameras being displayed, instead of cellular phones. Others chose “Electronics” from a pull-down menu

instead of “Cell Phones and Service.” Upon finding a cellular phone, many participants were not sure how to verify if it had a camera. Participants who owned camera phones or knew of camera phones opted to search by manufacturer, often shortening search times. The final part of the second task was to find a service plan that met certain criteria. Participants had difficulty locating the “Service Plans” link at the top of the screen, or did not know to input their zip code to check for phones and plans available in their area. Almost every participant clicked the “activation info” link that was near the image of the cell phone, expecting there to be service plan options listed there.

Common to both tasks was clicking the “Go” button after choosing a category without entering in any keywords. Upon doing so, participants were taken to a screen akin to an “advanced search.” Many just clicked the “back” button in their web browser to return to the previous search screen.

Participants may have behaved the way they did due to external motivation factors. Participants were informed that they were being watched in both test conditions, and this knowledge may have spurred certain behavior because motivation to complete the task came from this external “being watched” factor. Deci and Ryan (1985) report that individuals in this state perform not because of internal reasons, but to avoid an aversive external situation such as a reprimand or disapproval from the moderator. A state of introjected regulation (Deci & Ryan, 1985) may have occurred, where participants acted out of a sense of obligation. They wished to feel more confident performing the seemingly simple task of navigating a website.

### *Potential difficulties*

The differences found in task times due to complexity level may exist because the second task requires the user to interact with less-often used elements of Amazon.com, namely the cell phone marketplace for the second task. Many participants did not know that cell phones were available through Amazon.com. Few participants who were more familiar with the domain of cell phones would use specific manufacturer search keywords instead of the more common search terms such as “camera phone,” or “cell phone with camera.” Many participants were lead astray by a link entitled “Activation Info,” which did not lead to information that would allow them to complete the task.

Participants rarely reported critical incident data despite being trained and encouraged to report critical incidents. The experimenter repeatedly stressed that this study was an evaluation of the interface, and not the interface tester. Still, participants preferred to struggle through tasks and read sometimes irrelevant pages instead of reporting a difficulty with the interface. Many would scroll up and down excessively while reading non-pertinent information. A few participants would take as long as 14 minutes attempting to complete the second task and still not report difficulty or a critical incident. The evaluator never offered assistance until it was asked for by the participant, in an effort to keep the experiment consistent.

Computers used for the experiment differed in hardware configuration. The remote test computer was a Dell Optiplex 260 desktop computer with a 17” flat panel LCD display displaying a resolution of 1280 x 1024. The traditional computer was a Dell Dimension XPS R350 computer with a 17” CRT monitor displaying a resolution of 800 x 600.

Although the Amazon.com interface would adjust itself to accommodate both screen resolutions, the higher resolution setup offered a larger view of the interface. In addition, text would appear smaller on the higher resolution setup.

### *Study limitations*

The STAI may not have been a sensitive enough instrument to measure stress level differences that a usability study may elicit in participants. Strangely enough, the “everyday” trait anxiety score means were slightly but not statistically significantly higher than the state anxiety score means. These findings suggest that this particular usability study does not affect user stress or anxiety level. Traditional anxiety scores may have been higher if users’ performance was assessed during the test or if users were representing an organization when completing the usability study. Users had no extrinsic motivator such as a grade or ranking associated with their performance, and so may not have felt anxious during their usability tests. In real-world usability tests, a user may feel compelled to perform well; the user may feel that he or she is representing their organization or that somehow their job skills are being tested.

A better critical incident tool would have been desired. Other studies had a separate window showing at all times reminding the participant to report difficulties should they encounter them. In this study, the moderator would only record critical incidents if the participant reported having difficulty. In this study, the burden of finding and reporting critical incident data rests entirely with the participant. Capturing participant behavior through digital filming or screen captures would have allowed better critical incident reporting by the moderator.

The study's between-subject findings have a low power associated with them. This low power weakens the ability to make the claim that participant times differed significantly between testing locations. One method to increase power would be to facilitate increasing the effect size. Adding additional participants is not likely going to increase power, as the sixty tested so far performed rather similarly. Differences in between-subject task times were not very large, especially in a practical context. To say that usability study task times differed about thirty seconds to a minute is not a major issue in terms of practical usability test performance.

### *Practical Implications*

From the data found in this study, it appears that usability testing carried out in remote and traditional laboratory setups does not change the user experience significantly. Practically speaking, results from remote usability testing provide usability information as robust as traditional laboratory results. When recruiting users for a study, the compensation and travel expenses of those users may become a factor for the usability test administrator. If remote usability testing methods do not incur such expenses, yet provide similar desirable results, it appears that remote testing is a favorable alternative.

The results from this study failed to find differences in subjectively assessed user stress levels between the testing locations. This suggests that the corporeal presence of a passive moderator does not affect user anxiety levels as they perform the usability tasks. In an external real-world setting, a passive experimenter present in the room with users participating in a usability study will not affect their performance significantly according to this study's findings.

### *Future research*

In future iterations of the study, a better critical incident report tool should be used. In addition, perhaps a user frustration or other method of capturing user psychological state should be used. The study design can be altered to make the task complexity variable between subjects as well, to ascertain what types of tasks can cause user anxiety or frustration. Participants recruited were very similar; all participants were associated with Embry-Riddle either as students or as full-time employees.

If users are reluctant to report critical incident data, perhaps the moderator can collect such data in future research. Participants behaved in very similar ways in both conditions, and to an individual familiar with how to complete the task, ascertaining when user actions did not meet user expectations is very apparent.

Other interfaces may be tested as well; perhaps a software interface rather than a website will yield different results. An alternative method that automates data collection and therefore separating the moderator in time as well as space from the participant would be an interesting focus for future research.



## CONCLUSION

Software and website usability testing are becoming more frequent, necessary, and critical to the success of software or websites as their respective complexity increases. The proliferation of the internet and widespread integration of computers into everyday life secures a demand for usable interfaces. Ever-increasing complexity demands a usable interface. Usability studies continue to be the best method for finding potential issues with the interface; heuristics and expert design cannot hope to identify every interface problem, especially for novel interfaces and designs.

Usability studies work best when participants are varied in their makeup and plentiful in numbers. Recruiting varied users becomes simpler when users do not have to travel to a testing location. Remote testing allows a separation of user and examiner in space and potentially time. In today's high-speed networked world, remote software testing is potentially far less expensive than traditional laboratory testing, provided the user experience with remote testing yields good usability data, that is, data that uncover potential usability issues. Lower costs associated with testing allow more participants to be recruited, and these participants are potentially more varied as they can be recruited from more distant locations. Remote testing also stands to gain additional benefits if data collection can be automated, relieving the experimenter of time commitments to participants, and therefore increasing the amount of data collected. Such automation can reduce both time required and costs of usability testing, encouraging usability testing use, and therefore affording the creation of a more usable piece of software or website.

The results from this study suggest that remote and traditional laboratory testing do not differ practically in terms of participant task times and critical incident reporting.

Although between-subject task times differed significantly, in a practical setting the difference of about thirty seconds is negligible. Both test settings appear to capture usability issues, especially if the moderator can observe trends in participant behavior such as excessive scrolling or site navigation through the repeated use of the “back button.” Both test conditions seemed to reliably capture these difficulties. In an ideal testing scenario, both remote and traditional laboratory testing would be utilized during the design process.

## REFERENCES

- Armstrong, S. D.; Brewer, W. C.; and Steinberg, R. K. (2002). Usability Testing. In Charlton, Samuel G. and O'Brien, Thomas G. (Eds.). *Handbook of Human Factors Testing and Evaluation (2<sup>nd</sup> Ed)*: Lawrence Erlbaum Associates, Publishers. (pp. 403-432) Mahwah, New Jersey
- Byström, K. & Järvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing and Management*. Vol 31, 2. (pp. 191-213).
- Campbell, D. (1988). Task complexity: a review and analysis. *Academy of Management Review*. Vol 13. pp. 40-52.
- Castillo, J.C. (1997). *The user-reported critical incident method for remote usability evaluation*. Unpublished thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA.
- Castillo, J. C., Hartson, H. R., & Hix, D. (1997). *The user-reported critical incident method at a glance*. Department of Computer Science Technical Report TR-97-13. Blacksburg, VA: Virginia Tech.
- Castillo, J.C., Hartson, H.R., and Hix, D. (1998). Remote Usability Evaluation: Can Users Report Their Own Critical Incidents? *Summary of CHI'98 Human Factors in Computing Systems*. (pp. 253-254).
- Charlton, S.G. & O'Brien, T.G. (2002). *Handbook of Human Factors Testing and Evaluation. 2<sup>nd</sup> Edition*. Lawrence Erlbaum Associates, Mahwah, NJ.
- Deci, E.L. & Ryan, R.M. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior*, New York, Plenum Press
- Dorazio, P., & Stovall, J. (1997). Research in context: ethnographic usability. *Journal of Technical Writing and Communication*, Vol 27, Issue (1), pp. 57-67
- Dumas J.S. & Redish, J.C. (1993). *A Practical Guide to Usability Testing*. Norwood, NJ: Ablex Publishing Corporation.
- Fitts, P.M. & Jones, R.E. (1947). Psychological Aspects of Instrumented Display: Analysis of 270 "Pilot Error" Experiences in Reading and Interpreting Aircraft Instruments. In H.W. Sinaiko (Ed.), *Selected papers on human factors in the design and use of control systems* (pp. 1-38). New York: Dover Publications, Inc.
- Flanagan, J.C. (1954). The critical incident technique. *Psychological Bulletin*, (July), 51 (4), pp. 326-358

- Flanders, V. (2004). "What would amazon.com do?" Article on website.  
<http://www.websitethatsuck.com/>
- Friedman, H.S. (1992). Understanding hostility, coping, and health. In H.S. Friedman (ed.) *Hostility, Coping, and Health*.(pp.3-9). Washington, DC: American Psychological Association.
- Gray, B., Barfield, W., Haselkorn, M., & Spyridakis, J. (1990). The Design of a Graphics-based Traffic Information System Based on User Requirements, *Proceedings of the 34th Annual Human Factors Society Meeting*, Boston, 603-606.
- Guest, M.A. (2001). Assessment of time stress effects on decision accuracy at multiple phases of skill acquisition. *Dissertations Abstracts International*. 61(10-B): 5564
- Hammontree, M. Weiler, P., and Nayak, N. (1994). *Remote usability testing*. Interactions, (July), 21-25
- Institute of Electrical and Electronics Engineers. *IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries*. New York, NY: 1990.
- Lazarus, R. S. (1981). The stress and coping paradigm. In C. Eisdorfer, D. Cohen, & P. Marim (Eds.), *Models for clinical psychopathology*. New York: Spectrum.
- Liew, W.C. (2003). *Website usability testing, remote vs. traditional*. Master's Thesis, University of Nebraska, Lincoln, Nebraska.
- Nielsen, J. (1997). Usability Testing. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics, 2<sup>nd</sup> Edition*. (pp. 1543-1568). John Wiley & Sons, New York
- Nielsen, J. (2000). Why you only need to test with 5 users. Alertbox (<http://www.alerbox.com/>), March 19, 2000.
- Nielsen, J., and Mack, R.L. (1994). *Usability Inspection Methods*. John Wiley & Sons, New York, NY.
- Miller, D. (1981). The depth/breadth tradeoff in hierarchical computer menus. *Proceedings of 25<sup>th</sup> Annual Meeting - Human Factors Society*, NY, 296-300.
- Miller, K. (1963). The concept of crisis. *Human Organization*. 22, 195-201.
- Rubin, J. (1994). *Handbook of Usability Testing*. John Wiley & Sons, New York, NY.
- Sarason, I. G. (1984). Stress, anxiety, and cognitive interference: Reactions to Tests. *Journal of Personality and Social Psychology*, 46, 929-938.

- Schaab, B. (1997). The influence of time pressure and information load on rule-based decision making performance. *Dissertation Abstract International*. 58(5-B): 2713
- Speilberger, C.D. (1968). State trait anxiety inventory. Consulting Psychological Press Inc.
- Speilberger, C.D., Gorsuch, R.L., & Lushene, R.E.(1969). The state trait anxiety inventory manual. Palo Alto: Consulting Psychologists Press.
- Tullis, T.S., Fleischman, S., McNulty, M., Cianchette, C., and Bergel, M. (2002). *An empirical comparison of lab and remote usability testing of web sites*. Usability Professionals Association Conference, July 2002, Orlando, FL.
- Yerkes, R.M., Dodson, J.D. (1908). *The relation of strength of stimulus to rapidity of habit formation*. *Journal of Comparative and Neurological Psychology*. Vol 18, pp. 459-482

APPENDIX

**Task Complexity Questionnaire**

*Please circle the number that corresponds with your assessment on the question:*

During your completion of this task:

1. Did you feel that there were multiple possible ways to complete it?

Few ways      | 1      | 2      | 3      | 4      | 5      | 6      | 7      | More ways

2. Did you feel that the task had multiple goals?

One goal      | 1      | 2      | 3      | 4      | 5      | 6      | 7      | Multiple goals

3. Did the methods you used to complete the task depend on one another (i.e. you had to finish one step completely before starting another)?

Did not depend    | 1      | 2      | 3      | 4      | 5      | 6      | 7      | Did depend

4. Did you feel the methods used to complete the task conflicted with one another (i.e. one step seemed to interfere with completion of another step)?

Did not conflict    | 1      | 2      | 3      | 4      | 5      | 6      | 7      | Did conflict

5. Where you unsure of how to start the task?

Sure              | 1      | 2      | 3      | 4      | 5      | 6      | 7      | Unsure

6. Overall, rate how complex you thought this task was:

Simpler            | 1      | 2      | 3      | 4      | 5      | 6      | 7      | Complex

## **Informed Consent Form**

---

### INFORMED CONSENT FORM

#### **Project Identification**

To compare the effect of testing location (either remote testing or traditional laboratory testing) on usability study performance and user stress level.

#### **Purpose of Research**

You are hereby invited to participate in a usability evaluation of Amazon.com. You will be evaluating the interface used by Amazon.com by performing common tasks that your everyday Amazon.com user would perform. You may either be assigned to complete the experiment in a remote setting away from the evaluator or you may stay in the traditional laboratory testing environment, complete with moderator. It will be stressed repeatedly during this study that we are not testing your computer skills or familiarity with Amazon.com; we are testing ideas for improving the interface of Amazon.com as well as improving the act of testing such interfaces. You will be trained in critical incident reporting, so that you may report any difficulties you may experience using the Amazon.com interface.

#### *Remote Testing Situation*

You will be asked to follow instructions that will require you to navigate Amazon.com. No actual purchases or transactions will be made during the normal course of this study. Two different tasks will be required of you, if you recognize a difficulty in using the interface, or completing the task, we ask that you report the difficulty and its severity to the moderator. You will complete this part of the experiment in one of the campus computer labs; however you will still be able to communicate with the moderator through a chat program such as MSN or AIM. Also in the remote condition the moderator will be able to view your screen to track your progress, as well as assume control over your computer should you require assistance in completing these tasks.

#### *Traditional Testing Situation*

You will stay here in the lab and follow the same instructions as the remote condition, navigating Amazon.com. You are free to ask questions of the moderator and ask for assistance, but if you encounter interface difficulties it is asked that you report them using the report tool.

Upon completing either condition you will fill out a questionnaire describing your satisfaction with the site, as well as a stress-assessment questionnaire.

### **Risks and/or Discomforts**

There are no known perceived risks to you or your property as a result of participating in this research.

### **Confidentiality**

Any identifying information is kept confidential and not released when reporting results. Results from this experiment that may be published are published as aggregated data; no personal information is released.

### **Additional/further information**

You may feel free to ask questions of the experimenter and moderator at any time prior to, during, and after the experiment. You will be debriefed following the actual experiment. Furthermore, you may request additional information or an electronic copy of the study by corresponding with [andr635@erau.edu](mailto:andr635@erau.edu).

### **Freedom to withdraw**

You are not obligated to finish the entire experiment; your participation is voluntary, you may leave at any time. Your relationships with ERAU faculty, staff, and students are not affected by prematurely terminating your involvement with this study.

---

**Signature of Participant**

---

**Study Date:**



## User Training Sheet

Thank you for participating in this study. Please remember that *we are not testing you or your computer-related abilities*, we are testing the website interfaces you will be working with. Please take a moment to understand this; if you encounter difficulty, this is not an indicator of a lack of knowledge or ability on your behalf. We are looking for inconsistencies and difficulties that the websites may contain.

### *The User-Reported Critical Incident Method: Participant Briefing*

You will be using the Critical Incident Method to inform us of design issues with the websites. A critical incident is best defined by an example:

You find yourself wishing to enter the Lehman building. There is a set of double doors in front of you; you pull on the right one, finding it to be locked. The left one opens however. A critical incident in door opening has just occurred. In the context of door opening, an event happened that affected the failure or success of your goal: entering the Lehman building. In this case, the right locked door delayed (and perhaps annoyed you) your entrance.

The preceding example is trivial, perhaps, but either way an event occurred that directly affected your success or failure with a task. Designers can use such critical incident data to improve doors in the future.

These are the kind of problems we would like you to point out with websites in the experiment you are about to begin. Whenever you encounter a difficulty or issue with the performance of a task, please report it (instructions on how to do this will follow). *Please remember we are NOT testing you or your computer knowledge*; please report any possible difficulties you may encounter. At the same time, we do not require you to offer suggestions on how to fix the problem, please tell us what your difficulty is, as we are not always clear what the problem is if you offer just the suggestion on how to fix it.

Good Report Example: I am unable to find the “Submit” button.

Bad Report Example: Make the “Submit” button larger and make it flash rainbow colors.

**After you encounter a critical incident and document what it is, please rate it on a severity scale from 1 – 5 where 1 is “this is annoying” and 5 is “this kept me from completing the task.”**

As always the experimenters thank you for your time and participation in this experiment.

You will now be assigned two tasks; each task will have you following instructions. One task is assigned per sheet; each task has a few steps associated with it.

## User Profile Questionnaire

1. Please circle the choice that best describes how long you have been using computers:

1 year or less       5 years or less       10 years or less       More than 10 years

2. How often on average, during the academic year, do you use on-campus computer labs per week (C-Lab, LB 171, 172, etc)?

0 hours       ~2 hours       ~5 hours       ~10 hours (or more)

3. Please check all operating systems you are familiar with:

<input type="checkbox"/> Windows (1.0 thru 3.14)	<input type="checkbox"/> OS/2
<input type="checkbox"/> Windows 9x (includes Windows ME)	<input type="checkbox"/> Linux
<input type="checkbox"/> Windows 2000/XP	<input type="checkbox"/> UNIX (includes X-Windows)
<input type="checkbox"/> Windows CE	<input type="checkbox"/> Palm OS
<input type="checkbox"/> Mac OS (1.0 – 9.22)	<input type="checkbox"/> DOS (1.0 – 6.2)
<input type="checkbox"/> Mac OS X (Jaguar and Panther)	<input type="checkbox"/> Other: _____
<input type="checkbox"/> Don't Know/Not Sure (of any of these choices)	

4. Please check any and all application types you are familiar with:

<input type="checkbox"/> Word Processing	<input type="checkbox"/> Internet (web browser, FTP)
<input type="checkbox"/> Spreadsheets	<input type="checkbox"/> CAD/CAM
<input type="checkbox"/> Programming/ Coding (IDE environments)	<input type="checkbox"/> Animation (3D or 2D)
<input type="checkbox"/> Databases	<input type="checkbox"/> Drawing/Photo editing (Photoshop)
<input type="checkbox"/> Don't Know/Not Sure (of any of these choices)	

5. Have you ever Beta-tested any software?

Yes, please briefly specify \_\_\_\_\_  
 No  
 Don't Know/Not Sure (what this means)

6. Do you shop online (eBay, online merchants i.e. GAP.com)?

Yes  
 No

7. Do you create web pages (either by coding, software, or other means i.e. Xanga.com, Yahoo Page Builder)?

Yes  
 No

## User Task Instruction Sheet – Task Rogue

Participant ID: \_\_\_\_\_

Usability Process Date: \_\_\_\_\_

Website: http://www.amazon.com/

Please answer questions by writing your answer in the provided boxes.

Please open a new browser window (if one is not already open and navigate to <http://www.amazon.com/>).

*Step 1:*

- Choose “Books” under the “All Products” pull-down menu.

*Step 2:*

- In the search field, enter “new jedi order”

1. How many **total** results are returned (not Most Popular Searches)?

*Step 3.*

- Click on any one of the books you found. Add it to your “Shopping Cart.”

*Step 4.*

- View your shopping cart; now remove the item from the shopping cart.

Did you receive the message “Your Shopping Cart is empty.”?

Your first task is now complete; please proceed to the next page where you will be presented with the next task.

## User Task Instruction Sheet – Task Twin Suns

Participant ID: \_\_\_\_\_

Usability Process Date: \_\_\_\_\_

Website: <http://www.amazon.com/>

Please answer questions by writing your answer in the provided boxes.

Please open a new browser window (if one is not already open and navigate to <http://www.amazon.com/>).

### *Step 1:*

- Using the search function under the “All Products” pull-down menu, find a cell phone that has a camera feature.

Pick one of the choices that result, making sure it is a cell phone with a camera feature. Which camera phone model did you choose?

How much does it cost?

### *Step 2.*

- From the page depicting the camera cell phone, please find a service plan.
- The plan must include:
  - More than 500 Anytime Minutes Per Month
  - Have Unlimited Nights and Weekends
  - Have an activation fee LESS than \$36.00

Which plan did you choose?

How much is it per month?

You have completed this task. Thank you for your participation, please move on to fill out the post-evaluation questionnaires.

**User Reported Critical Incident Description Sheet (to be filled out by moderator)**

Participant ID: \_\_\_\_\_

Usability Process Date: \_\_\_\_\_

Website: http://www.amazon.com/

Please account for any critical incidents you encounter according to your previous training. Please do not feel that you have to fill out every single entry. More entries will be provided should you require them.

On the severity scale, please place an X in the gray boxes below that reflect the severity of the incident. Severity level 1 reflects a minor annoyance whereas Severity level 5 meant the incident kept you from completing the task.

*Critical Incident Report*

Please provide a brief description of the incident:

Severity Level 1    Severity Level 2    Severity Level 3    Severity Level 4    Severity Level 5

*Critical Incident Report*

Please provide a brief description of the incident:

Severity Level 1    Severity Level 2    Severity Level 3    Severity Level 4    Severity Level 5

*Critical Incident Report*

Please provide a brief description of the incident:

Severity Level 1    Severity Level 2    Severity Level 3    Severity Level 4    Severity Level 5

*Critical Incident Report*

Please provide a brief description of the incident:

Severity Level 1   Severity Level 2   Severity Level 3   Severity Level 4   Severity Level 5

*Critical Incident Report*

Please provide a brief description of the incident:

Severity Level 1   Severity Level 2   Severity Level 3   Severity Level 4   Severity Level 5

*Critical Incident Report*

Please provide a brief description of the incident:

Severity Level 1   Severity Level 2   Severity Level 3   Severity Level 4   Severity Level 5

*Critical Incident Report*

Please provide a brief description of the incident:

Severity Level 1   Severity Level 2   Severity Level 3   Severity Level 4   Severity Level 5

**Description:**

STAI - State Anxiety Form

Directions: A number of statements which people have used to describe themselves are given below. Read each statement and then circle the appropriate number to the right of the statement to indicate how you feel *right* now, that is, *at this moment*. There are no right or wrong answers. Do not spend too much time on any one statement but give the answer which seems to describe your present feelings best.

	NOT AT ALL	SOMEWHAT	MODERATELY SO	VERY MUCH SO
1. I feel calm	1	2	3	4
2. I feel secure	1	2	3	4
3. I am tense	1	2	3	4
4. I feel strained	1	2	3	4
5. I feel at ease	1	2	3	4
6. I feel upset	1	2	3	4
7. I am presently worrying over possible misfortunes	1	2	3	4
8. I feel satisfied	1	2	3	4
9. I feel frightened	1	2	3	4
10. I feel comfortable	1	2	3	4
11. I feel self-confident	1	2	3	4
12. I feel nervous	1	2	3	4
13. I am jittery	1	2	3	4
14. I feel indecisive	1	2	3	4
15. I am relaxed	1	2	3	4
16. I feel content	1	2	3	4
17. I am worried	1	2	3	4
18. I feel confused	1	2	3	4
19. I feel steady	1	2	3	4
20. I feel pleasant	1	2	3	4

- STAI – Trait Anxiety Form

Directions: A number of statements which people have used to describe themselves are given below. Read each statement and then circle the appropriate number to the right of the statement to indicate how you *generally* feel.

	ALMOST NEVER	SOMETIMES	OFTEN	ALMOST ALWAYS
21. I feel pleasant	1	2	3	4
22. I feel nervous and restless	1	2	3	4
23. I feel satisfied with myself	1	2	3	4
24. I wish I could be as happy as others seem to be	1	2	3	4
25. I feel like a failure	1	2	3	4
26. I feel rested	1	2	3	4
27. I am "calm, cool and collected"	1	2	3	4
28. I feel that difficulties are piling up so that I cannot overcome them	1	2	3	4
29. I worry too much over something that doesn't really matter	1	2	3	4
30. I am happy	1	2	3	4
31. I have disturbing thoughts	1	2	3	4
32. I lack self-confidence	1	2	3	4
33. I feel secure	1	2	3	4
34. I make decisions easily	1	2	3	4
35. I feel inadequate	1	2	3	4
36. I am content	1	2	3	4
37. some unimportant thought runs through my mind and bothers me	1	2	3	4
38. I take disappointments so keenly that I can't put them out of my mind	1	2	3	4
39. I am a steady person	1	2	3	4
40. I get in a state of tension or turmoil as I think over my recent concerns and interests	1	2	3	4



## User Site Structure Post-Evaluation Questionnaire

Please circle the number on the interval scale that matches your experiences with the websites that you used:

### 1. Site Design/Layout

Overall, I felt the design of website was easy to understand      Disagree    1    2    3    4    5    Agree

The site's design made the instructions easy to follow      Disagree    1    2    3    4    5    Agree

Links were labeled so I knew what was on the target page before I used the link      Disagree    1    2    3    4    5    Agree

I had a pleasurable experience using this site, I didn't feel frustrated      Disagree    1    2    3    4    5    Agree

### 2. Subjective Appraisal

I found the site design attractive      Disagree    1    2    3    4    5    Agree

I would use the same images and colors if I made this site      Disagree    1    2    3    4    5    Agree

I would arrange a website this way if I made one      Disagree    1    2    3    4    5    Agree

### 3. Task Appraisal

The instructions given to me were clear      Disagree    1    2    3    4    5    Agree

Understanding the instructions was easy      Disagree    1    2    3    4    5    Agree

I knew what the instructions wanted me to do      Disagree    1    2    3    4    5    Agree