2004

A Study of Fidelity Level and Task/Non-Task Based Testing Scenarios on the Effectiveness of Usability Testing

Joshua R. Dolecal
Embry-Riddle Aeronautical University - Daytona Beach

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

Part of the Ergonomics Commons

A Study of Fidelity Level and Task/Non-Task Based Testing

Scenarios on the Effectiveness of Usability Testing

Joshua R. Dolecal

M.S. Human Factors & Systems

Embry-Riddle Aeronautical University

2004

UMI Number: EP32063

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®

A Study of Fidelity Level and Task/Non-Task Based Testing
Scenarios on the Effectiveness of Usability Testing

by
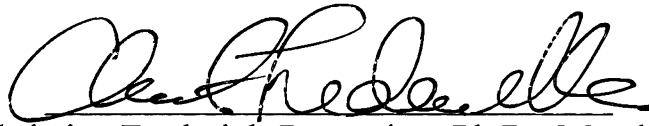
Joshua R. Dolecal

This thesis was prepared under the direction of the candidate's thesis
committee chair, Dahai Liu, Ph.D., 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 and Systems.

THESIS COMMITTEE:

_____
Dahai Liu, Ph.D., Chair

_____
Christina Frederick-Recascino, Ph.D., Member

_____
Remzi Seker, Ph.D., Member

_____
MS HFS Program Coordinator

_____
Department Chair, Department of Human Factors and Systems

_____
Associate Chancellor for Academic Affairs

Abstract

Author: Joshua R. Dolecal

Title: A Study of Fidelity Level and Task/Non-Task Based Testing Scenarios on the
Effectiveness of Usability Testing

Institution: Embry-Riddle Aeronautical University
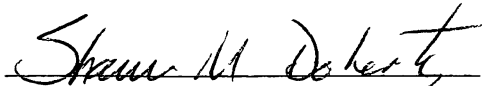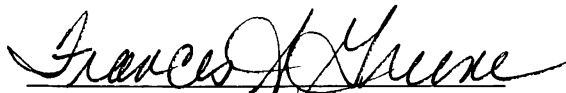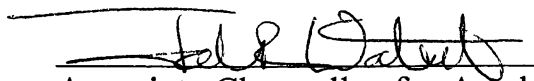
Degree: Masters in Human Factors and Systems Engineering

Year: 2004

Advances in usability are directing developers towards creating a better and all around friendlier environment for users. Unlike when buying a physical product where you purchase it first and then experience its usability, web site usability is immediately present. So, if a site doesn't meet our needs, we as users, are quick to look elsewhere. Some research has suggested that fidelity makes no significant impact on users' ability to locate errors or problems in a web site. This paper is intends to investigate the interaction between fidelity and task vs. non-task user testing with regards to the types of problems experienced by users. The goal is to identify the most efficient types of user testing. User testing is conventionally designed to emulate typical user situations and tasks. However the goal of testing is to find all possible problems with the interface's design and correct or improve them. Asking users to perform typical tasks may not uncover all of these problems. Asking users to perform a non-task or browse the interface may provide additional information. And the interaction between fidelity and task should suggest that

both types of tasks are needed at varying degrees of fidelity to ensure a quality user

interface.

Table of Contents

List of Tables

List of Figures

Introduction

The Web (WWW, Internet, Online, Cyberspace and Information Super Highway) is changing the way business and society functions. Friends now communicate over the internet while eating or writing a paper. The news, movie reviews and even shopping can be accomplished online. To take advantage of this shift organizations now display their products online. This can be done solely, or in supplement to their physical organization. No matter why we have decided to come online, we, as users, have come to expect certain things. This is the idea behind web usability. It is of great benefit to acquire large quantities of information without having to leave the comfort of your chair, but that alone is not enough. As web sites improve we want to see that our online experience becomes increasingly easy. Users expect to find information with little to no effort, and that information to be presented in such a way that it very quickly conveys a plethora of information. Many web developers have made the mistake of designing a web site much like a book or other written publication and this must change.

The principles of usability and systems engineering dictate that increased effort (up to a certain level) towards ensuring users' needs will shorten the product development time, cut costs, and create a better product. Research has been done that suggests that fidelity has little or no impact on a users' overall ability to find problems

with a web site (Walker, Takayama & Landay, 2001). But according to some usability professionals (Newman & Landay, 2000), the bulk of usability testing and evaluation is focused on information design and navigation design, while graphic design is left until the end or outsourced. What this means is that more focus and attention is placed on how users navigate and the global package of an interface, but the small details are overlooked. Graphic design refers to the smaller more detail oriented part of an interface such as colors used, the words used (terminology or clever phrases), and how well this information is outlined to convey information to the reader. If a different approach to user testing is taken such as using a non-task or a browsing task, it may be possible to capture the usability of this information and improve design.

This paper intends to study the interaction between fidelity and testing type (task-based vs. browsing). Many user tests involve giving users a typical task to perform and asking what problems they found, while an experimenter observes their behavior. This technique may be producing a psychological tunneling effect, such that researchers' focus is on the task, so their ability to find problems is structured around the task. If users are not given a specific task, and just asked what problems they see in a web site, users may be free to find problems that they would otherwise have not found. This is often done in industry with the use of heuristic evaluation (Nielson, 1994). Heuristic evaluation makes use of professionals or experts in usability. The suggestion of using a non-task or browsing user test is meant to allow the user this same freedom in identifying problems in an interface's design.

People have often found problems with various websites such that a function may not work as they expected, or they interpret things differently than how web developers

intend. A task test may overlook these flaws because a user is focused on the task, and

will notice problems they have while performing that task. But without a specified task,

the user may notice problems of a different sort. This makes clear the fact that we still do

not fully understand user's mental model while using a specific interface. This fuzziness

is in part due to the fact that software companies use many different interface styles and

they try to accommodate wide varieties of users. Some interfaces are designed for

specific people with a certain degree of training or education, but there is a lot of

software and interfaces that are meant to be used by everyone. This fact makes it difficult

for developers to create an ideal interface because the only real similarities between users

are that they are human, and can speak and read a specific language. However this is not

always the case as some designs even allow for handicapped individuals, or are designed

such that the user only needs to identify pictures instead of words. The freedom of a non-

task to browsing user test may allow developers to obtain more information about users

that we had not considered. Users may actually be more likely to find graphic design

problems in this case. If this is true, then it is important that both task and non-task based

tests with varying degrees of fidelity be performed on users specifically in the early

stages of web development.

<div align="center">Review of Literature</div>

Software Usability: What is Usability?

Usability has gone by many names (Nielson, 2000) such as: Usability, Usability

Engineering, Human Factors, Ergonomics, and User Centered Design (UCD) among

many others. Ergonomics and Human Factors have been popular terms in Europe for

usability, and the Human Factors and Ergonomics Society in the U.S. only recently added

the word ergonomics to its name. These names are all interchangeable and focus on the fact that the human or user is an essential part of a system; such that they should be considered in a system's design.

Woodson (1981) defines UCD as "the practice of designing products so that users can perform required use, operation, service, and supportive tasks with a minimum of stress and maximum of efficiency." Neilson (1993) describes Usability as the overall acceptability of a system. First a system must be socially acceptable. The new technology or product being used must be something that the public or users want, and how they want it. Examples of this are when the Glass Cockpits and Automation was introduced to the field of aviation. Older pilots were not accustomed to the technology and resisted it. Automation also began to make the pilot more passive than active. That resulted in a lot of accidents.

The practical characteristics of Usability that Neilson (1993) outlines are:

- Learnability
- Efficiency
- Memorability
- Errors
- Satisfaction.

Learnability refers to how easy a system is to learn and use. Some systems like an ATM (Automatic Teller Machine) have no learning curve. A user can almost instantly figure out how to use it. However many systems require time to learn. This learning process is different for different users. Three main classes of users, according to Nielson (1993) are the novice user, casual user, and expert user. A novice has little experience with the

particular system or those that are similar to it. This learning curve (see Figure 1)

proceeds with a rapid initial learning phase which eventually plateaus. The expert user

can have experience with the system or other systems that are similar to it. Their learning

curve (see Figure 1) shows a slower rate of learning. However experts plateau at a higher

efficiency level than the novice users. This difference is due to the fact that the user is

knowledgeable about other similar
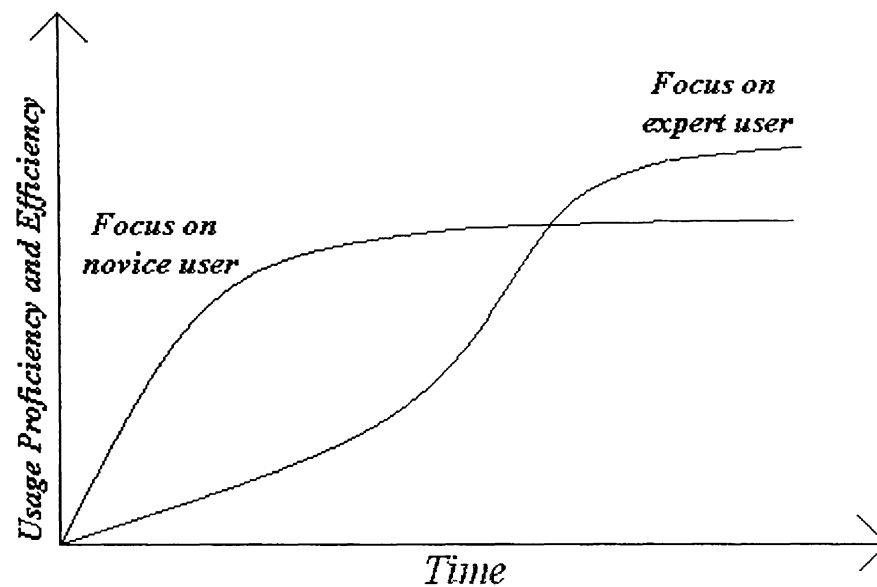
Figure 1 User Learning Curve from Nielson (1993)

systems. Expert learning involves the transfer of skills from one domain or system to the

next. The casual user is different from the other two, but shares some common

characteristics. A casuals user's experience can range anywhere from a novice to an

expert. When they do use a system they use it for a certain amount of time and then stop

using it for a while. It is this lack of prolonged use with a system that separates them

from expert users. This pattern also has an effect on their learning curve such that it is

different from a novice or expert's learning curve. However for practical purposes this

casual user is usually lumped in one of the other two categories based on their

experience. McCracken and Wolfe (2004) mention four user classes based on their behavior. Novices lack experience and knowledge of a system and generally have a fear of making mistakes. Advanced beginners are a little more knowledgeable than novices and no longer fear making mistakes. They feel more comfortable exploring the system. Competent performers are the next highest level. They can diagnose any problems found and perform long and complicated series of tasks. Lastly, experts can diagnose problems and perform tasks at a higher level of difficulty than competent performers. Experts are also interested in the system, and are not simply focused on accomplishing a task or goal. So looking at McCracken and Wolfe (2004) as well as Nielson's (1993) descriptions of users it is clear that we have yet to understand what various categories of users there are. If we can someday categorize them all, that will be a key component to understanding users' mental models, while interacting with various systems and interfaces.

Efficiency refers to user performance on the learning curve once it plateaus and they feel they have learned enough. At this point the user is very knowledgeable about the system with regards to their job. It should however be noted that most users do not learn every aspect of a system (Nielson, 2000). Instead they simply learn the aspects that are critical to their work or interests. Ergo, the user is an expert in that sense.

Memorability refers to a user's ability to use a system that they have not been in contact with for some prolonged amount of time. This is important because a project may require that someone do various different tasks. Users may only use certain systems for a while, and then use a different system. This is why it is important to recognize the existence of the casual user. Not all users spend long interrupted lengths of time with a system. It is important to know if individuals retain what was learned and can continue

using the system at the same level of efficiency as when they left off. Errors refer to error detection, prevention, and correction. A lot of time can be spent on error correction, so it is important that this issue is addressed. Some programs like Microsoft Word have features that automatically detect and correct possible errors. This can be a very valuable function and at the same time be problematic as it may correct what it believes to be an error, which in reality is not.

Last, satisfaction refers to how pleasant a system is to use. This is often accomplished through interviews and questions designed to elicit users' subjective feedback. Satisfaction is particularly important because as a system's "pleasantness" falls, users will want to spend less time using it. This information can be used in improving a system to users' standards and helping an organization design new systems in the future.

These usability criteria as outlined by Nielson are important and Gould and Lewis (1985) outlined a global structure for implementing usability into design. They referred to usability as UCD (user centered design) and their three main principles are:

- Early focus of target users and their typical tasks
- User Testing to gather behavior data on the system/product
- Iterative Design (design, modify, and test)

These principles should be integrated with Neilson's five characteristics. Neilson's characteristics are simply the elements of usability that should be used throughout system or product development, while Gould and Lewis have pointed out that Usability should be considered early in development. If usability is considered later in the development

process, then a lot of money will be lost trying to redesign the system to meet customer

and user needs.

Three specific areas of specialization within web usability are (Gould & Lweis,

1985) (see Fig. 2)

- Information Design

- Navigation Design

- Graphic Design

Information design refers to how information in a web site is structured. This is a global

view. Navigation design refers to the design of web navigation, or how a user finds their

way around the site. Graphic Design refers to how information is structured locally,

whether it is using color, pictures, text, or other forms of graphics.

Information Design

Navigation Design

Graphic Design
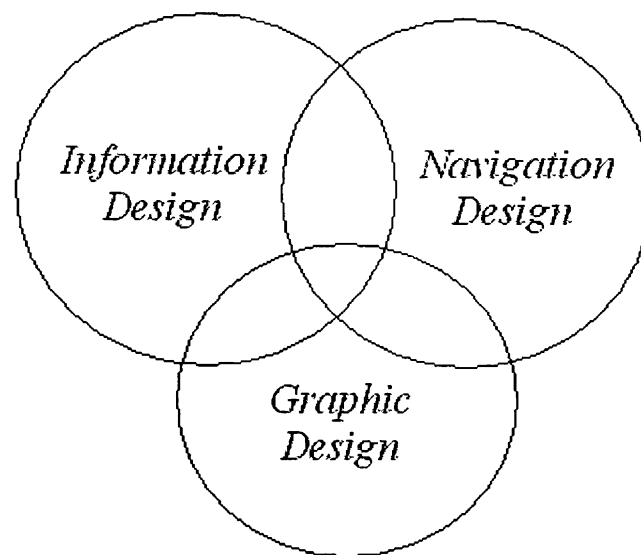
Figure 2 Web Usability Specializations from Newman & Landay (2000)

Newman and Landay (2000) outlined the above model while talking to professional web

designers. They also discovered that while these professionals used UCD, they stated

only doing so when needed. These professionals would often gather the necessary data

for information and navigation design, however graphic design was often left as the final

part and was outsourced. They also noted that usability evaluations touched partially on information and navigation design but almost completely neglected graphic design. However by making use of non-task user testing developers may be able to capture some of this neglected information.

Usability Considerations

> "Ultimately, users visit your website for its content. Everything else is just the backdrop. The design is there to allow people access to the content. The old analogy is somebody who goes to see a theater performance: When they leave the theater, you want them to be discussing how great the play was and not how great the costumes were." (Nielson, 2000)

This quote points out exactly the problem with a lot of web design. A lot of web sites try to dazzle and impress us with their use of the latest web technology and graphical expertise in order to draw our attention. But they are lacking when it comes to simplicity and usefulness. So, three important guidelines developers want to adhere to when designing web sites are (Nielson 2000):

- state information briefly and clearly

- scannability

- use hypertext to break up large quantities of information

Information on a web site should always be stated briefly and clearly. Users want to arrive at the desired information very quickly, and do not want to read through a lot of useless text. Because of this developers must design sites for scannability. By doing this, users can skim through a site and easily locate information. Full, complete, and grammatical sentences may not always be the best choice. The internet is a different medium than a physical publication like books, or magazines. So developers must think

differently when writing information. Bulleted points, phrases, and quotes can help

outline the information on a web site so that users can find and read things quickly. Also,

since the internet does not have the standards a publication might have, information on a

web site should be stated in less than half the space it would appear in a printed

publication.

Nielson (2000) states that a HOMERUN website has the following characteristics.

- (H) igh-quality content

- (O) ften updated

- (M) inimal download time

- (E) ase of use

- (R) elevant to users' needs

- (U) nique to the online medium

- (N) et-centric corporate culture

A lot of issues can be addressed by the technical expertise of the engineers and

programmers (who build web sites) however there is a real need for Human Factors

specialists to be involved in the development process. Such a specialist needs a

background in usability, research, testing protocol, and human behavior among many

other things. With this knowledge base, the Human Factors specialist can address

pertinent issues regarding the usability of the web site of interest. Another important

thing to note is that a Human Factors (HF) specialist must also be flexible, a good

communicator, and able to see the "big picture" or main idea while working with

development teams. New unforeseen problems may arise while designing a website or

system for usability, and a HF specialist should be able to adapt to these changes and make the necessary adjustments to tests, design, and implementation.

Usability Misconceptions and Problems

There are three popular myths or misconceptions surrounding software usability and they include (Mehlenbacher, 1993):

1. Eventually we will design and build such simple and innate systems that they no longer require support documents.

2. Increasing user feedback and usability testing will also increase the systems lifecycle (regarding development), and cost.

3. A perfect list of guidelines will eventually be produced, by research and design.

The systems that are already simplistic and appear almost innate are things that are imbedded in our lives socially and culturally. Things like pens, paper, books and clocks are immensely imbedded in our lives. We use them on a daily basis and don't think much about it. Take a pen for instance; the ink formula or color may change a little, but the basic function of the pen is still the same. Software is a different product entirely. Software systems have various different uses and functions, as well as different hardware systems enclosing them. Software systems are also constantly changing and evolving. Ergo, we can see that developing systems that are so simple and innate that they require no support documentation is an illusive dream.

The second misconception isn't completely wrong. In fact, gathering user feedback and conducting more usability tests does increase the time and cost of a systems development. However in a more general sense these costs prove to be more profitable,

and can sometimes shorten lifecycle development, as it takes longer to correct a mistake at the final stages of development than collecting the appropriate amount of data at the beginning and applying it early on. While sometimes it appears that putting the time, and resources into usability is just an unwanted expense, the alternative can prove more beneficial/profitable, and isn't the point of capitalist business?

The last myth suggests that eventually we will have perfect guidelines to help us develop a system. While, guidelines can be very useful, they can also be too rigid or too general. "Tailoring" is a very important part of software as well systems engineering (Blanchard and Fabrycky, 1998). With new demands from customers and new technology come new products and new ways of approaching problems. It has been found that the most important project for a systems engineer is his/her second project. Their first project is done by following and seeking help from a senior systems engineer. But in their second project they must not fall into the trap of mapping all the solutions of the first onto this new project. This is where an individual must learn to tailor solutions to issues and problems. So while guidelines can prove useful, they can also cause problems. It would be best to continuously keep people aware of this issue. A guideline may provide a way to get started, but if you truly tailor the project to the needs of the customer, almost any guideline can be changed.

The following are some common fundamental errors in usability of (Nielson, 2000; Rubin, 1994) user interfaces such as web sites:

- Emphasis on the system and not on the user
- Organizations not aware of the changes in their target users due to current trends in technology

- Designing for usability is not as easy as common sense

- Treating a web site like a brochure instead of a medium to conduct business

- Designing a website to reflect how the organization is set up instead of being designed to meet user needs

- Using the newest advances in technology to entice the user instead of meeting needs in a more natural and realistic way

- Writing text on a web site like you would a physical publication

- Treating your web site as if it is an isolated body of information instead of giving well placed links to other sites, companies, and information

As we can see Web Usability requires a unique focus. Many people are accustomed to writing in a physical medium (papers, books, magazines, reports, etc.) however the internet is something different. Developers need a new outlook when designing web sites. Users want information to be easy to understand short and able to convey a lot of meaning. This means developers may want to make use of visual information and broken text more than tradition forms of information presentation.

The Internet is far from a perfectly designed user interface. Users often spend a lot of time trying to locate information (Nielson 2000; Ratner, 2003) and may end up finding themselves lost. The internet also changes at a very rapid pace without the restrictions that paper publications have. Thus, it is important to understand as much as possible about the user so that as technology and demands change we can present users with information in the most efficient manner.

Usability Tests

The importance of usability testing is that it is done early enough to use the information in product development. In the past, usability tests were done towards the end of a product's development. This meant that what was learned was used along the lines of making simple cosmetic changes to the product, or to show that the product met the customer needs. But this isn't very useful. Many people today agree that usability testing is needed all throughout the development lifecycle. Bailey's (1982) Human Performance Model (See Fig. 3) highlights the main idea behind usability testing. The user (Human) interacts with their environment (Context) by performing some task(s) (Activity). Each of these three things should be given a
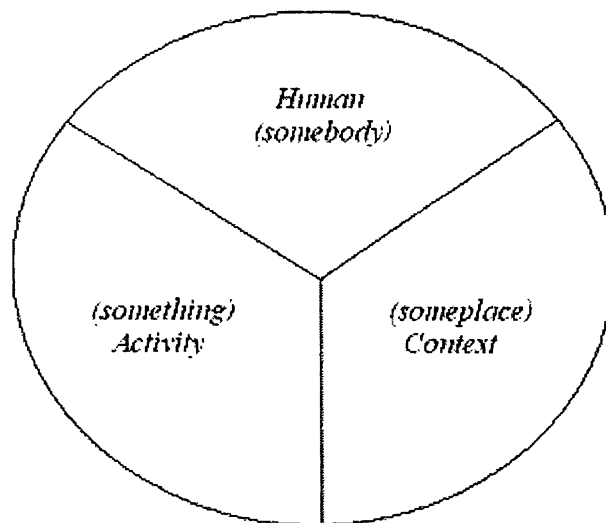
Figure 3 Bailey's (1982) Human Performance Model

considerable amount of attention as well as their relationships in usability. By observing this model it becomes clear that problems can arise if developers only focus on the activity and context to acquire any necessary information about the user (Human). By doing this a simple or even large issues with the user can be overlooked causing the system or product to fail or not perform as well as it could have.

If a developer happens to be a user of the system being developed he/she might

think, "why do I need to test users when I am one." This is a good point; however

developers spend a large amount of time studying a system. So user testing is performed

in order to uncover any details about a system that have been overlooked. McCracken

and Wolfe (2004) describe the iterative nature of usability testing (See Fig 4) in three

main steps:

- Design UI (User Interface) to satisfy user requirements or needs

- Build a prototype of the UI

- Proceed with user testing
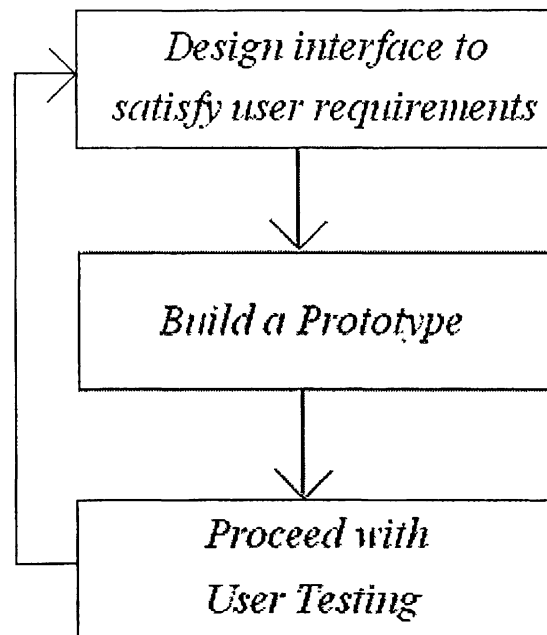
Figure 4 User Interface Refinement Cycle from McCracken and Wolfe (2004)

These steps should be repeated as many times as necessary in order to meet those

originally specified user needs.

A few examples of iterative testing (See Table 1) are as follows. Users review the

documentation while using the application, comfortably. Users perform a pre-planned

task(s) using the product and its documentation. Users look for specific information in the documentation. Users paraphrase and/or summarize portions of the documentation. Users maintain a record of critical events. Again, these tests can prove very helpful if preformed all throughout the system's lifecycle. So, the question now is how to implement iterative tests. Talk-aloud protocols are one example. This is often done by setting up a scenario or a sequence of tasks for users, and asking them to speak aloud and let the experimenter know what they're doing. This is useful because it lets us know what is on the users' minds, what they are doing and why they are doing it. Videotaped sessions are a less intrusive method of capturing this kind of information. Another important difference is that videotaped sessions are more visual while talk-aloud protocols are more auditory. Videotaped sessions can however make use of a talk aloud protocol, but then it becomes a mix of the two styles. Since videotaped sessions are more visual, we can now look at what our users are doing. If it takes them a long time to perform a certain function, or if they repeatedly perform a function incorrectly, all these things can be seen. Even though the users are being watched they also work more comfortably in this way because they aren't being watched by someone who is in the room taking notes of their actions. In a sense, the camera isn't as intrusive allowing the user to become absorbed by the task at hand. User-log analyses are the least intrusive method of collecting feedback. In this situation, the user is again given some sequence of tasks to perform. The software or system itself monitors the user's progress. So, time to complete a problem, reading/writing speeds, and how often a menu or button is accessed are examples of data that is collected in this type of test.

Interviews are another approach to collect some data. An interview can be conducted via, in person, e-mail, or pen and paper. Rather than conducting this while a person is performing a task, an interview is done preferably immediately afterwards. This type of test assesses more of what a user thinks are important issues and solutions to problems. This lacks the objectivity of user-log analysis, but it gives developers insight to what the user is thinking. We must ultimately please the user, hence it's important to get their input on the product. User surveys are also conducted when the user is not currently using the product. These surveys are often a list of questions with a multiple choice answer gauging how positive (or negative) users felt about various features of a system. This is a simple way to collect data; however there is always a problem with the response rate to surveys. This response rate always begs the issue of some people being biased in their response; hence you only get information of a subpopulation of users.

System benchmarking involves comparing systems to some standard or agreed upon benchmark. If a system processes a certain task slower than the current benchmark than there may be a problem, however if it processes the information faster than the benchmark, than something was done correctly, a that can possibly be used as selling point.

Last, Beta testing involves observing a system as it is used in the user's environment (home, workplace, etc.). This can be a very useful test if different solutions have been found to solve the same problems within a system. By putting the system to the test in its intended environment you can obtain data that will closely resemble how users will view the system. Again, these are just a few tests that can be performed to assess the usability of a software system. Each test has its own problems and advantages.

But we can not choose to perform one kind of test and ignore the others, because while some tests may have some unavoidable problems they also provide us with different types of information that otherwise would not be gathered.

User Test	Description	Examples
Talk a-loud protocol	User speaks aloud letting the experimenter know what they are thinking (also known as think aloud)	Nielson (1993)
Video taped session(s)	Experimenter uses a video recording device so that user behavior can be later analyzed	Sazegari, Sara (1994)
User log analysis	A device monitors user performance during a test (e.g. time, errors, and # of tasks)	Nielson (1993)
Interview	Experimenter asks questions of the user in order to access their reactions	Greenbaum, Thomas L. (1997)
User survey	Questionnaires written by the experimenter and given to groups of users	Alreck, Pamela L. & Settle, Robert B (1994)
Beta testing	Testing user performance with a fully functional prototype	Rubin, Jeffrey
Coaching method	Problems are identified and solved by allowing the user to ask questions An expert or "coach" is present to answer these questions	Nielson (1993)
Co-discovery learning	Two users work together in performing a set of tasks. (designed to simulate workplace environment)	Rubin, Jeffrey
Question asking protocol	An experimenter asks questions of the user in order to access the user's understanding	Dumas, JS & Redish, Janice (1993)
Shadowing method	The expert makes use of an "expert" to explain the user's performance (e.g. behavioral scientist or psychologist)	Nielson (1993)
Teaching method	One user is allowed to become familiar with a system in order to then teach a new or naïve user how to use the system	Vora & Helander (1995)
Heuristic Evaluation	Experts (e.g. Human Factors Engineers) examine the nature of a system for problems	Nielson (1994)

Table 1 Usability Tests

Task vs. Non-Task Testing

Using the "task method" in usability testing refers to user testing. This involves putting the user in a typical environment performing typical tasks (usually with the system or product being considered). This is a very simple method of usability testing and is done often (Hong, Li, & Lin, 2000). "Non-task testing" refers to heuristic evaluation. Heuristic evaluation (Nielson, 1994) is the process of having three to five Usability experts inspect some system or product against usability standards and look for any possible problems or flaws. These experts make these evaluations alone. After this is done they may get together and discuss what they have found, how to address it, or possibly highlight new problems/issues.

It makes sense to use user testing for web design or for a system in terms of usability because developers are concerned with how well users use their product. However the difference between evaluating a system or product with or without tasks may be very important. Psychology tells us that we deal with situations by building schemas or mental models (Sternberg, 2003). These schemas help us to process information and to organize it, however it is not full proof. If a person needs to go shopping, he/she might choose to remember what he/she needs in terms of what he/she plans to cook that night. So if he/she plans to make lasagna, he/she might look for cheese, meat, sauce, and noodles. However this causes problems when things exist that do not fit neatly within that schema. What happens if a friend asked you to pick up some herbal tea while you are out? You may end up not getting the tea because you were focused on the lasagna and what you needed to make it. The same thing can happen while testing users with a new system or product. If users are focused on a task, they may not be aware of

other problems or flaws; hence they can not report them to the tester. This does not mean that user testing is not valuable. It only means that user testing (using tasks) do not help developers see the entire picture. Similar problems may exist with heuristic evaluation, such that these evaluators are professionals and are looking for specific things, and comparing what they see against known standards. A non-task test for users may provide some information that is not accessible using the other testing methods. A non-task test would be designed such that users would look for problems with a web site without having to perform a task such as navigation, downloading, or making an online purchase.

User testing is focused on emulating the user environment and the tasks they would perform. This is a valid way to approach testing; however it may limit the user's response. By allowing users to explore or browse a system or web page, users can focus on looking for any and all problems without having to worry about completing a task. This freedom may result in users missing some errors, while discovering others.

According to Nielson (2000), user testing is designed to mirror typical tasks that a user would perform with a system. Developers must first identify these tasks before constructing a scenario for users to enact. During these user tests, experimenters monitor user behavior from an objective and subjective viewpoint. Data is often collected on performance time(s), problems found in a design, user level of performance, and user impressions. This data, once collected, can help developers identify any potential problems in a system.

Fidelity in Web Usability Testing

Fidelity is term used to describe how closely something resembles, that which it is emulating. In the example of high fidelity, a mock up would very closely resemble the

original such that they may even be indistinguishable, and the opposite is true for low fidelity. User testing has been performed for web site usability with varying degrees of fidelity (See Table 2) (Walker, Takayama, and Landay, 2001). Low fidelity is used earlier in the development process because of its simplicity to gather user needs and help develop a framework for a web site. Later some medium fidelity "mock-ups" are used to further identify user needs as well as identify any potential problems. Eventually a high fidelity prototype is developed which is a fully functional web site. Users can now navigate the site and use its various functions. User testing has shown that fidelity has little or no impact in task based user testing (Walker, Takayama, and Landay, 2001). This suggests that it would be best to use low fidelity testing as much as possible, because it is cheaper and faster to create and manage.

Level of Fidelity	Description
Low	Pen & paper sketches, or very simple sketches using visual programs
Medium	Color printouts of the website missing specific details such as dates, prices, names, etc.
High	A fully functional web site prototype

Table 2 Web Site Fidelity

Low fidelity testing (LFT) is important at first because low fidelity is rough and ambiguous (however that ambiguity can also be problematic). LFT gives users a chance to look at the main points of a web site without the distractions of graphics and templates. But this type of testing has some flaws as some professionals mention (Newman and Landay, 2000). Rough sketches (low fidelity) are useful and cheap to produce, but are hard to manipulate. If users like certain features or want to see certain features change slightly, the entire sketch has to be redrawn. This along with the fact that testers will need to eventually create something that looks more professional pushes them to use graphical

tools earlier and earlier. Programs like DENIM (Lin, Newman, Hong, and Landay, 2000) and SILK (Landay and Myers, 2000) allow developers now to create rough sketches with a GUI (Graphical User Interface) rather than just using pen and paper. This makes it even easier to modify designs for users, and to move between varying levels of fidelity. Because the information is stored digitally web developers can use these tools to overlay or remove levels of fidelity and test it with potential users. These tools will make a significant impact on web development similar to the movement to using DOORS instead of Microsoft Word for developing system requirements documentation.

The advantages and disadvantages of low and high fidelity testing are apparent to developers, and it often drives how they design their user tests (Gould and Lewis, 1985). Low fidelity offers certain advantages (see Table 3) that high fidelity design (see Table 4) can't offer. The reverse is also true. Management and users expect to see certain things that work well under certain conditions. So developers must cater to these needs when designing what types of tests to use at which level of fidelity.

Low Fidelity Pro's & Con's	
Pros	**Cons**
Easy to build	Ambiguity
Easy to modify	Fail to indicate to management what the actual product will look like
Its less professional look does not affect user opinions	Fail to portray time spent doing actual tasks
Does not require a professional in order to create	
Maximizes refinement before a prototype is developed	

Table 3 Low Fidelity Advantages and Disadvantages

High Fidelity Pro's & Con's	
Pros	**Cons**
Has a professional look	Creation requires time and a skilled professional
Portrays to management how the system will look and work	Modification is not a simple task
	A small error or "bug" can have a huge impact on tests

Table 4 High Fidelity Advantages and Disadvantages

According to McCracken and Wolfe (2004), building a paper prototype (low

fidelity) consists of the following.

- Gather some simple materials: blank paper, markers, glue, tape, and sticky

    notes

- Outline the system template (in the case of web design copy the UI of a browser

    like Internet Explorer or Safari)

- Construct a model of the system or web site.

In creating a low fidelity design it is important, to outline to the user what information

represents hyperlinks or functions. This can be done by underlining them or highlighting

them with a marker. Other than a vague and simple representation of a web site, a low

fidelity design can also be a script that outlines to the user what information will be

present in the design. A high fidelity design is a working prototype. This is accomplished

by hiring the necessary professionals (e.g. software engineers and programmers) to create

the design outlined in the lower levels of fidelity. A high fidelity design is also a

completely functional product such that all information and functionality is present

within it. Medium fidelity is generally referred to as a mock up or low level prototype.

This design is very similar to the High fidelity prototype, however is can be missing

detailed information and some functionality. This allows developers to give management

something that has a professional look as well as test users on something that looks closer to the actual prototype.

Research Objectives and Rationale

The internet is a powerful tool and has changed the way companies do business, and the manner in which information is shared. So it is important to understand the best means to with which to test individual performance on web sites so that we can outline to web developers the most effective way to construct a web site. A web site is a unique medium and is not to be confused with paper publications such as magazines and books. Users look towards information on the internet to be brief, important and meaningful (Neilson, 2000; McCracken & Wolfe, 2004). It is all too easy to go somewhere else if you do not find what you are looking for at a specific web site. So a web site must provide the user with information in such a way that it is important and easy to find. A web site is successful when a user frequently re-visits the site. It is however unfortunate that many web designers are not aware of the importance of usability in designing a web site. This is evident in the vast amounts of web sites on the internet that contain annoying "pop-up" advertisements, or excessive use of the most recent technology such as video or macromedia flash. While it is true these things can all provide useful information they are often ignored or avoided by users (Nielson, 2000; Ratner, 2003).

Current research done on the effects of fidelity on task based usability testing (Walker, Takayama, &Landay, 2002) suggest that fidelity has little to no impact on user performance to identify problems and help developers create a better web site. Knowing this fact it would then be reasonable for a company to use low fidelity testing as long as possible during testing because it is cheaper and faster to implement. A low fidelity web

design is all on paper using pen, pencils markers, or other writing utensils. So anyone can make modifications and it is not necessary to hire a web designer for this.

There has been no research as of yet looking into the effectiveness of non-task based testing. Most information that is gathered in a non-task fashion is heuristic evaluation. Heuristic evaluation is a method of examining a web site by using experts such as Usability Engineers or Human Factors specialists (Nielson 1994). The expert's task is simply to examine the website and inform the developers how well it adheres to current usability standards and if there are any potentials problems that need attention. This is a useful source of information however it may also be useful for users to be offered this same freedom in examining a website for problems. It is typical that user testing for websites is done with tasks that closely resemble what task the user would typically perform (Nielson, 1994; Nielson, 2000; McCracken, & Wolfe, 2004). The idea is that the test should closely resemble the user's environment and tasks, and performance is measured. However web usability testing is also interested in identifying problems with the design and proposed solution. This is the main or essential task, so it may be beneficial for some user's to identify these problems without having to worry about performing some list of pre-specified tasks. It is possible that this freedom will allow users to identify problems that they were previously not as aware of, because they were more focused on performing the pre-specified tasks.

This study intends to investigate:

- the difference in performance (ability of users to identify problems they experience) due to a task and non-task based test
- the interaction between fidelity and testing method for finding problems

The specific hypotheses for this study are:

- There is no significant difference for fidelity between groups for finding

  problems (user difficulties in using website)

- There is a significant difference between testing methods for finding problems

  (between task and non-task conditions)

- There is an interaction effect between testing method and fidelity for finding

  problems

Ability to identify problems for each condition will be summed up by distinctly different problems. So, no repeated problems will be added to the group's overall score. Problems will then be categorized as referring to navigation design, information design, and graphic design. The intent here is that the number of problems identified may not differ altogether, however the differences may lie within the navigation, information, and graphic design categories. The expectation is that users in the non-task condition will be likely to identify more problems, than the task condition users. This comparison between task and non-task conditions is the testing condition that will be referred to. It will also be called the task/non-task condition or variable. An interaction is expected such that the non-task group will identify more problems in higher fidelity conditions because more information is present, and vice versa.

Methods

Participants

Thirty undergraduate students at Embry-Riddle Aeronautical University volunteered to participate in this study. Each participant signed a consent form indicating

their willingness to participate in this study and that they can leave at any moment if they feel uncomfortable. Also, gratitude was expressed towards each volunteer for their participation.

Apparatus

A *Dell dimension XPS T500,* equipped with a Pentium IV processor, 19 inch Dell monitor with a resolution of 1024 X 768 pixels, with Windows NT. The web site was created with pen and paper as well as with Microsoft Front Page. Questionnaires used were printed out on paper to be filled out during the test (See Appendix). Also a VHS camcorder was used to observe behavior.

Design

This is a 3x2 between subjects design.

	Task-based	Non Task-based
Low-Fidelity	5	5
Medium Fidelity	5	5
High Fidelity	5	5

Table 5 Subjects Used in each Condition

Five participants were used in each condition in accordance with Neilson's suggestion (Nielson, 2000) for sample size in usability testing (see Figure 5). 30 participants were used in all (19 male and 11 female). Each participant was currently a student of Embry-Riddle Aeronautical University. The web site used in this study is the Embry-Riddle Aeronautical University Department of Human Factors and Systems which is located at http://www.humanfactorsandsystems.com/. This site was fully functional and was used as the high fidelity model for testing. The medium
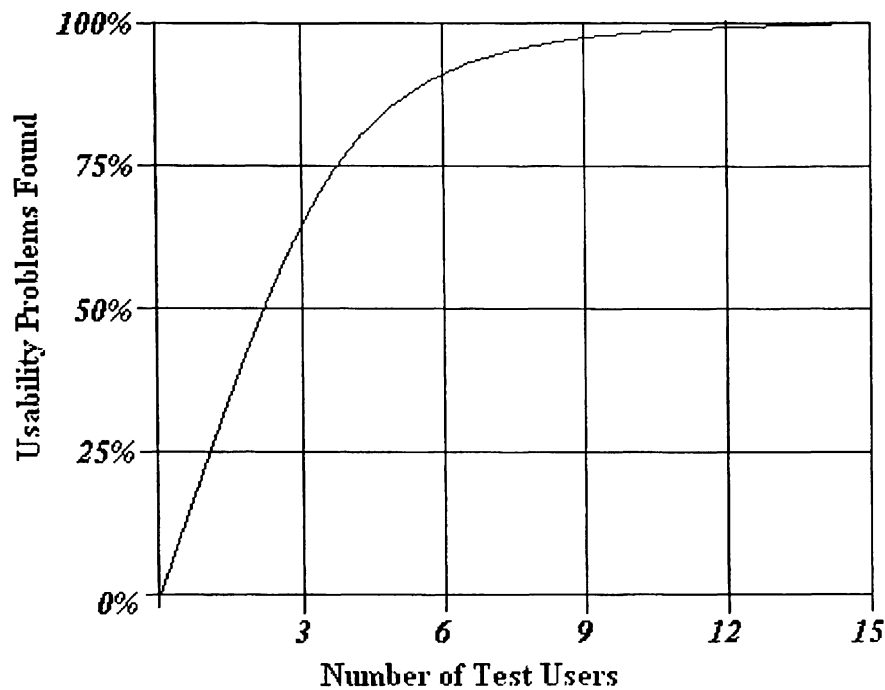
Figure 5 Usability Problems Found Per User (Nielson, 2000)

level of fidelity was created using this site and Microsoft Notepad. Medium fidelity was

similar to the final prototype. It was in black and white had the basic look and structure of

the web page, however it lacked specific information such as the exact pictures and

graphics of the high fidelity design. Low fidelity was a simple paper and pen sketch of a

web site. It contained the same information, and basic layout as the medium level of

fidelity, however it contained no graphics. The low fidelity design was on paper, and the

medium and high fidelity design was displayed on a computer monitor. Both medium and

high fidelity models accessed the computer's hard drive rather than accessing a web

address.

During this test, users were given a simple user profile questionnaire, a scenario

(task or non-task), and a post test questionnaire (see Appendix), as well as a consent and

debriefing form. The task based scenario consisted of a list of questions requiring the user

to identify and record information found on the web site. While performing this task the

user is asked to identify any problem(s) found in the web site, as well as the severity of that problem on a one-to-five scale (with 5 being the most severe, and 1 being the least severe). Problems are defined as any difficulties a user encounters or any elements of the web site that could be improved some way. These problems generally fall into three groups: Graphic Design, Information Design and Navigation Design. Users in the non-task condition were asked to browse the web site and simply identify any problems they find. Participants in both task and non task condition were given exactly 20 minutes to complete the questions and/or identify problems. They were all asked to continue doing so for the entire 20 minutes. Lastly the post test questionnaire was a small list of questions accessing the user's impressions of the website in general. During the course of these tests users' behavior was recorded with a Digital 8 Sony video recorder to ensure similarities in behavior among subjects and to note any unusual behavior.

The data was analyzed using an ANOVA with a Tukey HSD post-hoc analysis. SPSS was used to calculate this data.

Results

User Profile Results

The following (Figures 6 – 10 and Table 6) show the results of the participant profile questionnaire. Looking at Figure 7, most participants were upperclassmen or graduate students, and so an average age of 23.9 (See Table 6) is expected of such a group. Also, most participants were Human Factors majors or Aeronautical Science majors (See Figure 6). This is possibly the case because data collection occurred during the summer semesters. The proportion of male to female (19:11) participants happened to

be the proportion of males to females on Embry-Riddle Aeronautical University's campus (See Figure 8). Also, it was found that most participants have been using computers for longer than five years. This is also an expected result as more and more people use computers everyday, for work or recreation. This result includes use or exposure to desktops, laptops, and computers of any Operating System (e.g. Windows, MAC, and Linux). Also, most participants browsed the internet between six and ten hours a week, which could mean one hour a day on average. This profile appears to describe the typical college student.

Age Statistics	
Mean	23.9
Median	22.5
Mode	22

Table 6 Age Statistics
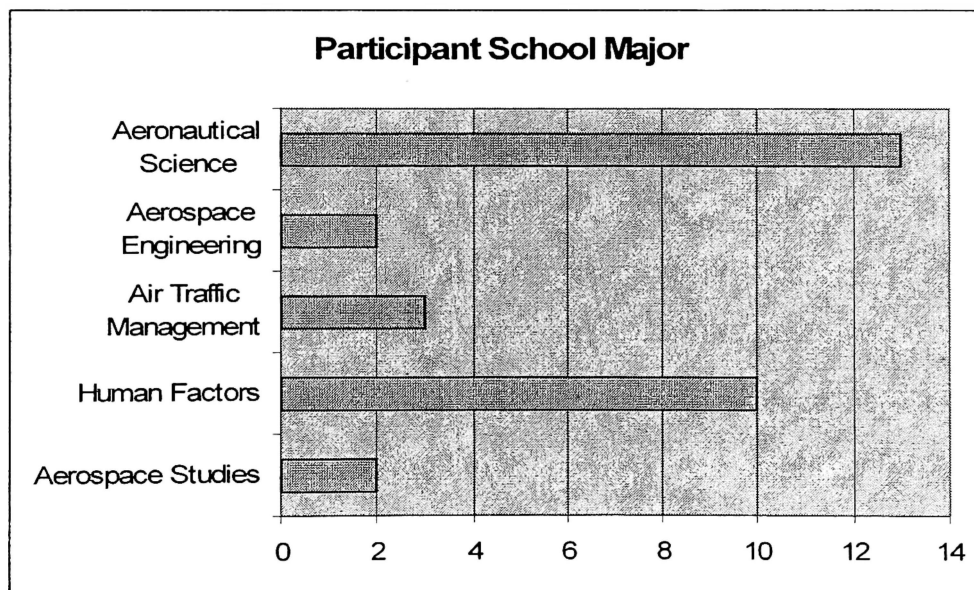
Figure 6 Participants' Major

Participant Class/Year

Graduate

Senior

Junior

Sophmore

Freshman

0   2   4   6   8   10
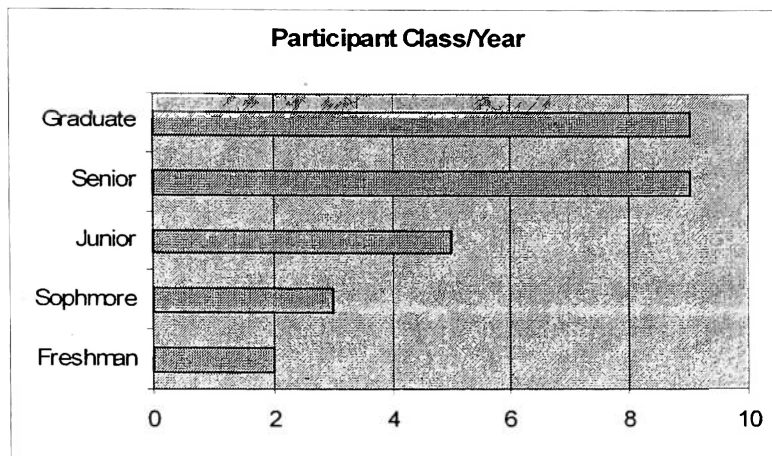
Figure 7 Class/Year of Participants

Participant Gender

Male

Female

0   5   10   15   20

# of Occurences
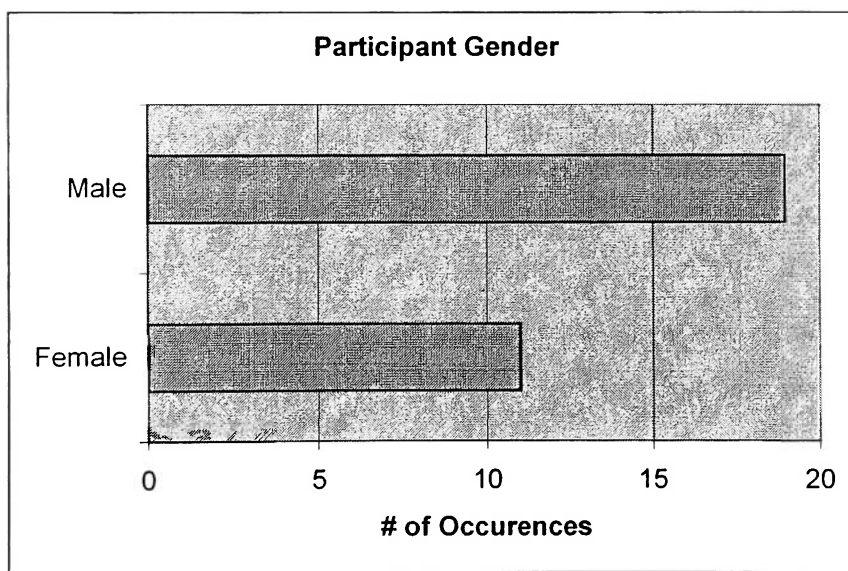
Figure 8 Gender of Participants

Years Spent Using Computers
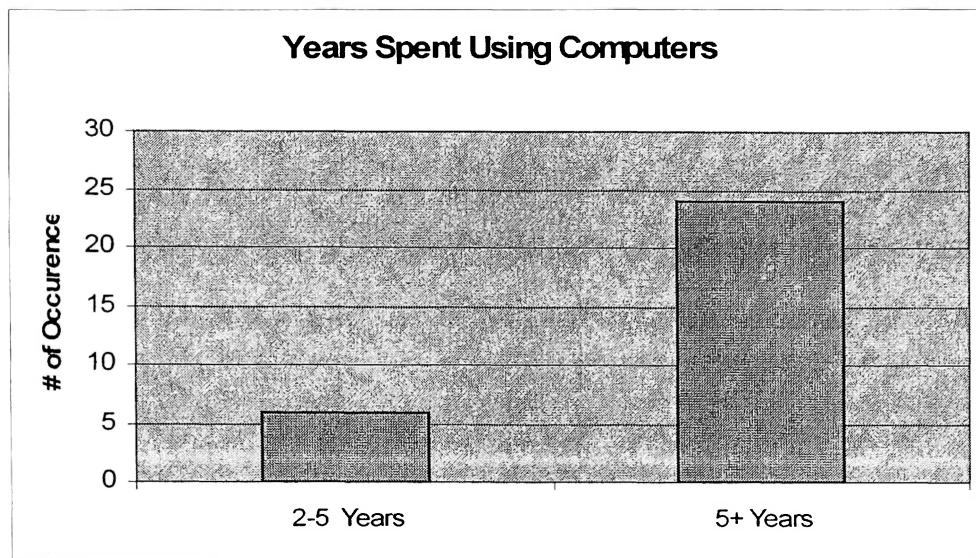
Figure 9 Amount of Years Participants Have Used Computers

Hours Spent Browsing the Internet
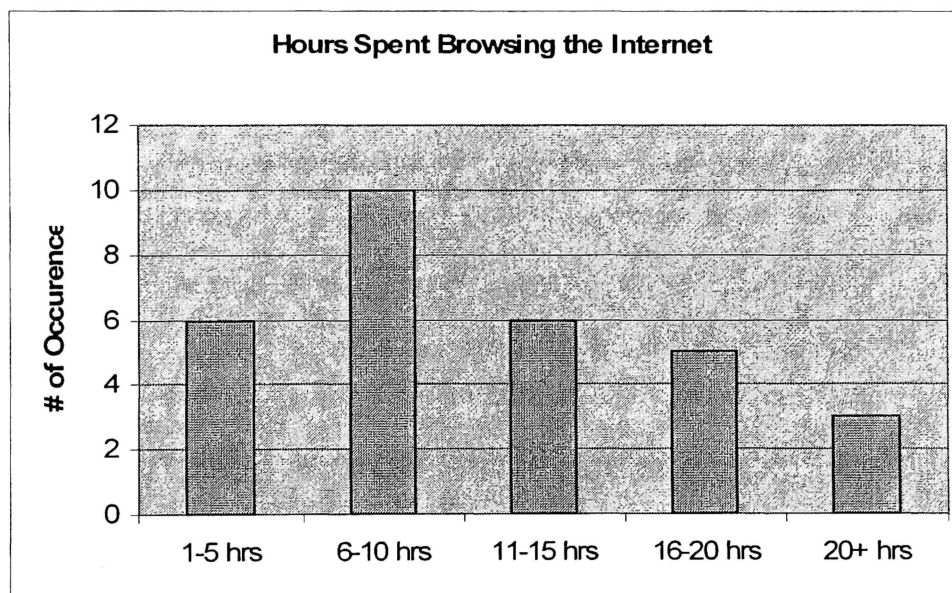
Figure 10 Hours Per Week Participants Spent Browsing the Internet

Results on Number of Identified Problems

A factorial ANOVA was then performed on the collected data regarding problems

identified and the task/non-task condition was found to be statistically significant ($p =$

.02) (See Table 8) such that the non-task condition identified more problems than the task

condition. Fidelity ($p = .86$) (See Table 8) and the interaction effect ($p = .49$) were not

found to be significant. This implies that only the task condition has an impact on

performance confirming the hypotheses that there would be a significant difference for

the task/non-task condition on performance and there would be no significant difference

for fidelity condition. However the final hypothesis was rejected. There was in fact no

significant difference for the interaction between these two variables. The effect size

found for the interaction was 6% which means that the interaction did not have a strong

effect on the number of problems identified. The fidelity condition only accounted for 1%

of the variation in the scores, which is a very low percentage. The Interaction effect had

an effect size of 6%. The means for problems identified under each condition can be seen

in table 7.

Average # of Problems Identified	Low Fidelity	Medium Fidelity	High Fidelity	Total
Task Condition	2	4.4	5.6	4
Non-Task Condition	9	8.4	7.6	8.33
Total	5.5	6.4	6.6	6.17

Table 7 Average Number of Problems Identified

Table 8

Analysis of Variance				
Source	df	F	η	p
(A) FIDELITY	2	0.16	0.01	0.86
(B) TASK/Non-Task CONDITION	1	6.39*	0.21	0.02
AxB Interaction	2	0.73	0.06	0.49
S Error	24			

* $p < .05$

Power was observed for these variables given an alpha of .05 (See Table 9). The power

for the condition of interest, task condition, was 0.686 which is a mid to high level of

power. This means that 68.6% of the time, if this experiment was run repeatedly the same

results would occur. If a larger sample size had been used a greater value of power would have been observed. Increasing the alpha level would also be another method to increase power. The power for the fidelity condition and interaction was very low.

Condition	Power
(A) FIDELITY	0.072
(B) TASK/Non-Task CONDITION	0.686
AxB Interaction	0.159

Table 9 Observed Power for # of Problems Identified

Post-hoc pairwise comparisons using Tukey's HSD (Honest Significant Difference) yielded no difference between any of the groups (See Table8). The significance value observed for all of these comparisons was greater than 0.05. This method of pairwise comparison was used because all possible comparisons were made after observing the recorded data. This result is possibly due to the fact that a small sample size was used. Had a larger sample been used a significant difference might have been observed.

Tukey HSD _p_ vaues			
	Fidelity Comparisons		
	High – Medium	High – Low	Medium – Low
Problems Identified	.995	.858	.903

Table 10 Tukey HSD Significance Values for Problem Severity

Results on Problem Severity

Problem severity and participants' subjective impressions of the web site from the Post Scenario Questionnaire (See Appendix) were then analyzed using ANOVA. Severity refers to the average severity participants ranked their problems. The subjective impressions refer to the four categories that questions were broken up into in the Post Scenario Questionnaire. Overall impressions refer to their overall opinion of the website.

Information Presentation refers to how the information was laid out on the web site. This refers more to bodies of text, descriptions, and outlines. Navigation refers to refers to participants' ability to navigate throughout the website; this generally refers to the function and visual salience of links, and graphics refers to imagery such as logos photographs and other images used to enhance the website.

Looking at the average problem severity given by participants (See Table 12), a significant difference was only at the fidelity condition, $p = .01$. This was found with an effect size (See Table 12) of 0.35 and a power (See Table 13) of 0.864. This is a considerably large power and a relatively large effect size. So, this again means that the condition of fidelity accounted for about 35% of the variation in scores. Looking closer at this data, using Tukey's HSD, a significant difference was found (See Table 14) between the high and low fidelity ($p = 0.01$), and the medium and low fidelity ($p = 0.03$). No significant difference was found between the high and medium fidelity conditions. This is likely due to the fact that both occurred on the same medium (a computer monitor). From the results it can be seen that problems were more severe as fidelity increased. The means for average problem severity under each condition can be seen in table 11.

Average # Problem Severity	Low Fidelity	Medium Fidelity	High Fidelity	Total
Task Condition	1.2	2.7	2.9	2.5
Non-Task Condition	2.4	3.3	3.5	2.9
Total	1.8	3.0	3.2	2.7

Table 11 Average Problem Severity

Table 12

Analysis of Variance (for Avg. Problem Severity)				
Source	df	F	η	p
(A) FIDELITY	2	6.45*	0.35	0.01
(B) TASK/Non-Task CONDITION	1	1.17	0.05	0.29
AxB Interaction	2	2.06	0.15	0.15
S Error	24			

* $p < .05$

Condition	Power
(A) FIDELITY	0.864
(B) TASK/Non-Task CONDITION	0.180
AxB Interaction	0.382

Table 13 Observed Power for Avg. Problem Severity

Tukey HSD p vaues			
	Fidelity Comparisons		
	High – Medium	High – Low	Medium – Low
Severity	.825	.007	.026

Table 14 Tukey HSD Significance Values for Problem Severity

Results of User Impressions

Looking at participants' overall impressions of the website (See Table 16), a significant difference was found only at the fidelity condition, $p = 0.01$. An effect size of 0.39 and a power (See Table 17) of 0.915 was observed for participants' overall impressions. A power of 0.915 is a considerably large power. The observed effect size of 0.39 is a relatively large effect size. The condition of fidelity accounted for about 39% of the variation in scores. Looking closer at this data, using Tukey's HSD, a significant difference was again found (See Table 18) between the high and low fidelity ($p = 0.03$), and the medium and low fidelity ($p = 0.01$). No significant difference was found between the high and medium fidelity conditions. So, it can be seen that participants' overall

impressions of the website were better in lower fidelity conditions than in higher fidelity conditions. This is likely due to the idea that users have higher expectations of a higher fidelity or finished product than they do of a low fidelity product or draft. The lower fidelity generally looks less professional, and users may get the impression that this is someone's starting idea for a website. The higher fidelity may conversely give the impression that this website is almost done and ready for use. This is not always the case but a higher fidelity has a more professional look than a lower fidelity and hence can cause this impression. The means for users' average overall impressions under each condition can be seen in table 15.

Average Overall Impressions	Low Fidelity	Medium Fidelity	High Fidelity	Total
Task Condition	4.6	2.9	3.2	3.6
Non-Task Condition	3.7	2.5	3.0	3.1
Total	4.1	2.7	3.1	3.3

Table 15 Average Overall Impressions

Table 16

Analysis of Variance (for Overall Impressions)				
Source	df	F	η	p
(A) FIDELITY	2	7.57*	0.39	0.01
(B) TASK/Non-Task CONDITION	1	2.87	0.11	0.10
AxB Interaction	2	0.53	0.04	0.60
S Error	24			

* $p < .05$

Condition	Power
(A) FIDELITY	0.915
(B) TASK/Non-Task CONDITION	0.370
AxB Interaction	0.127

Table 17 Observed Power for Overall Impressions

Tukey HSD p vaues			
	Fidelity Comparisons		
	High – Medium	High – Low	Medium – Low
Overall Impression	.591	.028	.003

Table 18 Tukey HSD Significance Values for Overall User Impressions

Participants' impressions of information content on the website (See Table 20), were similar to that of their overall impressions. A significant difference was found for the condition of fidelity with an effect size of 0.25 and a power (See Table 21) of 0.646. However Tukey's HSD revealed no significant differences (See Table 22) between the varying levels of fidelity, and observing the power that was calculated it appears that this was not due to chance. The means for users' average impression of information presentation under each condition can be seen in table 19.

Average Information Presentation Impressions	Low Fidelity	Medium Fidelity	High Fidelity	Total
Task Condition	4.2	3.2	2.9	3.4
Non-Task Condition	3.9	2.8	3.0	3.2
Total	4.0	3.0	3.0	3.3

Table 19 Average Information Presentation Impressions

Table 20

Analysis of Variance (for Information Content)				
Source	df	F	η	p
(A) FIDELITY	2	3 89*	0.25	0.03
(B) TASK CONDITION	1	0.30	0.01	0.59
AxB Interaction	2	0.18	0.01	0.84
S Error	24			

* $p < .05$

Condition	Power
(A) FIDELITY	0.646
(B) TASK/Non-Task CONDITION	0.083
AxB Interaction	0.074

Table 21 Observed Power for Information Content

Tukey HSD p vaues			
	Fidelity Comparisons		
	High – Medium	High – Low	Medium – Low
Information Presentation	.993	.053	.067

Table 22 Tukey HSD Significance Values for User Impressions of Information Presentation

Participants' impressions of the navigation component of the website (See Table 24) were similar to that of their overall impressions. A significant difference was found for the condition of fidelity with an effect size of 0.57 and a power (See Table 25) of 0.999. Looking closer at this data, using post-hoc pairwise comparisons such as Tukey's HSD, a significant difference was again found (See Table 26) between the high and low fidelity ($p = 0.01$), and the medium and low fidelity ($p = 0.00$). The means for users' average impression of navigation under each condition can be seen in table 23.

Average Navigation Impressions	Low Fidelity	Medium Fidelity	High Fidelity	Total
Task Condition	4.5	1.9	3.3	3.2
Non-Task Condition	3.8	2.0	2.2	2.7
Total	4.2	2.0	2.8	3.0

Table 23 Average Navigation Impressions

Table 24

Analysis of Variance (for Navigation Impression)				
Source	df	F	η	p
(A) FIDELITY	2	16.00*	0.57	0.00
(B) TASK/Non-Task CONDITION	1	3.11	0.12	0.09
AxB Interaction	2	1.20	0.09	0.32
S Error	24			

* $p < .05$

Condition	Power
(A) FIDELITY	0.999
(B) TASK/Non-Task CONDITION	0.359
AxB Interaction	0.283

Table 25 Observed Power for Navigation Impression

Tukey HSD p vaues			
	Fidelity Comparisons		
	High – Medium	High – Low	Medium – Low
Navigation	.126	.004	.000

Table 26 Tukey HSD Significance Values for User Impressions of Navigation

Participants impressions of the graphics content yielded no significant difference

(See Table 28) in this case. A relatively low power was also observed (See Table 29). So

it is likely that these results are due to a low sample size. A larger sample would be

needed to access if fidelity and task condition affect users impressions of the graphic

component of a web site. The means for users' average impression of graphics under

each condition can be seen in table 27.

Average Graphics Impressions	Low Fidelity	Medium Fidelity	High Fidelity	Total
Task Condition	3.4	2.4	3.0	2.9
Non-Task Condition	2.3	1.9	3.0	2.4
Total	2.9	2.2	3.0	2.7

Table 27 Average Graphics Impressions

Table 28

Analysis of Variance (for Graphics Impression)				
Source	df	F	η	p
(A) FIDELITY	2	2.84	0.19	0.08
(B) TASK/Non-Task CONDITION	1	2.94	0.11	0.10
AxB Interaction	2	1.05	0.08	0.37
S Error	24			

* $p < .05$

Condition	Power
(A) FIDELITY	0.504
(B) TASK/Non-Task CONDITION	0.377
AxB Interaction	0.211

Table 29 Observed Power for Graphics Impression

Summary of Results

Table 25 illustrates the significance value results for all dependent variables under the two independent variables (and their interaction) investigated in this study. The results revealed that the first hypothesis was accepted. The first hypothesis, which is a null hypothesis, stated that there would be no significant difference for fidelity between groups for identifying problems. The second hypothesis was also accepted. The second hypothesis, which is an alternative hypothesis, stated that would be a significant difference between task/non-task conditions for finding problems. The null hypothesis is rejected for the second hypothesis. The third hypothesis was rejected. The third hypothesis, which was an alternative hypothesis, stated that there would be an interaction effect between task/non-task condition and fidelity for identifying problems. There was a failure to reject the null hypothesis for the third hypothesis.

With regards to severity, a significant difference was only observed for the condition of fidelity. The null hypothesis was rejected for the variable of fidelity.

Participants' overall impressions resulted in a significant difference for fidelity in all four cases except for graphics. The null hypothesis was rejected in these cases for fidelity. In the situation with graphics, the null hypothesis was accepted, but it may not be justified because the power and effect size were low in this case.

Dependent Variables	Independent Variables		
	Fidelity	Task/Non-Task	Interaction
# of Problems Identified	N	Y	N
Avg. Problem Severity	Y	N	N
Overall Impressions	Y	N	N
Impression of Information Content	Y	N	N
Impression of Site Navigation	Y	N	N
Impression of Graphic Design	N	N	N
Y = Significant difference | N = No Significant difference			

Table 30 Significant Differences Found for all Variables

Discussion

The purpose of this research was to investigate the effects of fidelity and task/non-task scenarios on usability testing. The first two hypotheses were accepted. A significant difference was observed for task/non-task condition, and no significant difference was observed for the fidelity condition. The third hypothesis however was rejected. No significant difference was observed for the interaction effect between the two variables (fidelity and task/non-task condition). This suggests that a non-task condition is a better way to obtain data regarding problems from users in usability testing. The results from severity and user impressions suggest that users have higher expectations of high fidelity websites than of low fidelity web sites.

The condition of task was found to be significantly different with a relatively high power, and a substantial effect size. This means that the condition of task accounted for 1/5th of the difference in the scores for the groups. However, the condition of fidelity made no significant impact performance. Since the non-task condition identified more problems than the task condition, this suggests that in the early stages of web development, an organization or individual does not need to create scenarios or tasks for participants to perform. The result of the post-hoc comparisons did not yield a significant difference. This could be a result of the small and limited sample size used, seeing that each condition contained five participants. It is also important to recognize the data from the participant profile. Usability tests are run on a wide variety of people. Who are typical users of this product or system? What are their characteristics (age, education, profession, geographic location, etc.)? These are some important questions that are addressed when usability testing is performed. So the information gathered here, while valid, may not be generalized to all usability tests. In the case of any usability testing situation it is important to understand the unique set of users that you are dealing with. Users can be the vast majority of people across the world or a tiny subset of people who have a specific education, age, and other characteristics. So these results can be generalized to a population that closely fits the profile data.

The difference observed between task conditions is may likely due to psychological issues. When people are asked to perform a task and asked to identify all problems while performing that task, it appears that they give too much focus to the initial task. But when participants are only asked to look for problems, they appear to look everywhere. By giving participants a task to perform while looking for problems,

they may develop a tunnel vision, which hampers their performance. Instead of looking at all possible problems everywhere participants may only be looking at problems that relate to their tasks. The non-task condition appears to put participants in a seemingly more flexible condition to find problems, such that they find more than the task condition. However this task and non-task condition only accounted for about 20% of the variance. This means further research is needed before an informative model can be developed to explain the user experience such that we can maximize their performance in usability testing.

The result of severity scores showed that there was a significant difference in the fidelity condition, such that as fidelity increased perceived problem severity increased. This would suggest that if the same problems are present in all levels of fidelity, users believe the problem is more severe if it exists in a higher level of fidelity than a lower level of fidelity. This in some way is something we would naturally expect. When an early draft is prepared we may view problems as less severe because there is time to change and improve the draft. However when a final report is presented that same problem may be viewed as severe because it should have been identified and changed before the final report was finished. As Usability Engineers learn more about the users needs, more things can be addressed, in the case of a web site or software tool. The more the organization or expert learns the better the product becomes, hence there should be less problems. So we have a lot higher expectation for the final product than an earlier prototype.

User impressions of the site were significant (with the exception of graphics) such that the higher fidelity received more positive impressions than lower fidelities. People

are generally more impressed with the higher fidelity models than the lower fidelity ones. But results from severity scores suggest that users are judging the overall layout of the site somewhat independently of the severity scores they provided. Users feedback could possibly be a result of the nature of the web site used and the time period given for the task of identifying problems. Had a much larger site with a lot more problems been used, this plethora of problems should influence users enough such that their overall impressions reflect the severity scores they provided.

Limitations

This research also has some limitations which need to be addressed if it is to be understood entirely. The sample size used was small, however it was sufficient to find statistical significance for the task/non-task condition with a relatively high power. If a larger sample was used power could be increased. An alpha of .05 is the standard according to the (APA) American Psychological Association. It may also be very meaningful to use an alpha of .10 or .15 depending on the situation. Another limitation is the sample used. The results of the profile questionnaire describe the type of users that participated. Participants were generally very similar which is expected from a college population. However, many different types of users view web sites. These users vary in age, education, and culture among other things. So the results may only apply to usability testing with a group similar to the one used in this study. Further research is needed to discover if this is true for all users. The type of web site used was designed to inform people about the Human Factors department at Embry-Riddle Aeronautical University. This site was small, contained a lot of information, and had a small amount of

functionality. However, the internet is populated with many different types of web sites. Some make use of multimedia, while others might make use of functions such as purchasing a product or accessing a forum. It is important to notice these differences, as the results found here may or may not apply to those instances. Again, further research will be needed to ascertain the validity of these results across a wide variety of web sites. The time used for the task of identifying problems, whether users were in a task or non-task group, is also a limitation. Given more time users may likely find more problems, however this particular website was small, and not a lot of time was needed. Given a different type of website, more or less time will likely be needed.

Future Direction

In conclusion what we can learn from this research is that given this type of user group the condition of task/non-task influences (about 20%) their ability to identify problems on a website. These results may help to lead future research such that a model of user performance can be developed. This model would then drive the development of more refined usability testing. And the better the usability testing, the better information is abstracted from users. The end results are better products for users and increased profits. Fidelity, had no impact on identifying problems, however it does effect users' impression of a web site. As a website moves further through development users expect it to adhere to higher standards. This is why when the same problems persist in various fidelity levels, users have a better impression of the lower fidelity than the higher fidelity design. This should somewhat guide usability testing such that there should not be a rush to jump to

high fidelity testing unless a good amount of information is gathered and implemented first.

Some ideas researchers may be want to investigate are the effects of culture or education on users' ability to identify problems. These particular two variables may account for the nature in which people approach problems. Seeing how people approach problems in many different ways this may provide some useful information so that a mental model can be developed and direct future research. The ultimate goal here would be that an in depth understanding of users would be available to help guide development of the most effective usability tests.

Practical Implications

The practical implications of this study is such that if the identification of problems is the primary concern of web developers that a non-task scenario should be used. Giving users a task only appears to lessen their performance. The nature of usability testing is such that it must conform to time and money, so developers have to get things right and spend as little time as possible acquiring information from users. Task based scenarios appear to be more appropriate in the situation where developers are concerned with users performance on an almost completed product. This performance could be measure in many different ways such as time, accuracy, and error made. This research was intended to give developers as well as researchers some insight as to what conditions maximize user performance. In usability testing, developers wish to gather information from users as quickly and as accurately as possible. So it is important to know which kinds of tests are the most efficient. Most usability tests have users perform typical tasks

or be placed in typical situations. These tasks are used to guide users in giving feedback as to how a website can be improved.

Another important thing to consider is that users appear to have high expectations of high fidelity products. A lower fidelity appears to be more like a rough draft or an unrefined idea. So users appear to be more lenient. They rate problems less severe and have general higher impressions of a low fidelity web site. A high fidelity web site appears more like a final or completed product. User will judge this more strictly, and have lower impressions of this level of fidelity if important problems have not been addressed. With this in mind, it may be more beneficial to use low fidelity testing until developers are confident they have addressed most or all of the problems and issues with the web site. Making a rush to higher fidelity testing while it looks more professional also costs more money to maintain or manipulate. A lot of information can be gathered using lower fidelity designs so that when developers move to high fidelity testing only minor problems may be left to address.

References

Alreck, Pamela L., and Settle, Robert B. (1994). *The Survey Research Handbook.* Irwin

Professional Publishing, Chicago, IL.

Bailey, Robert W. (1982). *Human Performance Engineering: A Guide for System*

Designers. Englewood Cliffs, NJ: Prentice-Hall.

Dumas, JS, and Redish, Janice (1993). A Practical Guide to Usability Testing, Albex:

Norwood, NJ.

Gould, J. D. and Lewis, C. Designing for Usability: *Key Principles and What Designers*

Think. Communications of the ACM, 2 (3), March 1985, 300 – 311.

Greenbaum, Thomas L. (1997). The Handbook for Focus Group Research. Thousand

Oaks CA: Sage Publications Inc.

Hong, Jason I., Li, Francis C., and Lin, James (2001). *End-User Perceptions of Formal*

and Informal Representations of Web Sites, Adjunct Proceedings of CHI 2001,

Seattle WA, 385-386..

Landay, J. A. (1996). *Interactive Sketching for the Early Stages of User Interface Design.*

Technical Report CMU-CS-96-201, Carnegie Mellon University, Pittsburgh, PA

Lin, J., Newman, M. W., Hong, J. I. and Landay J. (2000). A. *DENIM: Finding a Tighter*

Fit Between Tools and Practice for Web Site Design. In Proceedings of Human

Factors in Computing Systems: CHI 2000. The Hague, Ntherlands.

McCracken, Daniel D. and Wolfe, Rosalee J. (2004). User-Centered Website

Development. Upper Saddle River, NJ: Pearson Education, Inc.

Mehlenbacher, Brad. (1993). *Software Usability: Choosing Appropriate Methods*

*for Evaluating Online Systems and Documentation.*SIGDOC 93: The 11th Annual

International ACM SIGDOC Conference Preceedings. NY, NY: ACM, 209-222.

Newman, M. W. and Landay, J. A. Sitemaps, Storyboards, and Specifications: A Sketch of Web Site Design Practice. In Proceedings of Designing Interactive Systems, DIS 2000 New York City 263 – 274.

Nielson, Jacob. (1993). *Usability Engineering.* Chestnut Hill, MA: Academic Press.

Nielson, Jacob. (1994). Heuristic Evaluation. In Nielson, J., and Mack, R. L. (Eds.), *Usability Inspection Methods.* John Wiley & Sons, New York, NY.

Nielson, Jacob. (2000). *Designing Web Usability.* Indianapolis, Indiana: New Riders Publishing.

Nielson, Jacob. (2000). Why *You Only Need to Test With 5 Users.* Communications of Jacob Nielson's Usable Information Technology. http://www.useit.com/alertbox/

Ratner, Julie. (2003). Human Factors and Wed Development. Mahwah, NJ: Larence Erlbaum Associates, Inc.

Rubin, Jeffrey. *Handbok of Usability Testing.* John Wiley & Sons, New York, NY.

Sazegari, Sara (1994). *Designing a usability lab: a case study from Taligent.* Behaviour & Technology 13(2): 20 – 24.

Sternberg, Robert J. (2003). Cognitive Psychology. Belmont, CA: Thompson Learning Inc.

Vora P. and Helander M. (1995). *A Teaching method as an alternative to the concurrent think-aloud method for usablity testing,* in Proceedings of the Sixth International Conference on Human-Computer Interaction, Tokyo, Japan.

Walker, Miriam, Takayama, Leila, and Landay,James A. (2002). *High-Fidelity or Low-*

Fidelity, Paper or Computer? Chosing Attributes When Testing Web Prototypes.

In Proceedings of the Human Factors and Ergonomics Society 46th Annual

Meeting Baltimore, Maryland.

Appendix

User Profile Questionnaire

1. Age:_____

2. Gender:_____

3. Major:_____

4. Year: Freshman   Sophomore   Junior   Senior   Graduate

5. How long have you been using desktop and/or laptop computers (PC/Macintosh)

    a. less than 6 months
    b. less than 1 year
    c. less than 2 years
    d. less than 5 years
    e. more than 5 years

6. How many hours per week do you spend browsing the internet (surfing the web)?

7. Please circle all that apply

    a. I rarely use a computer outside of school work
    b. I am familiar with a few applications/programs
    c. I am familiar with a wide variety of applications/programs
    d. I have a limited (novice) understanding of the applications/programs I use
    e. I have an intermediate understanding of the applications/programs I use
    f. I have an in depth understanding of the applications/programs I use
    g. I have experience with programming language(s)
    h. I have experience with multiple Operating Systems
                    (e.g. Windows, Linux, and MAC OSX)

8. Do you have any experience in Web Design? (if yes please elaborate)

    a. yes_____
    b. no

Post Scenario Questionnaire

Please circle the answer that best describes your experience.

Impressions

This site was well organized

        1 – I strongly disagree
        2 – I disagree
        3 – Neutral
        4 – I agree
        5 – I strongly agree

This site was helpful and useful

        1 – I strongly disagree
        2 – I disagree
        3 – Neutral
        4 – I agree
        5 – I strongly agree

This site was easy to use

        1 – I strongly disagree
        2 – I disagree
        3 – Neutral
        4 – I agree
        5 – I strongly agree

Information and Overall Content

This site offered a lot of useful information about the Human Factors and Systems program.

        1 – I strongly disagree
        2 – I disagree
        3 – Neutral
        4 – I agree
        5 – I strongly agree

The information on this site was well laid out and easy to read.

        1 – I strongly disagree
        2 – I disagree
        3 – Neutral
        4 – I agree

5 – I strongly agree

Navigation

It was easy to move from page to page and find where I needed to go.

    1 – I strongly disagree
    2 – I disagree
    3 – Neutral
    4 – I agree
    5 – I strongly agree

All hyperlinks were meaningful and straightforward (allowed me to go to where I needed to be)

    1 – I strongly disagree
    2 – I disagree
    3 – Neutral
    4 – I agree
    5 – I strongly agree

Graphics

The Graphics (color and images) assisted in making this site easier to use

    1 – I strongly disagree
    2 – I disagree
    3 – Neutral
    4 – I agree
    5 – I strongly agree

The Graphics (color and images) made this site appear professional

    1 – I strongly disagree
    2 – I disagree
    3 – Neutral
    4 – I agree
    5 – I strongly agree

Consent Form   #_____

      Thank you for participating in this research experiment. Your evaluation can help improve the testing of web sites and other user interfaces. Your performance during this evaluation will be recorded via video camera. It is emphasized that your participation will be kept confidential. If at any point of this experiment you feel uncomfortable or object to the procedure, you are free to leave. Also, in order to participate in this experiment you must not have taken IT 240 or higher, or have extensive knowledge of web development. By signing below you agree to these terms.

Name_____

Date_____

Debriefing Form

You have just participated in a study which is looking at the effectiveness of different testing techniques. Your participation is greatly appreciated, and will be kept confidential. If you have any concerns feel free to e-mail me at dolec744@erau.edu Also, I ask that you do not discuss the details of this experiment with anyone for the duration of the summer.

Task Scenario (for High and Medium Fidelity)

While you are answering the questions below, I ask that you thoroughly identify all problems and difficulties you encounter with this web site. Please do so, on the blank sheets of paper you are provided in this packet. I ask that you list each problem along with a description of that problem, where the problem was located, and how severe that problem was on a scale of 1 – 5 (1 = minute problem, 5 = extremely severe problem). You will be provided with more paper if needed. You will have 20 minutes to complete this, <u>so please continue to search for problems until the time is up</u>.

Example: when I went to this site I had problems on the home page finding my way around. I could not locate the contact information. I rate this problem as a 4. It was a severe problem, but not extremely severe.

Please ask any questions you have NOW before we begin !!! If you have no questions inform the instructor that you are ready.

1. You are now at the homepage. Locate the "Faculty & Staff" link, and click it to go to that page.
2. You may now answer the following question with the information present on this page.

What is the name of the Department chair of the Human Factors and Systems department?

---

1. Now, locate the "Home" link, and click it to go to that page. You may now answer the following question with the information present on this page.

You are interested in this department and are looking for a phone number with which to contact someone for further information. Write down the department phone number.

---

1. Now, locate the "Research & Consultation" link, and click it to go to that page. You may now answer the following question with the information present on this page.

How many research activities involve aircraft pilot testing of some kind? _____

1. Now, locate the "Home" link, and click it to go to that page.
2. Locate the "Program Information" link, and click it to go to that page.
3. Locate the "Bachelor's Degree" link, and click it to go to that page. You may now answer the following question with the information present on this page.

What is the minimum required amount of Human Factors and Psychology classes needed

for the Bachelor degree in Human Factors Psychology? _____

1. Now, locate the "Programs" link, and click it to go to that page.
2. Locate the "Five Year Master of Human Factors and Systems" link, and click it to go
to that page. You may now answer the following question with the information present
on this page.

How many degrees will a student be awarded for successful completion of the Five-Year

Masters Degree Program in Human Factors and Systems? _____

1. Now, locate the "Programs" link, and click it to go to that page.

2. Locate the "Master of Human Factors and Systems" link, and click it to go to that
page. You may now answer the following question with the information present on this
page.

What two main choices is a graduate student in the Human Factors and Systems faced

with? _____

1. Now, locate the "Home" link, and click it to go to that page.
2. Locate the "*To find out more about HF*" link, and click it to go to that page. You may
now answer the following question with the information present on this page.

What exactly is Human Factors?_____

Task Scenario (for Low Fidelity)

While you are answering the questions below, I ask that you thoroughly identify all problems and difficulties you encounter with this web site. Please do so, on the blank sheets of paper you are provided in this packet. I ask that you list each problem along with a description of that problem, where the problem was located, and how severe that problem was on a scale of 1 – 5 (1 = minute problem, 5 = extremely severe problem). You will be provided with more paper if needed. You will have 20 minutes to complete this, so please continue to search for problems until the time is up.

Example: when I went to this site I had problems on the home page finding my way around. I could not locate the contact information. I rate this problem as a 4. It was a severe problem, but not extremely severe.

Please ask any questions you have NOW before we begin !!! If you have no questions inform the instructor that you are ready.

**page numbers are located at the bottom right of each page.

1. You are now at the homepage. Locate the "Faculty & Staff" link, and go page 3.
2. You may now answer the following question with the information present on this page.

What is the name of the Department chair of the Human Factors and Systems department?

---

1. Now, locate the "Home" link, and return to page 1. You may now answer the following question with the information present on this page.

You are interested in this department and are looking for a phone number with which to contact someone for further information. Write down the department phone number.

---

1. Now, locate the "Research & Consultation" link, and go to page 13. You may now answer the following question with the information present on this page.

How many research activities involve aircraft pilot testing of some kind? _____

1. Now, locate the "Home" link, and return to page 1.
2. Locate the "Program Information" link, and go to page 14.

3. Locate the "Bachelor's Degree" link, and go to page 15. You may now answer the following question with the information present on pages 15 and 16

What is the minimum required amount of Human Factors and Psychology classes needed

for the Bachelor degree in Human Factors Psychology? _____

1. Now, locate the "Programs" link, and go to page 14.
2. Locate the "Five Year Master of Human Factors and Systems" link, and go to page 17.
You may now answer the following question with the information present on this page.

How many degrees will a student be awarded for successful completion of the Five-Year

Masters Degree Program in Human Factors and Systems? _____

1. Now, locate the "Programs" link, and go to page 14.
2. Locate the "Master of Human Factors and Systems" link, and go to page 18. You may
now answer the following question with the information present on pages 18 20.

What two main choices is a graduate student in the Human Factors and Systems faced

with? _____

1. Now, locate the "Home" link, and return to page 1.
2. Locate the "*To find out more about HF*" link, and go to page 2. You may now answer
the following question with the information present on this page.

What exactly is Human Factors?_____

_____

Non-Task Scenario (for all levels of fidelity)

Now please browse the site as if you were a student interested in this department. I ask that you thoroughly identify all problems and difficulties you encounter with this web site. Please do so, on the blank sheets of paper you are provided in this packet. I ask that you list each problem along with a description of that problem, where the problem was located, and how severe that problem was on a scale of 1 – 5 (1 = minute problem, 5 = extremely severe problem). You will be provided with more paper if needed. You will have 20 minutes to complete this, so please continue to search for problems until the time is up.

Example: when I went to this site I had problems on the home page finding my way around. I could not locate the contact information. I rate this problem as a 4. It was a severe problem, but not extremely severe.
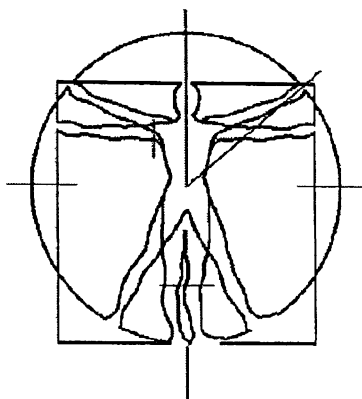
Low Fidelity Prototype
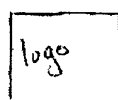
| logo | Embry- Riddle Aeronautical University | | Department of Human Factors & Systems |

Faculty & Staff

Research & Consultation

Program Information

Important Links

Human Factor

logo

Opening your mind to the field of Human Factors & Systems design ... To find out more about HF

Embry- Riddle Homepage

For information regarding Embry-Riddle University visit us at: www.erau.edu

Design Challenge

Coming soon the Design Challenge Here's your chance to test your skill in human centered design!

New on this Site

HPSAA II Conference March 22-25, 2004 Daytona Beach, FL

Department of Human Factors & Systems
Embry- Riddle Aeronautical University, 600 South Clyde Morris Blvd., Daytona Beach FL 32114-3900
Tel: (+1)(386) 226-6790 Fax: (+1)(386) 226-7050

(1)

Medium Fidelity Prototype

EMBRY-RIDDLE
AERONAUTICAL UNIVERSITY

DEPARTMENT OF
HUMAN FACTORS & SYSTEMS

**FACULTY &
STAFF**

**RESEARCH &
CONSULTATION**

**PROGRAM
INFORMATION**

**IMPORTANT
LINKS**

To find out more about HF

EMBRY-RIDDLE HOMEPAGE

For information regarding Embry-
Riddle University visit us at
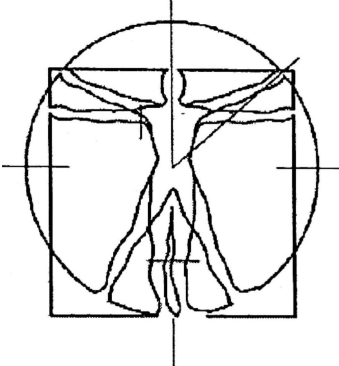www.erau.edu

DESIGN CHALLENGE

Coming soon the Design Challege " Here's
your chance to test your skill in human
centered design!

New on this Site

HPSAA II Conference
March 22 - 25, 2004 Daytona Beach
FL

Embry Riddle Aeronautical University, 600 South Clyde Morris Blvd., Daytona Beach, FL 32114-3900
Tel (+1) (386) 226 6790 Fax (+1) (386) 226-7050

63

High Fidelity Prototype