TEXAS STATE
UNIVERSITY
SAN MARCOS

Department of Computer Science
San Marcos, TX 78666

Report Number TXSTATE-CS-TR-2008-9

An Effort-Based Approach to Measuring Software Usability

Carl J. Mueller   Oleg V. Komogortsev   Dan Tamir   Liam Feldman

2008-10-21

Contents

Table of Figures

Abstract

Design and implementation of usable human computer interface (HCI) systems involves expensive, primarily cognitive based, usability testing and evaluation techniques. This complicates the development process and may cause software companies and software engineers that are more familiar with objective testing methodologies to reduce or completely avoid the usability testing stage, reverting to best practice techniques, and producing HCI systems that lack usability. This research is based on the assumption that usability of HCI systems is directly related to the amount of mental and physical effort expended by the user throughout the interaction. It explores and exploits the utility of an objective, relatively easy to measure, and engineering oriented usability metric. A mathematical model of interaction effort is formulated. The model transforms data related to primitive interaction events such as keyboard keystrokes, mouse key clicks and Mickys traversed by the mouse along with eye tracking data into an effort metric. A carefully crafted set of user interaction goals employing scenario based test design techniques is implemented. Data is collected using logging programs that record goal completion time along with keyboard, mouse, and eyes interaction events. The recorded information is reduced to a statistically meaningful data-set that is used to evaluate the validity of the research assumptions. Experimental results support the hypothesize. Furthermore, they are prompting several interesting finding that merit further research and investigation. This is the first research that carries the intuitive idea of relation between effort and usability all the way to the “field” by recording and processing effort based metrics obtained from subjects while interacting with real complex systems.

1 INTRODUCTION

Poor software usability not only causes user dissatisfaction but also can lead to substantial development cost overruns. Software developers can use a wide variety of tools (prototyping, inspection, usability testing, iterative processes, etc.)[23] assuring the software they produce has the usability desired. Considering the large number of user complaints about software usability, these techniques may not address the problem efficiently. Furthermore, the challenges presented by usability issues may not lie solely in the tools and techniques used in the development process. Physiological and psychological characteristics and sociological conditioning heavily influence software usability making it possibly one of the most subjective attributes of software quality. Many software engineers are not familiar with the factors influencing usability and are frequently uncomfortable with the entire topic. One approach that may make software engineers more comfortable with the topic of usability is to recast it into terms and concepts that are more familiar to the software engineering community. Investigating an objective and engineering-based methodologies of evaluating software usability is the focus of this research.

Because software usability is highly subjective, evaluation requires observing a number of human subjects while engaged in using the system. Interpreting these observations necessitates adding a psychologist or a person skilled in psychological evaluation to the testing team. Some developers do not view usability testing as productive evaluation because these evaluations usually indicate an area where the subjects had problems and does not necessarily point to a specific issue with the software. Because these evaluations do not necessarily identify specific defects or issues, it can make developers and managers extremely frustrated. Being close to the project deadline can amplify this frustration, especially when there is no way to determine how much time and effort are required to identify and modify the usability issues with the software. Because of the uncertainty and expense of usability evaluations, some managers are reluctant to include formal usability testing in their development plan. Instead of using testing, these managers prefer to rely on best practices, templates, and inspections to establish software usability.

The actual challenge of developing usable software may lie in the lack of a clear and concise understanding of what too many software engineers view as a fuzzy concept. Not all authorities on software quality provide a definition of usability. Some authorities recommend usability testing but only provide a checklist of things to investigate [9, 15, 20]; and these authorities are, for the most part, balancing between systems with "card input" and interactive systems. Most quality models [1, 14, 19, 21] provide a relatively consistent and concise definition of usability, but the attributes used to characterize the many facets of usability are not consistent. This research uses the characterization of usability provided in the ISO/IEC 9126 because it is one of the more recent quality models, it is an industry standard, and it provides a measurement system for each of their quality attributes and characteristics. This standard defines usability as "the capability of the software product to be understood, learned, used, and attractive to the user when used under specified conditions" [1], with the following characteristics: Understandability, Learnability, Operability, Attractiveness, and Compliance.

Understandability is the ability of a user to understand the capabilities of the software and if it is suited to accomplish specific goals. It is measured by providing the user with a tutorial or software documentation and then evaluating the users' knowledge to determine the users' level of understanding of the software's functionality, operation, and input/output data [2]. It also recommends using cognitive monitoring techniques to evaluate the subject's response. Cognitive monitoring techniques are using one-way mirrors or concealed cameras to record the subject's behavior along with evaluation of the findings by a psychology professional.

Learnability describes how easy it is for a subject to learn to use the software. For this characteristic, the standard measures how long it takes the user to learn and perform a task, the number of functions used correctly, and the utility of the help facility [2]. In addition to the measurements, the standard proposes cognitive monitoring techniques. Of the characteristics used by the ISO Quality model to describe software usability, only learnability has deep roots outside of software quality. A German psychologist first introduced a learnability model in the 19th century [13], as a model describing the time required to memorize something. Figure 1 is an illustration of the learning model. In the 1930's, research at Wright-Patterson quantified the notion; and in the 1960's it evolved into an experience curve [3]. This research applied the learning model to an industrial setting by comparing cost per unit verses units developed.

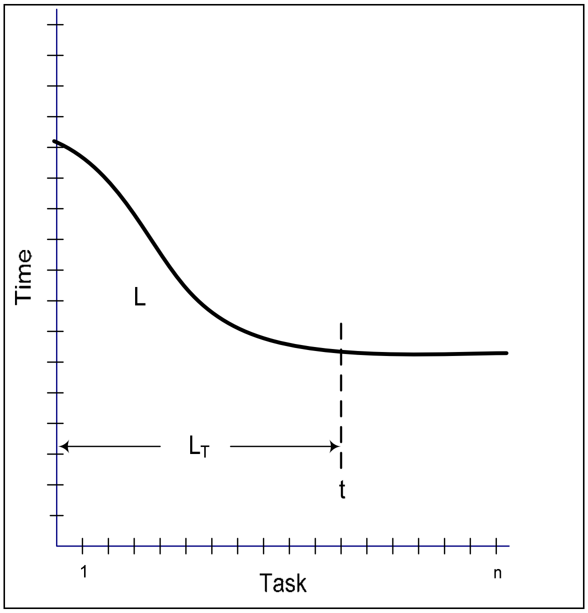


Figure 1 Hypothetical Learning Curve

Operability is the capability of a user to use the software to accomplish a specific goal. Assessing operability requires measuring the following characteristics: Operational consistency; Error correction; Error correction in use; Default value availability in use;

Message understandability, Self-explanatory error messages, Operational error recoverability in use, Time between human error operation in use; Undoability, Customizability, Operation procedure reduction, and, Physical accessibility [2]. Some of these metrics are objective measurements, but many require cognitive monitoring techniques to evaluate.

As the name implies, attractiveness is the appeal of the software to a user. Attractiveness is possibly the most subjective of all the usability characteristics, involving not only sociological and psychological issues but gender and personal taste issues. The ISO/IEC 9126 standard characterizes attractiveness by providing subjects with a questionnaire to evaluate the interface and by observing subjects customizing the appearance to their satisfaction [2].

Compliance measures how well the software adheres to external and internal rules and regulations relating to usability. It is also the most straightforward characteristic to evaluate. Developers compile a list of the required standards, conventions, style guides, and regulations. Then using functional testing techniques, verify that the software complies with them [2].

Even from this short description of the metrics necessary to evaluate usability, it is apparent that designing a usability test would be an extremely time consuming and, therefore, an expensive task. This is a test with potentially a very high cost, which may not identify any specific design or implementation defects or issues. Reducing the high cost of usability testing is difficult because each of the measures proposed by the ISO/IEC standard are good and identify specific problems, and it is not possible to eliminate the use of human test subjects. Another problem with the number of measurements is how to create objective specifications for so many diverse characteristics. Setting objective measurements for all of these characteristics would definitely increase the time necessary to specify requirements. What may be possible is to take a slightly different approach to usability testing using techniques that developers and testers are more comfortable with and could administer without requiring cognitive evaluation techniques.

One possible approach to usability testing might be to design a set of goals or tasks and to measure the effort and time necessary for group of subjects to accomplish each goal [26]. If developers estimated the effort and time necessary to complete each goal, then it would be possible to compare the observed effort with the estimated effort. If the observed effort is greater than the estimated effort, then there is a problem requiring further investigation. After identifying the existence of a problem, developers could trace the observation logs to find where the subjects experienced a problem causing the expenditure of additional effort. Trying to evaluate every characteristics of usability from effort and time to complete a goal may be a bit overreaching, but understandability, operability, and learnability should be within the scope of these measurements.

Section 2 of this report presents the theory surrounding this notion of an effort based usability measurement. Much of the material in this section was presented at the 6th Workshop on Software Quality [25]. From the positive feedback at the workshop, the researchers decided to continue to the next step and began experimentation into the concept. Section 3 describes an experiment conducted to provide empirical data to support the notion that it is possible to measure usability based on an effort based model. Section 4 presents a discussion of the experiment's results. Conclusions and future research topics are presented in Section 5.

The appendices contained within this report provide the details of the experimental procedures used in this experiment. Appendix A contains the goals used by subjects in the experiment. In Appendix B, the forms used for the experiment are illustrated. Appendix C contains the scripts used by facilitators during the execution of the experiment. Raw data from the experiment is described in Appendix D.

2 A HYPOTHESIS FOR EFFORT-BASED USABILITY

2.1 Overview

Many software publishers are claiming that their product requires less effort than the competition. Even though these advertisers provide no objective substantiation for these claims, the fact that this may entice a buyer to purchase the product gives some credibility to the notion that there is a relationship between usability and effort. For this hypothesis, E denotes all the effort required to complete a task with computer software, as defined by the following equations:

$$E = E_{mental} + E_{physical}$$

$$E_{mental} = E_{eye\_mental} + E_{other\_mental}$$

$$E_{physical} = E_{manual\_physical} + E_{eye\_physical} + E_{other\_physical}$$

Where:

E_{eye_mental} the amount of mental effort to complete the task measured by eye related metrics.

E_{other_mental} the amount of mental effort measured by other metrics.

$E_{physical}$ the amount of physical effort to complete the task.

$E_{manual\_physical}$ the amount of manual effort to complete the task. Manual effort includes, but is not limited to, the movement of fingers, hands, arms, etc.

$E_{eye\_physical}$ the amount of physical effort measured by eye movement related metrics.

$E_{other\_physical}$ the amount of physical effort measured by other metrics.

With current technology, directly measuring cognitive activity for a specific task is not practical, therefore, measuring the total effort (E) is not possible at this time. Logging keystroke and mouse activity approximates the manual effort ($E_{manual\_physical}$) expended by a subject. An eye-tracking device allows logging eye position data to estimate amount of mental effort (E_{eye_mental}) and physical effort ($E_{eye\_physical}$) in terms of eye movement metrics. Terms such as E_{other_mental} and $E_{other\_physical}$ are presented as estimation of all other factors that might be contributing to the final effort for task completion.

2.2 Measuring Effort

2.2.1 Manual Effort

Several informal studies indicate that many system users associate the “physical” effort required for accomplishing tasks with the usability of the software. In the case of interactive computer tasks, it may be possible to calculate effort from a weighted sum of mouse clicks, keyboard clicks, Mickeys, etc. The term Mickey denotes the number of pixels (at the mouse resolution) traversed by the user while moving the mouse from a point(x_0, y_0) to a point (x_1, y_1).

The definition of effort uses continuous functions. In practice, given the discrete nature of computer interaction, these measures are quantized by converting integrals to sums. Assume that an interactive task R starts at time t_0. We define the effort at time t as:

$$E_{manual\_physical}(t) = \frac{1}{t - t_0}\int_{t_0}^{t}(w_1 \times mic(t) + w_2 \times mc(t) + w_3 \times mk(t) + w_4 \times p(t))dt$$

Where: $mic(t)$, $mc(t)$, $mk(t)$ are (respectively) the number of Mickys[1], the number of mouse clicks, and the number of keystrokes by a subject during the time interval $t - t_0$. Furthermore, $p(t)$ is a penalty factor that measures the number of times the user switched from mouse to keyboard or vice versa during the interval. Note that $E(t)$ is a monotonically increasing function.

2.2.2 Mental Effort

Mental effort is essentially the amount of brain activity required to complete a task. The human brain is an extremely complex system processing multitudes of sensory inputs in parallel, and requires sophisticated equipment to assess cognitive activity related to a specific task. Brain activity related to a task can be approximated by processing eye movement data recorded by an eye tracker [11]. Modern eye trackers are similar to web cameras, without any parts affixed to the subject's body. Eye trackers provide useful data even in the absence of overt behavior. With this device, it is possible to record eye position data and classify it into several eye movement types useful for eye related effort assessment. The types of the eye movements are: 1) fixation – eye movement that keeps an eye gaze stable with regard to a stationary target providing visual pictures with high acuity, 2) saccade – very rapid eye movement from one fixation point to another, and 3) pursuit – stabilizes the retina with regard to a moving object of interest [11]. The Human Visual System without dynamically moving targets does usually not exhibit pursuits. Therefore, parameters related to smooth

[1] A Mickey is the number of screen pixel traversed by the mouse between two mouse events.

pursuit are not discussed in this paper. In addition to basic eye movement types, eye tracker systems can provide biometric data such as pupil diameter.

Many researchers consider the following metrics as a measure of the cognitive load. Hence, these metrics facilitate the estimation of mental effort.

Average fixation duration: Average fixation duration, measured in milliseconds, indicates cognitive load that can be interpreted as a difficulty in extracting information or as an indication that an interface object is more engaging in some way [22], and is.

Average pupil diameter: Eye tracking systems enable measuring biometric data such as pupil diameter. Pupil size, measured in millimeters, can be indicative of the high cognitive effort [22].

Average saccade amplitude: Saccade amplitude, measured in degrees, indicates meaningful cues, and relates to task completion effort [22]. To certain extent large average saccade amplitude, represent lower mental effort.

As with the definition of manual effort, the definition of mental effort uses continuous functions that are quantized by converting integrals to sums. Assume that an interactive task R starts at time t_0. The mental effort at time t is defined as:

$$E_{eye\_mental}(t) = \frac{1}{t - t_0} \int_{t_0}^{t} \left(w_5 \times fix\_dur(t) + w_6 \times \mathrm{pup\_d}(t) + w_7 \frac{1}{\mathrm{sac\_amp(t)}} \right) dt$$

Where: fix_dur represents fixation duration, pup_d is the pupil diameter and sac_amp represents saccade-amplitude. Occasionally, eye-tracking devices produce data that is below a reliability threshold. Periods including non-reliable data were excluded from integration.

2.2.3 Eye Physical Effort

Ideally, effort expended by the Human Visual System (HVS) to complete a task is represented by the amount of energy spent by HVS during the task. The energy expanded is dependent on the amount of eye movements exhibited by the HVS, the total eye path traversed and the amount of force exerted by each individual extraocular muscle force during each eye rotation. These terms are defined below:

Number of saccades: High number of saccades indicates extensive searching, therefore less efficient time allocation to task completion [22]. Increased effort is associated with high saccade levels.

Number of fixations: Due to non-optimal representation, overall fixations relate to less efficient searching [22]. Increased effort is associated with high amounts of fixations.

Total eye path traversed: This metric, measured in degrees, presents the total distance traversed by the eyes between consecutive fixation points during a task. The length of the path traversed by the eye is proportional to the effort expended by the HVS to achieve the goal.

Extraocular muscle force: The amount of energy, measured by grams per degrees per second, required for the operation of extraocular muscles relates to the amount of force that each muscle applies to the eye globe during fixations and saccades. Based on the Oculomotor Plant Mechanical Model, it is possible to extract individual extraocular muscle force values from recorded eye position points [16]. The amount of force from each muscle can be summed to calculate the total force.

The total eye physical effort can be approximated by:

$$E_{eye\_physical}(t)$$

$$= \frac{1}{t - t_0}\int_{t_0}^{t}(w_8 \times fix\_count(t) + w_9 \times \text{sac\_count}(t) + w_{10}$$

$$\times \text{ eye\_distance(t)} + w_{11} \times \text{extraocular\_force(t)}dt$$

Where: fix_count , sac_count , eye_distance, and extraocular_force represent the total amount of fixations, the total amount of saccades, the total amount of eye distance traversed, and the total amount of force exerted by the extraocular muscles respectively. The integration excludes periods of time that include non-reliable data.

2.3 Formalization of an Effort-Based Usability Model

Consider the following example. Assume a set of n subjects selected at random complete a set of k tasks or goals. Further, assume that the subjects are computer literate but unfamiliar with the application under evaluation. The objective of each goal is to make travel reservations, *and each goal requires about the* same effort. After the subjects complete all of the goals, an average of the effort (E_{avg}) and the time (T_{avg}) for each goal is calculated. If it is possible to measure the effort expended by each subject, then plotting the average effort (E_{avg}) for each task should produce a graph similar to the one illustrated in Figure 2, when using subjects that have limited knowledge of the application. Like learning and experience curves, an effort curve is plotting the expenditure of a resource to accomplish a task. It is the

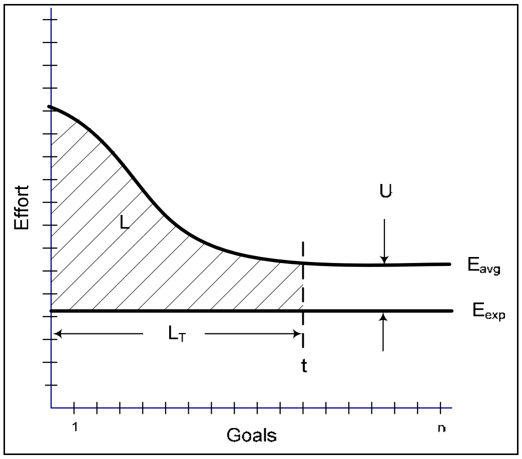


Figure 2 Hypothetical Effort Model

hypothesis of this research that usability, and specifically, operability, learnability, and understandability are functions of effort.

It is possible to view usability from a static and dynamic perspective. Static usability is established when the human interface is designed and does not change with user customization or activity. Under this assumption, it is possible to ignore the “shape” of the curve of $E(t)$, and only use the “final” effort, that is, the accumulated effort at time of completion of tasks. In order to derive the relation between $E(t)$ and static usability, we define the effort associated with an interactive task R in the following way:

$$\widehat{E_R}(t) = \frac{1}{t_c - t_0}\int_{t_0}^{t_c}\big(w_1 \times mc(t) + w_2 \times mk(t) + w_3 \times mic(t) + w_4 \times p(t) + w_5 \times e(t)\big)dt$$

Where t_c is the time of task completion for the task R. Note that the division by the factor $t_c - t_0$ eliminates the dependency of $\widehat{E_R}$ in the ‘time to completion’ factor. Assuming that $mc(t_0) = mk(t_0) = mic(tt_0) = p(t_0) = e(t_0) = 0$ and define the total effort associated with R, to be:

$$\widehat{E_R} = \frac{1}{t_c - t_0} \times \big(w_1 \times mc(t_c) + w_2 \times mk(t_c) + w_3 \times mic(t_c) + w_4 \times p(t_c) + w_5 \times e(t_c)\big)$$

Normalizing $\widehat{E_R}$, and defining E_R (the normalized effort associated with R) by setting $t_c = 0$ and $t_0 = 1$. At this stage of the research, it is possible to assume that $w_1 = w_2 = w_3 = w_4 = w_5 = 1$. Finally, denoting $f(1)$ by f to obtain:

$$E_R = mc + mk + mic + p + e$$

Where $mc, mk, mic,$ and p denote the total number of mouse clicks, keyboard clicks, mickeys, and mouse to/from keyboard switches throughout the process of completing the task R. The total physical eye effort is represented by e.

One feature added to the effort model not found in the learning model (see Figure 1) is the notion of expected effort (E_{exp}) or designer effort. At the time of an application's deployment, the people who know the software best are the developers. Therefore, they should expend less effort in completing specific tasks, and provide a point of comparison. Thus, the designer expected effort is a single number that represents the "ideal" (with respect to minimum effort) way to interact with the system in order to accomplish a task. In order to associate E_R with operability, learnability, and understandability it is necessary to formalize the concept of design expect effort. Let $R(\bar{x})$ be a task with a parameter vector $\bar{x}$ and let S be a sequence of user interactions that can accomplish $R(\bar{x})$ from scratch. For example, $R(a, b)$ can denote the task of reserving a flight from a city a to a city b; then S can include interactions related to subtasks such as checking prices for different airlines, at slightly different arrival / departure times, or from different airports within the cities. Defining the designer expected effort for $R(\bar{x})$ as:

$$E_{exp} = \min_{s \in S} E_{R(\bar{x})}$$

It is now possible to define operability, learnability, and understandability in terms of designer expected effort (E_{exp}).

2.3.1 Operability

To elaborate, consider two possible designs ($D^{(1)}$ and $D^{(2)}$) of an interactive system for flight reservation enabling the task $R(a, b)$. Let $E_{exp}^{(1)}$ and $E_{exp}^{(2)}$ denote the designer expected effort for the designs $D^{(1)}$ and $D^{(2)}$ respectively and assume that $E_{exp}^{(1)} < E_{exp}^{(2)}$. Then, per the definition of operability, the operability of design $D^{(1)}$ is better than the operability of design $D^{(2)}$.

2.3.2 Understandability

Lack of understandability may result in non-efficient usage of the system or using the system for a task that is different from any task defined at design time. In this case, the user effort may converge to a value that is higher than the designer expected effort (E_{exp}). The difference between the user actual effort (E_{act}) and the designer expected effort (E_{exp}) may be a useful measure for understandability, and is depicted in Figure 2.

2.3.3 Learnability

It is possible to measure learnability as the rate of convergence of the average user effort (E_{avg}) to the ideal effort E_{exp}. Alternatively, we can define learnability in terms of the root mean square error. Here the error is the difference between the average user effort (E_{avg}) and the designer expected effort (E_{exp}) at a given task. Yet another (and similar) measure can be the area of the difference between the learning curve and the curve formed by the fixed line at $y = E_{exp}$. Figure 2 depicts the learnability (and understandability) curve. Due to understandability deficiencies, it is possible that the user learning curve does not converge to the designer expected effort (E_{exp}). Hence, the subject is said to have "learned" the system where the curve flattens.

3 EXPERIMENTATION

3.1 Overview

Selecting an objective for the first effort-based metric investigation posed the first research challenge. Determining an appropriate objective proved to be challenging because not all quality models view software usability as a quality attribute. Among those quality models which include software usability as a characteristic [6, 19, 14, 21], each uses slightly different definitions and sub-characteristics to describe usability. One notion consistent in all of these quality models is “ease of learning”, frequently referred to as learnability. Since there is consensus among most quality models that “ease of learning” correlates with one of the characteristics of software usability, the researchers felt that this would be a good place to start the investigation. Since it is not practical to work with varying definitions, researchers elected to use the IS0-9126 Quality Model’s [6] description of software usability for all other characteristics and attributes. Selection of the ISO quality model was made on the basis that its standards provide a complete description of how to measure each usability characteristic and attribute [7]. Additionally, it is one of the more contemporary attempts to describe software quality.

3.2 Planning

On the topic of designing usability tests, there are a number of references that provide limited guidance for the planning of a usability test [12, 14, 21, 24]. These references provide a few general insights into the preparation necessary for a usability test, but they are predicated on using cognitive evaluation techniques and do not address the issues of constructing a usability test by logging actual activity. In the chapter on Higher-Order Testing [20], Glenford Myers provides a list of eight (8) things a developer should consider when designing usability tests. These eight (8) rules are useful, but they do not provide much guidance on how to construct a specific test. After much discussion among the researchers, it was felt that developing an informal research plan in lab-books and meeting notes would probably provide a sufficient basis for an experimental protocol. The Texas State University-San Marcos Human Subject Policy further guided the research formulation.

Researchers decided to evaluate learnability for two similar but varying applications. One reason for evaluating two systems is that it gives researchers some confidence that an effort-based measure of learnability will permit system comparison. Another rationale for this decision is that it will provide more data, and possibly more insight, into using this technique in the evaluation of production software.

For this research, defects were defined as a behavior of the software deviating from a specification. Since researchers did not have specifications for the systems under evaluation, defect identification was not possible. This situation frequently occurs in the evaluation of Commercial-Off-The-Shelf (COTS) software. When evaluating COTS components, anomalistic behavior is sometimes referred to as an issue. Because this test closely models a COTS evaluation, any missing facility or anomalistic behavior is identified as a usability issue.

Conducting this type of research in a university setting offers a number of advantages and disadvantages. A large subject pool is one of the advantages. College students are usually computer literate and are willing to participate in an experiment for a slice or two of pizza. Universities require experiments involving human subjects be approved by the university administration, usually in the form of a review board. Some may view this as disadvantageous, but the approval process requires researchers to think about how the experiments will affect the subjects. One of the most obvious concerns is the safety of the subject, but the anonymity of the subject is also a major concern in the approval processes.

After some discussion, researchers decided to evaluate the learnability and operability of two web-based travel reservation systems. One reason for selecting this application was that it requires subjects to complete a number of none-trivial tasks, such as booking flights, hotel accommodations and rental cars under strict budget constraints. Making travel reservations is a task understood by most of the individuals in the subject pool. After selecting an application, the next step was to determine the number of subjects required. Researchers decided that two separate groups of ten subjects would provide an adequate number of subjects. Nielsen describes a number of different usability testing techniques [21]; the one closest to the type of experiment being conducted is called “logging actual use” and requires six (6) subjects. There is some literature that favors using more subjects [10], based on a probable number of usability defects an individual subject might discover. Because this was similar to a COTS evaluation, the defect probability was assumed low and therefore not an issue. However, since the algorithms used to calculate eye effort [17] require a minimum of ten (10) subjects, two groups of 10 subjects were required.

When developers are designing a test, they usually divide the test objects into a series of test cases. In usability testing, developers design a set of goals or tasks based on a specific usage scenario. Unlike a traditional test script, the goal or task describes only the desired objective. It is the responsibility of the subject to determine the best method to accomplish the objective. For this experiment, researchers created and tested 10 goals. Since the primary objective of the experiment or test was to observe learnability of the two travel reservation systems, each goal or task was designed to contain similar elements. Appendix A.1 illustrates the template used to create tasks or goals.

The decision to forego a small-scale or simulated environment in favor of a commercial application permitted subjects to interact with the system in an unconstrained manner, similar to the software utilization one might observe “in the field.” However, this choice carried with it a set of unique challenges and constraints not typically encountered in the course of usability testing.

The objective of each goal was to induce system interaction while minimizing negative motivation factors such as frustration or confusion. Goals were meant to challenging, but not inordinately difficult to achieve within a reasonable amount of time. All goals had to be completely feasible (i.e. achievable/able to be fulfilled) utilizing both System A and System B.

The task sets for the various goals had to be designed based solely on “trial-and-error” designer interaction with the system interface, as opposed to the more common scenario where test designers would be provided with guidance from system developers. In a sense, the interaction learnability goals had to be “reverse-engineered” from the goal designers’ own process of interacting with and learning about the systems under test. This process, despite being cumbersome, successfully determined a number of goal design requirements, such as:

Subjects should be required to book non-stop flights for the air travel sub-goal. Without this constraint, goals proved to be too trivial to fulfill, causing a decrease in data generation due to reduced subject interaction with the travel system.

Departure and arrival points for the air travel sub-goal should be non-hub, secondary-market cities such as Columbus, Ohio or Scranton, Pennsylvania. Booking travel involving large hub locations such as Chicago or Dallas proved to be too simplistic, but goals involving travel to or from out-of-the-way, poorly-serviced locales were determined to be extremely difficult if not impossible to fulfill. In the case of one of the goals, a destination city was accidentally selected which rendered the goal feasible using System A, but infeasible using System B, necessitating a mid-study correction.

Departure and arrival dates for the air travel sub-goal need to be at least three months into the future. This substantially reduces the risk that fluctuating flight availability or conditions would prevent goal completion.

Amenities for the hotel room sub-goal should be available in a wide range of accommodations. Examples of amenities meeting this requirement included high-speed Internet access, in-hotel restaurant or dining room, and exercise facility. Goals adhering to this requirement compelled subjects to interact in greater depth with the system without causing much searching

Even though the experiment was investigating an objective measure of operability and learnability, it was necessary to have subjects complete a questionnaire after each goal. This document provided a place to record the amount of time it took to complete the goal, whether or not the subject completed the goal, and some information about the subject’s physical state. Information about the subject’s physical state provides insight about fatigue, a factor that can affect performance. As shown in Appendix B.2, the post-goal evaluation form contains only an experiment reference number in lieu of the subject’s name. The Texas State Human Subject Protocol requires maintaining the anonymity of each subject. Only the subject and the test facilitator know the subject’s ID code. A list of participating subjects was maintained to ensure that a subject did not participate more than once, and destroyed when testing concluded.

To familiarize themselves with the eye-tracking equipment, researchers and facilitators attempted to complete at least one goal. The eye-tracking equipment could only be calibrated for one researcher and one facilitator. Although the precise reason for the calibration failures is unknown at this time, the disqualified researchers and moderators had in common the fact that they all wore prescription lenses. Each prescription differed, but two recurring factors were lens prisms and astigmatism correction. Researchers decided to add a subject profile questionnaire to investigate these calibration failures further. The profile captures some basic information about each subject, as shown in Appendix B.1. Like the goal evaluation form, this form identifies the subject only by ID number.

Due to the nature of the testing being conducted, a determination was made in advance that it would be necessary to exclude subjects from participation if they failed to meet any of the following criteria:

1. Must be able to see a monitor and read the adjacent instruction sheet, when seated within a few inches.
2. Must be able to use a mouse and keyboard while resting chin in chin-rest.
3. Must not have ocular condition or vision correction interfering with the calibration of the eye-tracking device.
4. Must not have extensive experience with travel reservation applications.

Based on these criteria, it proved necessary to exclude one subject due to inability to meet criterion 2 and two subjects due to inability to meet criterion 3. No subjects were excluded due to inability to meet criterion 1.

3.3 Physical Facilities

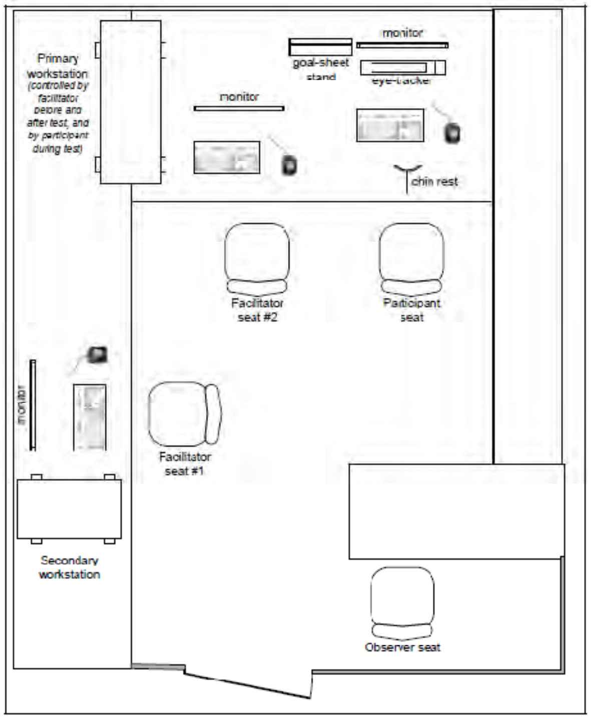


Figure 3 Testing laboratory layout

Figure 3 details the usability laboratory configuration used in both studies. An ideal usability laboratory facility would have the facilitator and subject in separate spaces [12]. In the laboratory used for this research no provision was made to separate the facilitator and the subject due to space limitations. The lack of isolation does not appear to have added a significant level of distraction. With the layout, one additional source of distraction occurred when the facilitator and the subject were of different sexes. In this case, the test protocol required leaving the door partially open; and if requested by either the subject or a facilitator an older female staff members was available to chaperone the session. Using the eye-tracker requires the subject to keep their chin in a fixed position, preventing them from looking around. It appears that is posture also reduces some of the effects of distractions. In the next experiment, facilitators will not share physical space with any subject during test administration to minimize distractions.

Using a Tobii x120 eye-tracker, utilizing Tobii Studio^{TM}, provided eye movement data. A Tobii x120 has the following characteristics: sampling rate - 120Hz, accuracy 0.5°, spatial resolution 0.2°, and drift 0.3°. Subjects viewed the stimulus on a 19-inch monitor located 70cm. from the chin rest, as shown in Figure 3.

3.4 Execution

The current body of literature pertaining to usability testing does not provide an exemplar or template for logging actual use employing an eye-tracker. In designing the test protocol, the general guidelines provided by Nielsen [21] were adhered to as closely as possible. Whenever guidance proved to be lacking, the protocol endeavored to maximize consistency and facilitate future test reproducibility.

The protocol employed is as follows: After directing the subject to sit at the subject's workstation, as shown in Figure 3 the test facilitator places a "Do Not Disturb" sign on the door of the testing facility and closes the door. The door is left slightly ajar if the facilitator and subject are of differing genders. The facilitator then thanks the subject for their participation and asks the subject to review, sign and date a statement of informed consent. After the subject signs and returns the form, the facilitator reminds the subject that they may withdraw from testing at any time, and asks to be notified if the subject wishes to discontinue. The facilitator then assigns a unique code number to the subject. All data for the subject will be associated only with this code number from this point forward.

A subject completes a subject profile, like the one shown in Appendix A.1. After completing the profile, the facilitator requests the subject to remove any hats or non-prescription sunglasses that they are wearing to avoid obstructing the eye-tracking device. Removal of a subject's prescription visual aids such as glasses or contact lenses is not required. If the subject has a cell phone, the facilitator requests that it is turned off or placed in silent/non-vibrating mode, since a ringing or vibrating cell-phone might cause distraction or involuntary eye-movement. After making the request of the subject, the facilitator does likewise. The subject places their chin on the chin rest, and if necessary, adjusts the chin-rest (see Figure

3), height and chair position so that the subject is comfortable and is looking directly at the monitor.

When beginning a session, it is necessary to calibrate the eye-tracking device by having the subject follow an on screen dot only moving their eyes. During this process, the subject may move their eyes and blink normally, but that they should not remove their chin from the chin-rest unless they wish to discontinue the test. These exercises serve two purposes: To calibrate the eye-tracking device, and to determine if the device is able to track the eye movements of the subject.

During the initial calibration, any one of the following conditions occurring would make the subject unqualified for the test.

1 The eye-tracking device will not accurately track a subject’s eyes;

2 The subject cannot see the monitor with good acuity;

3 The subject is unable to use a mouse and keyboard while keeping their chin on the chin-rest.

When disqualified, the subject is thanked and the session is concluded. Otherwise, the facilitator informs the subject that they are now going to continue with the first of a series of tasks.

Subjects are advised that the system is under evaluation and not their skills. The facilitator then explains to the subject the task contains certain requirements, but each task may be completed without precisely fulfilling every requirement. Subject are informed that they will be using an “actual travel website”, but that they should not book any travel, make any purchases, or enter any personal information into the system at any point. Subjects are then informed that the facilitator cannot assist them in any manner and to carry out whatever actions they feel are correct.

Finally, the facilitator tells the subject that they will be completing ten (10) tasks, that there will be periodic breaks, and that the testing session should last for approximately two hours. After offering to review the test instructions (and if necessary re-reading instructions) and confirming that the subject is ready to proceed, the facilitator uncovers the first task, starts the keyboard/mouse input recorder, opens a web browser preset to navigate to the travel system being evaluated, and initiates the eye-tracking recorder.

While the subject carries out each task, the facilitator monitors from either the facilitator's workstation or the observer's station, as illustrated in Figure 3. When the subject arrives at a test termination condition, the facilitator notifies the subject that they may now remove their chin from the chin rest. At this point the facilitator stops the input and eye-tracker recorders, logs the start time, stop time and elapsed time on a post-goal survey, as shown in Appendix A.2, and asks the subject to complete the remaining fields in the survey.

After the subject completes the first task, they are informed that they may take a fifteen-minute break between any of the tasks if they wish. To minimize subject fatigue, the subject takes a fifteen-minute break after tasks four (4) and seven (7). Before each of the remaining nine tasks, it is necessary to recalibrate the eye-tracking device. These subsequent recalibrations are similar to but shorter than the initial calibration.

At the conclusion of the testing, the facilitator once again thanks the subject for their participation, offers to answer any questions that they may have, and ensures that they take a copy of the consent form.

In developing the test administration protocol, a careful balance was struck between keeping subject stress and discomfort levels as low as possible and ensuring that all tests were administered in a consistent and unbiased fashion. A number of dry-runs and walkthroughs were performed in an effort to deal with as many contingencies as possible in advance. Nonetheless, it proved necessary to expand and enhance the testing protocol in response to unanticipated contingencies. Researchers took care to ensure that any expansions or enhancements conformed to the existing testing protocol. However, in one instance, it became necessary to revise, rather than expand upon, one of the goal instructions. A goal, which was feasible using System A, was later discovered to be infeasible using System B. A decision was made to remedy this situation by making the smallest possible change to the goal in order to restore feasibility. The goal’s destination city was changed from Anchorage, Alaska to Spokane, Washington.

The eye-tracker used in the experiments is a high-sensitivity device. It requires recalibration whenever the position of a subject’s head shifts substantially from a fixed position. This necessitates that subjects rest their chin on a chin-rest during testing. Longer testing intervals require that subjects spend more time in an awkward physical position, increasing the probability of physical discomfort and fatigue. On the other hand, shorter testing intervals require frequent recalibrations, increasing the overall length of the testing session.

In light of this dilemma, a decision was made to provide subjects with a short (2-5 minute) break between tasks. Additionally, subjects could request a longer break (5-15 minutes) between any of the tasks.

To keep test conditions as objective as possible, facilitators did not interact with subjects in any fashion while a task was under way. When encountering an issue or condition not covered in the testing protocol, facilitator's were authorized to resolve those issues at their discretion and then to document the condition and action taken. On these rare occasions, one or more researchers reviewed actions of the facilitator to determine if the data collected was usable or if the condition represented a usability issue.

At the start of each testing session, subjects were advised that they would not receive any information or guidance during a test task. They were further informed, both verbally and in writing, that neither their aptitude nor their abilities were being evaluated, and that if they got "stuck" at any point, they should simply proceed as they deemed best. If asked a question during a task, facilitators were directed to either restate directions contained in the facilitator instruction sheet, or to reply: "I apologize, but I cannot help you." After a number of instances of subjects giving indications of discouragement, a decision was made to inform subjects, if necessary, that they were "doing fine" and to remind them that it is the software under evaluation and not them.

Before the commencement of subject testing, a number of possible contingencies were foreseen and appropriate facilitator responses determined. However, some unforeseen circumstances did arise which required the facilitator to determine an appropriate response "on-the-fly." Facilitators were told to be as friendly and helpful as possible, but also to be extremely careful not to take action or provide information, which might bias results.

Certain subjects asked to know why a task was being terminated or under what circumstances a task would be terminated. The decision was made that this query should be replied to as follows: "I am monitoring the tasks and will end a session under certain circumstances." A terse response of, "I'm sorry, I can't tell you," was deemed to be too impolite, potentially causing bias by negatively motivating the subject. However, a full disclosure of termination conditions was decided against because it would provide subjects with “inside information” which might bias subject performance.

Subjects frequently, in various ways, attempted to determine whether or not they had successfully completed a goal or whether or not they had carried out a task “correctly”. Facilitators accordingly had to come up with a number of polite variations on the response, “There are no ‘incorrect’ actions; it is up to you to determine how to carry out a task and to decide whether or not you have successfully completed a goal.”

A few subjects asked for clarification as to the meaning of certain items in the “Vision” section of the subject profile. Since the information requested was of a general factual nature and not specifically germane to the experiment, facilitators would provide definitions of terms such as “near-sighted,” “far-sighted” or “prisms.”

During the design of the test protocol and the dry-run sessions, a set of polite and curious responses developed for facilitators. However, it is not possible to predict everything that can happen during a usability test. For example, one ambidextrous subject wanted to change hands during the experiment. Moving the mouse from the right-hand to the left and vice-versa probably would not affect the data. Since the facilitator had no instructions on this situation, they correctly requested that the subject use the same hand for all of the tasks.

In a number of instances, subjects were observed disregarding certain task constraints. For example, one of the sub-goals of the tasks was to book a hotel room within a certain geographical distance from another hotel. This sub-goal was frequently ignored because neither System A nor System B contained a feature which could directly provide this information. It was necessary for subjects to infer the distance between hotels based on each hotel’s distance from the destination city. Data from these subjects was not excluded from the study because this type of behavior may indicate a possible software usability issue.

Although the organizations offering these travel reservation systems would disagree, several subjects complained about distraction from “banner ads” and other content extraneous to the functionality of the travel system. Even though the browser used in the study had “pop-up” windows disabled, there were still a significant number of advertisements presented to the subject. This may be a situation where the additional revenue generated by these distractions may outweigh the impact on usability.

Because facilitators could not prompt or question subjects, it was necessary to carefully determine in advance the conditions whereby a task would be considered to be at its conclusion point. Facilitators were to terminate a test task if any of the following events occurred:

1. The subject arrives at a “log in to continue” screen. Subjects were not to provide any personal information, which required terminating the session.
2. The system becomes non-responsive for a period of two minutes.
3. The subject is unable to make any progress toward goal completion for two minutes.
4. The subject states that they are finished with the task.
5. Ten minutes have elapsed.

3.5 Summary

After collecting data from 20 subjects, ten (10) on each system, the researchers and facilitators met to discuss the experiment. In general, the facilitators felt that the experiment went smoothly despite resource limitations and facility constraints. However, a number of improvable shortcomings did become apparent during the research review.

In the planning process, the goal template was verified on both System A and System B; but when the specific tasks were developed, not every task was verified using System B. As stated in the previous section, one of the goals was not achievable using System B because that system did not provide air travel to the specified destination. A simple solution to this problem would be to check every goal on every system under evaluation.

Since the investigating team was small, five (5) individuals, a less formal plan proved adequate. However, a larger experiment with more principles, i.e. researchers, facilitators, and observers, requires a formal plan to ensure effective communication among the participants. A formal plan would also be necessary when conducting a formal validation, like those conducted for medical device and avionics.

Facilitators monitored subject interaction with the travel systems to ensure that no personal information was entered at any time and to check for task termination points. In future studies, it would be beneficial to modify the front-end of the system to proscribe the input of personal information and automatically terminate tasks when achieving all of the tasks goals.

After reviewing the session durations and some informal discussion with subjects, reviewers felt that the tasks required too much time to complete and the average session duration of 2.5 hours was excessive. Unless the subjects are paid, the time spent on each session may be excessive and make it difficult to acquire subjects. Although eye tracking adds much information about the subject during a usability test, it has a relatively high cost in both time and equipment.

3.6 Recommendations

After discussing the strengths and weaknesses of the experiment, the research team elected to take the following actions to improve the usability testing practice:

1. Develop a format for a formal plan for future usability tests.
2. Improve the usability laboratory and its equipment.
3. Improve goal design parameters and techniques.

Acquiring additional specialized equipment could reduce subject fatigue and make the experience more enjoyable for the subject. It may be possible to improve the testing process by adapting existing techniques such as IEEE Test Documentation Procedures [5].

One challenge of an effort-based usability test is data reduction. Data reduction, viewed by many experts as the primary limitation of an actual data logging approach [21], requires a great deal of investigation. It is believed that a major contribution of the new effort-based metric research will resolve this limitation.

3.6.1 Planning

Adapting the IEEE Test Documentation [5] requires making changes to some document outlines, and in some cases, adjusting the terms generally used in software testing to those used in usability testing. Many usability researchers have a psychology or sociology background but limited experience with software engineering. This may partially explain why usability testing terminology differs from traditional testing vocabulary.

Before addressing the additional outline items required for a usability test plan, it may be best to resolve some of the terminology differences between software testing and usability test practice. Three potentially confusing terms are “goal”, “task” and “protocol”. “Goal” and "task" are used interchangeably to mean work performed by a subject, using a specific set of software, in order to provide an observation. This definition is not significantly different from that of a test case. The IEEE Software Engineering Glossary [4] generally defines a test case as containing inputs, operating conditions, and expectations. Although the IEEE definition is much more precise, a test case is nonetheless a task being performed using a specific set of software for making an observation. In the case of a test case, the observation is whether actual results match expected results.

When software engineers use the term "protocol", they think about the rules necessary for two processes to interact [4], e.g. a communications protocol. In social sciences such as psychology, a protocol is conceptualized as a strictly followed detailed-procedure [8]. Software engineers simple use the term "procedure" to identify the instructions for a manual process. In software testing, test designers assume that testers strictly follow all written instructions.

Of the eight (8) documents described in the IEEE Test documentation standard [5], four (4) are used to document the test planning and test design. These four documents are the test plan, test design specification, test case specification, and test procedure specification. Only a few changes are necessary to make these documents consistent with the terminology and practice employed in a usability study.

The recommended outline for a test plan does not provide a section describing or referencing policies for human-subject experimentation. There are three (3) options for describing the human use policies:

1. Add a new main section.

2. Add a subsection to the Environment section.
3. Add a subsection to the Staffing section.

It is possible to support arguments for adding the necessary information to any of these sections. The only issue that is important is that the usability test plan describes the measures used to ensure the safety and anonymity of the subject. Safety and anonymity are the primary concerns for this section, but test designers might also want to document the procedures used in recruiting and compensation of subjects. These two (2) issues may have some bearing on the veracity of the test data.

Planners should insure that the test plan requires test designers to test every goal/task/test case on every system under evaluation. This will prevent changes to goals/tasks/test case during the testing process.

One of the more successful parts of this experimental test was the use of a goal template or test scenario. Approaching goal design using scenario based test case techniques greatly simplified goal creation and insured that there was a consistency among all of the goals or test cases. Scenario based test case design is interchangeable with use-case test design techniques. The template provides the basic characteristics of the use-case and the goal or task is just different input data. Probably the best place to put this information is in the Test Case Specification document, which should be renamed the Goal Specification Document.

Other than changing the name of the Test Procedure Specification to Test Protocol Specification there are no additional changes that should be necessary. Future experimentation will tell if more changes need to be made to the test documents.

3.6.2 Physical Facility

For the current studies, task instructions were placed on a stand directly to the left of, and in the same horizontal plane as, the system display. It is believed that shifting the location of the instruction sheet to a position directly above and in the same vertical plane as the display will enhance eye-tracker accuracy as well as reduce visual and physical fatigue in subjects.

Use of an adjustable desk should increase a subject's comfort and provide additional options for test setup configuration. Additionally, some sort of mounting scheme should be employed to ensure that the chin rest, eye-tracker and system display remain firmly fixed in place. A shift in any one of these components after eye-tracker calibration has been executed can cause invalid data to be recorded.

Even though researchers and facilitators felt the facilities were adequate for this type of experiment, a dedicated usability test facility would eliminate some of the challenges in conducting usability tests. Further research related to this effort-based measure of usability will require a large number of tests, and an experiment where both cognitive and objective methods are compared has been planned. A usability facility like the ones described by Joseph Dumas [12] will probably be required going forward.

3.6.3 Execution

The test design specified that each task must be terminated after ten minutes, regardless of subject progress towards goal achievement. It would have been infeasible and undesirable to allow an indefinite amount of time for completing each task. A "hard" limit of ten minutes attempted to strike a balance between maximizing the generation of test data while minimizing the possibility of physical discomfort or emotional frustration. It is now believed that either ten minutes is too much time for a specific task and ten tasks maybe excessive for a session. The overall length of the testing sessions is not believed to have compromised test data, but it is perceived that sessions lasting upwards of ninety minutes are overly tedious for both test facilitator and subject. In future research, the duration and number of tasks will need to be reduced accordingly.

3.7 Data Reduction

3.7.1 Manual Data

An event driven logging program obtains details of mouse and keystrokes activities from the operating system event queue. The program saves each event along with a time stamp into a file. The logged events are: Mickeys, keystrokes, mouse button clicks, mouse wheel rolling, and mouse wheel clicks. In the reported experiments, the program has generated about 60,000 time stamped event per task (about 10 minutes).

A data reduction program applied to the event data, counts the total number of events (e.g., Mickeys) per task. Based on the data logged, the right mouse button was not used. Only three subjects occasionally used the mouse wheel, and therefore, mouse wheel events were excluded. With 20 subjects, each completing 10 tasks; the data reduction program generated 200 data points. The data obtained from the data reduction stage is averaged per task per travel reservation system. Hence, a set of 20 points is generated where each point denotes the average count of events per task per reservation system.

3.7.2 Eye Movement Data

The saccade and fixation detection algorithm is based on the Velocity-Threshold Identification (I-VT) model [22], and measured in degrees. Hence, the velocity-based classification approach has more synergy with oculomotor mechanics than dwell time [22]. An eye position sample belongs to a saccade if the calculated eye velocity exceeds 30°/s. It is fixation, if the eye velocity falls below this threshold. This threshold value (30°/s) is suggested by the oculomotor research literature [18]. Saccades with parameters such as onset, offset, amplitude, and duration are detected by merging a sequence of continuous eye position samples with calculated velocities exceeding 30°/s. Saccade amplitude, measured by the Euclidean distance between the first and last point in the sequence, is represented by degrees of visual angle. Saccades with amplitudes of less than 2° and saccades where the eye tracker failed to detect an eye position even for a single sample were discarded from the analysis. A fixation is defined as a sequence of eye position samples with velocity of less than 30°/s and not more than 2° apart from each other. Sequences of eye tracking failures or blinks less than 75ms were considered as a part of fixation duration. Minimum fixation duration is 100ms. Fixation location is defined as the average of all the valid eye position coordinates excluding micro saccades (saccades with amplitude of less than 1°). Eye fixation duration is defined as a time difference between the first and last samples in the fixation sequence.

4 RESULTS

4.1 Observed Results

To paraphrase Glenford Myers [20], a good evaluation is one that finds issues. Using this as a measure of quality of the evaluation, one could say that this evaluation was very successful.

In a number of instances, subjects ignored certain task constraints. For example, one of the sub-goals of the tasks was to book a hotel room within a certain geographical distance from another hotel. Subjects ignored this constraint because neither System A nor System B contained a feature, which could directly provide this information. It was necessary for subjects to infer the distance between hotels based on each hotel's distance from the destination city. Although the organizations offering these travel reservation systems would disagree, several subjects complained about distraction from "banner ads" and other content extraneous to the functionality of the travel system. Even though the browser used in the study had "pop-up" windows disabled, there were still a significant number of advertisements presented to the subject. This may be a situation where the additional revenue generated by these distractions may outweigh the impact on usability.

Figure 4 contains the average task completion time for each task with each system. The average experiment length for "System B" follows the hypothesized learning curve presented in Figure 2. "System B" has a jittered trend, yet it follows a similar slope. In addition, the task completion time for "System A" is more than twice longer than the completion times for "System B". The standard, deviation values computed for "System A" are higher than the standard deviation values of "System B".

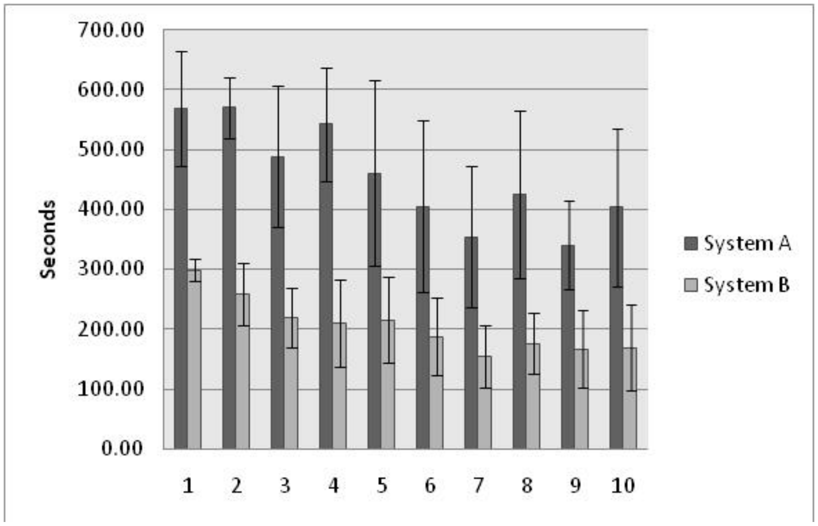


Figure 4 Average Experiment Length

4.2 Manual Effort Results

A set of figures generated from these points is used to evaluate the data, compare the usability of the two systems, and assess the correlation between the obtained data and the research hypothesis. Figure 7 contains the average number of Mickeys for each task with each reservation system. In addition, it includes the number of Mickeys expended by each of the two facilitators on each task with each of the two reservation systems. It can be observed that System B requires less activity than System A. A spike in activity with respect to task 3 in system A can be used as an example of the capability of the metrics to pinpoint potential interface shortfalls. In a similar way, Figures 5, 6, 7, and 8 contain the average number of keystrokes, left mouse clicks, and transitions for each tasks and reservation system. The results depicted in Figures 5 to 8 are highly correlated and show that more manual effort is expended in system A. It is evident that system B is more operable than system A and that the results are in agreement with the hypothesis that usability is related to effort. Additional research is required to refine the model and verify the agreement of results with the learnability part of the hypothesis.

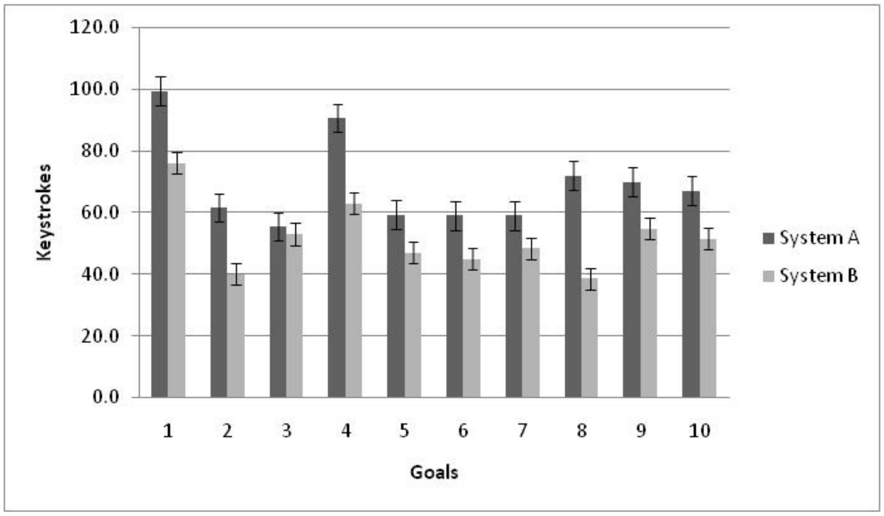


Figure 5 Average Keystrokes

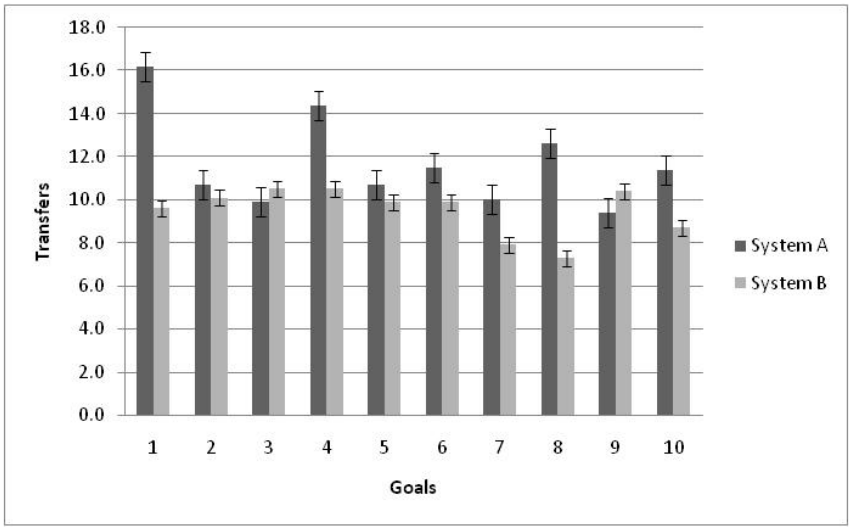


Figure 6 Average Transitions

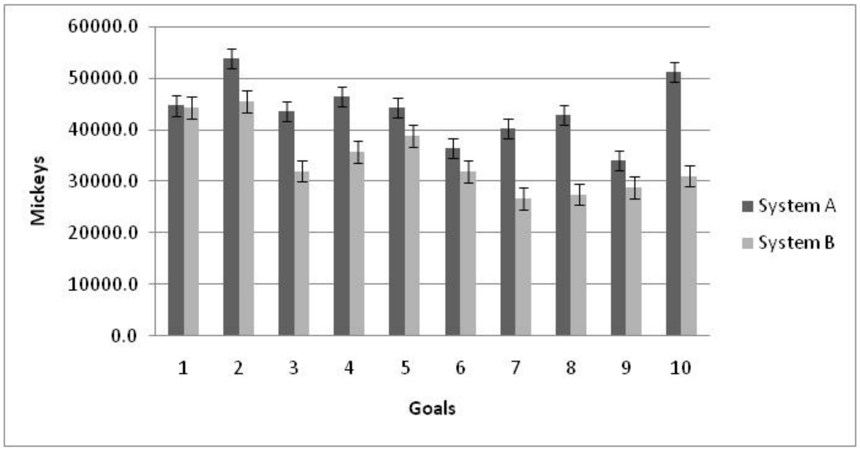


Figure 7 Average Mickeys

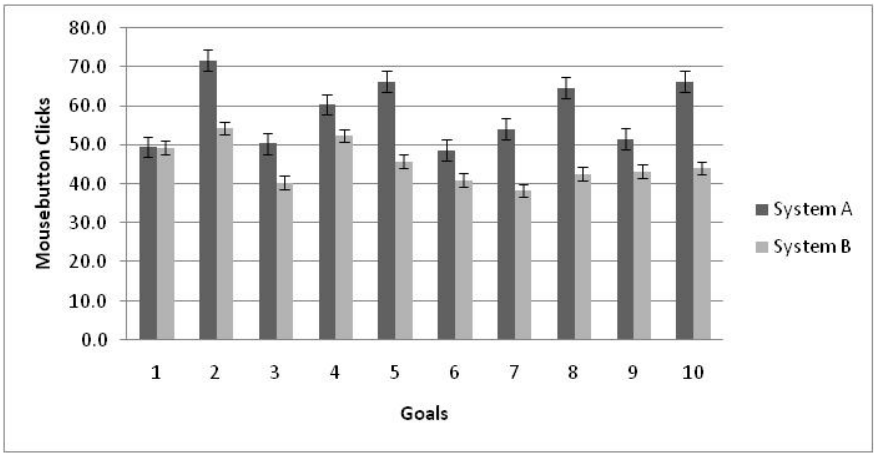


Figure 8 Average Left Mouse Button Presses

4.3 Eye Effort Results

Figure 9, depicts the average length of the path traversed by the eyes. The metric achieved its local minimum at the seventh trial and the length of the path traversed increased thereafter. This may be attributed to fatigue. Eye effort results are highly correlated with the completion time results. They support the learning curve hypothesis and show evidence of learning with respect to system B. In the case of “System A” the learnabilitty trend is less clear with average eye path traversed showing large variations. Figure 9, depicts average length of the path traversed by the eyes. The metric achieved its local minimum at the seventh trial and the length of the path traversed increased thereafter. This may be attributed to fatigue. Figure 10 illustrates the results for the average fixation count. This metric almost mirrored the results of average task completion time (see Figure 4) and the average path traversed by the eyes (see figure). Compared to “System B”, “System A” had substantially higher number of fixations for all tasks; indicating that “System B” presents the information more efficiently and provides better opportunities for organized search. According to Figure 11, a higher number of saccades is recorded for System A, unlike other measurements, higher results indicate less effort. These results indicate that more searching is done when using System B; therefore the may be presented less efficiently than in System A.

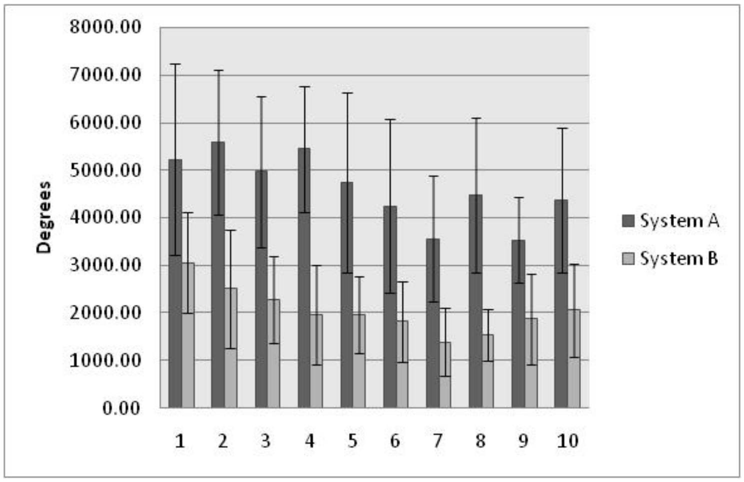


Figure 9 Average Eye Path Traversed

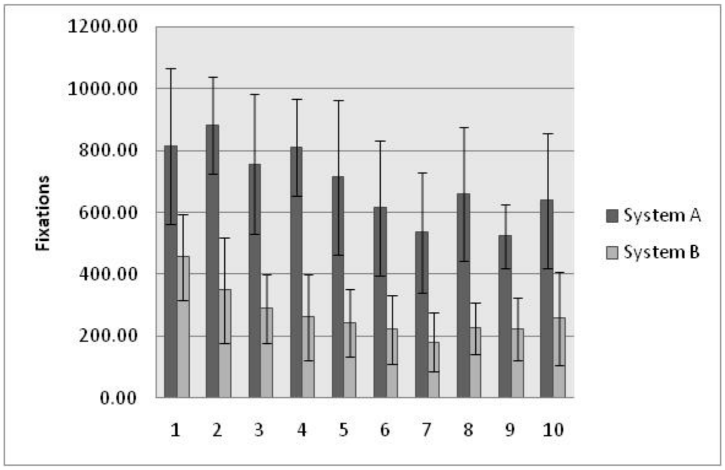


Figure 10 Average Fixation Count

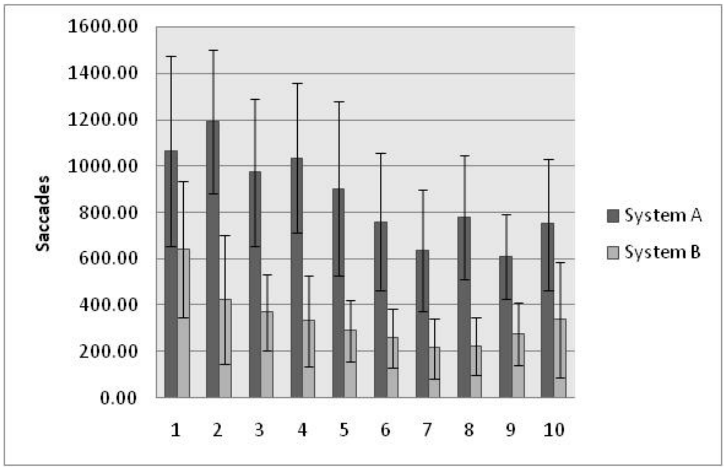


Figure 11 Average Saccade Count

4.4 Mental Effort Results

Figures 12 illustrates the average fixation duration. For System B, the average fixation duration does not change much, and is approximately 250ms. This indicates that despite the fact that the average-task completion-time decreased, the amount of attention allocated to each task is about the same. In the case of System A, higher cognitive load of more than 400ms was recorded. Again, this cognitive load did not change substantially from task to task possibly supporting the assumption that System A has higher cognitive load. The results obtained for System B are significantly lower than System A's results. This may indicate that System B requires less cognitive load, and may be indicative of subject's level of confusion.

Figure 13 shows that in terms of the pupil dilations metric, System A requires higher mental effort than System B. Figure 14 illustrates the average saccades amplitude results. Saccades amplitude is inversely proportional to mental effort. System B provides more meaningful cues to the user than System A therefore requires less mental effort. A slight learning curve for System A, depicted by an upward trend in the average saccade amplitude, which grows from 4° to 4.6°, can be observed. In the case of “System B”, the average saccade amplitude remained close to 5° without a specific trend.

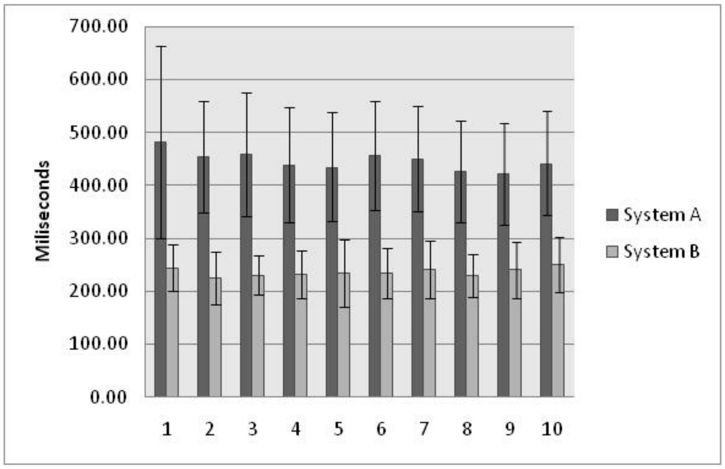


Figure 12 Average Fixation Duration

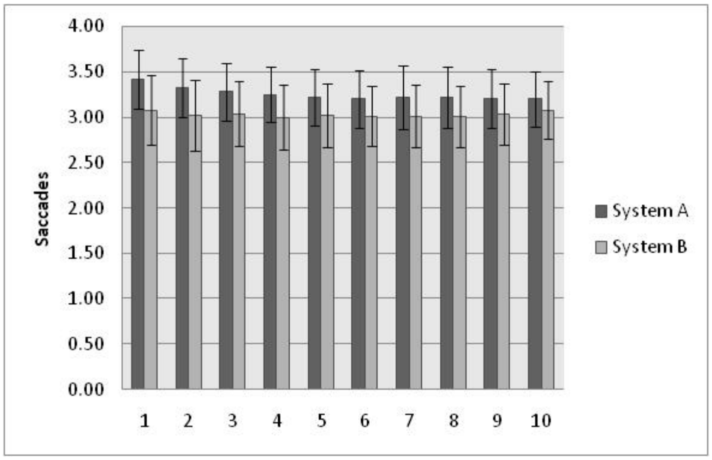


Figure 13 Average Pupil Dialation

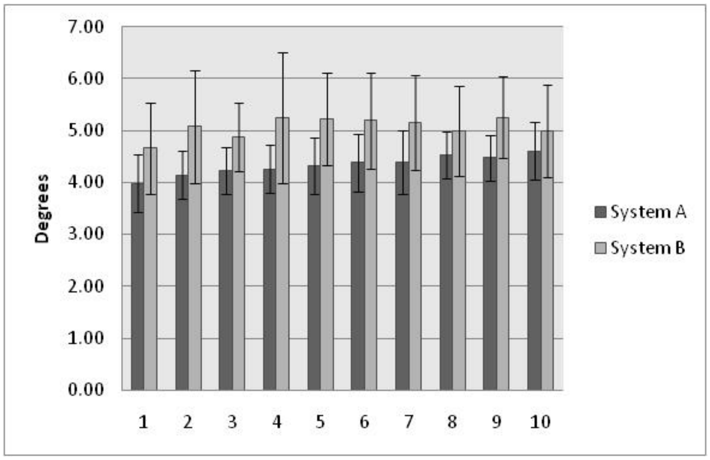


Figure 14 Average Saccades Amplitude

4.5 Findings

The research results support the following important observations:

1. The research illustrates that logging of interaction events such as Mickeys, mouse button clicks, and keystrokes, along with eye tracking, provides a great deal of useful information. There is a clear correlation between the effort approximation model presented and usability. This correlation can be exploited and used to evaluate the usability of existing user interfaces as well as user interfaces that are at a relatively advanced stage of design. Nevertheless, further experimentation is required in order to refine the hypothesis, the model, and the effort approximation procedures that are guiding the research.
2. The research shows that logging and processing interaction events is feasible and useful. Previously it was stipulated that the volume of data obtained through logging interaction events is un-reducible and therefore useless. This research indicates, however, that this may no longer be a problem.
3. An important contribution of this research is that it can enable pinpointing GUI design and implementation defects and shortfalls. For example, consider a goal, which yields an

excessive amount of effort in a given test (or set of tests). The reported method enables identifying the specific lines of code in the GUI implementation that generate the excessive effort and evaluating possible modifications to remedy the shortfall.

4. Careful design of the tests performed enables obtaining high quality results despite working with very limited resources and no funding.
5. One of the important aspects of the usability testing strategy is the utilization of a use-case scenario based test design technique. This technique is instrumental in facilitating the usage of appropriate goals and test procedures. Moreover, it is an important component of the ability of the proposed effort based metrics to pinpoint design and implementation shortfalls.
6. While the manual based interaction activities (mouse, keyboard, gloves, etc.) are strictly related to physical effort, the eye movement data is related to both physical and mental effort. It can be utilized for enhancing the physical effort model. Currently, it is the only type of data correlating with mental effort. Hence, the research opens the door for a layered approach to GUI usability testing. At the lower layer, only manual data is recorded and used for fast and relatively inexpensive usability evaluation. At the next layer, eye tracking devices provide means for mental effort evaluation and refinement of the physical effort approximation techniques. A potential future research relates to the utilization of brain wave measurements to further enhance the mental effort evaluation procedures.

5 CONCLUSIONS, AND FUTURE RESEARCH

This paper presents significant results of a current and ongoing research effort, which is gaining a noteworthy momentum. The almost intuitive idea that usability corresponds to effort and effort can be approximated through objective metrics such as the number of Mickeys, mouse clicks, and keyboard key clicks was considered about fifteen years ago. It was abandoned, however, due to a sentiment that the amount of data obtained through logging of these events (Mickeys etc.) is overwhelming [21]. This research shows that obtaining and processing effort data is possible, meaningful, and useful. Moreover, additional data obtained through eye movement metrics can be collected and processed using reasonable computation and logging resources. The eye movement data contributes to the physical effort approximation and is paramount for mental effort approximation.

The results do not completely match the hypothesis laid down at the beginning of the research. The main mismatch is in the area of learnability. Nevertheless, the results provide significant evidence that operability, learnability, and understandability can be approximated using interactions event logging and utilized to usability evaluation.

This research, which is in the beginning stages of development, can become a breakthrough in the approach toward usability evaluation. Nevertheless, it is important to stress that the new approach does not exclude the current cognitive based usability evaluation techniques and is intended to complement them.

Usability is a huge and important area of research and development and one paper or research effort cannot cover the multitude of relevant issues. Several of these issues, which will be addressed in future research, include:

1. Further investigation into scenario-based test design techniques appears warranted, based on the results from the initial experiment. With additional test cases and an improved test case design technique, it may be possible to shed more light on the usability model and

its utility, and reduce unknowns such as the influence of fatigue. Furthermore, additional research is in progress to devise a set of experiments to assess the usability of individual GUI widgets and their combinations.

2. This paper treats every metric individually. Considering the metrics individually, it is evident that the initial hypothesis of the research is partially established. The initial hypothesis of the research, however, includes an assumption that it is possible to derive a procedure to combine the individual metrics into a single approximation for the effort $E(t)$ and correlate this approximation with traditional measures of operability, learnability and understandability. Further research is required to determine whether it is possible to reduce the individual metrics into one measure that approximate usability. In this respect, additional experiments can provide data to assess this question and identify the appropriate weight values (w_i), in equations that are associated with individual metrics.
3. Additional research on metrics for mental effort evaluation, specifically brain wave related metrics, can be an interesting extension of this research.
4. Another direction of future research is to consider a dynamic scenario where the system adapts to the user and enables user specific improvements in usability at run time.

APPENDIX A GOALS OR TASKS

A.1 TEMPLATE

A.1.1 GOAL

Dr./Ms./Mr. ____________ is presenting a paper at the ______________ conference being held in _______________ at the _________________. He/she is presenting his/her paper at 10A.M., but he/she must be there for the opening session at 8:30 A.M. The conference will end at 6P.M. on ____________ and Dr./Ms./Mr. _____________ must be there for the closing session.

Dr./Ms./Mr. ________________ is traveling from ______________, and would like a non-stop flight to ___________________.

The conference is at the __________________ hotel on ___________ to ____________, but Dr./Ms./Mr. __________ feels that this hotel is outside of the range of his/her budget of ________ for the travel. Because of the high cost of the hotel he/she wants to stay at a hotel within ____________ miles of the conference center with the following amenities:

1. ____________________________
2. ____________________________
3. ____________________________
4. ____________________________

He/she will need a car to get around at conference city. Again, because of budget constraints, he/she does not want to spend more than ________/day for the car.

A.1.2 DIRECTIONS

Using the web browser already opened, make a flight, hotel, and car rental reservation for Dr. Waterford based on the below information. You should make every attempt to comply with the budget, distance, amenities, and travel time constraints given. Both the departure and return flights *must* be non-stop. Ensure that the airline and hotel reservation is for one adult only. Do not open additional browser windows/tabs, and do not navigate away from System A/System B. You may, however, click on any links provided by System A/System B if they are necessary for, or related to your search.

A.2 GOALS

A.2.1 GOAL 1

Dr. Vornoff is presenting a paper at the *Pikes Peak* conference being held at the Broadmoor hotel in Colorado Springs, Colorado. He is presenting his paper at 10:00 am on Thursday, October 16, but he must be present for the opening session at 8:00 am on Wednesday, October 15 and remain for the duration of the conference, which ends at 3:00 pm on Friday, October 17. He has a travel budget of $800.

Dr. Vornoff is traveling from Salt Lake City, Utah and insists on a non-stop flight to Colorado Springs. Since he feels that the Broadmoor is out of his price range, Dr. Vornoff would like a room at a less-expensive hotel within 10 miles from the conference. This hotel should have the following amenities:

1. Exercise room

2. Internet (wireless or wired)

3. Restaurant/dining room

Dr. Vornoff will need to rent a car during his stay in Colorado Springs. He does not want to spend more than $50 per day, or $180 total for the car rental.

A.2.2 GOAL 2

Dr. Jones is presenting a paper at the *Yellow Brick Road* conference being held at the Hyatt Regency hotel in Wichita, Kansas. She is presenting her paper at 10:00 am on Thursday, October 30, but she must be present for the opening session at 9:00 am on Tuesday, October 28 and remain for the duration of the conference, which ends at 3:00 pm on Friday, October 31. She has a travel budget of $900.

Dr. Jones is traveling from Houston, Texas and insists on a non-stop flight to Wichita. Since she feels that the Hyatt Regency is out of her price range, Dr. Jones would like a room at a less-expensive hotel within 8 miles from the conference. This hotel should have the following amenities:

1. Restaurant/dining room
2. Internet (either wired or wireless)
3. Exercise room

Dr. Jones will need to rent a car during her stay in Wichita. She does not want to spend more than $50 per day, or $250 total for the car rental.

A.2.3 GOAL 3

Mr. Smith is presenting a paper at the *Big Metal Arch* conference being held at the Omni Majestic hotel in St. Louis, Missouri. He is presenting his paper at 10:00 am on Tuesday, October 21, but he must be present for the opening session at 8:00 am on Monday, October 20 and remain for the duration of the conference, which ends at 4:00 pm on Friday, October 24. He has a travel budget of $1400.

Mr. Smith is traveling from San Antonio, Texas and insists on a non-stop flight to St. Louis. Since he feels that the Omni Majestic is out of his price range, Mr. Smith would like a room at a less-expensive hotel within 10 miles from the conference. This hotel should have the following amenities:

1. Restaurant/dining room

2. TV with premium cable channels

3. Exercise room

Mr. Smith will need to rent a car during his stay in St. Louis. He does not want to spend more than $70 per day, or $350 total for the car rental.

A.2.4 GOAL 4

Dr. Waterford is presenting a paper at the *Paul Bunyan* conference being held at the Minneapolis Grand hotel in Minneapolis, Minnesota. He is presenting his paper at 11:00 am on Wednesday, October 15, but he must be present for the opening session at 9:00 am on Tuesday, October 14 and remain for the duration of the conference, which ends at 4:00 pm on Friday, October 17. He has a travel budget of $1000.

Dr. Waterford is traveling from Albuquerque, New Mexico and insists on a non-stop flight to Minneapolis. Since he feels that the Minneapolis Grand is out of his price range, Dr. Waterford would like a room at a less-expensive hotel within 10 miles from the conference. This hotel should have the following amenities:

1. Wireless Internet

2. Restaurant/dining room

Dr. Waterford will need to rent a car during his stay in Minneapolis. He does not want to spend more than $70 per day, or $250 total for the car rental.

A.2.5 GOAL 5

Ms. O'Hara is presenting a paper at the *Tara and Twelve Oaks* conference being held at the Marriott Marquis hotel in Atlanta, Georgia. She is presenting her paper at 3:00 pm on Thursday, September 25, but she must be present for the opening session at 9:00 am on Wednesday, September 24 and remain for the duration of the conference, which ends at 4:00 pm on Friday, September 26. She has a travel budget of $1000.

Ms. O'Hara is traveling from Shreveport, Louisiana and insists on a non-stop flight to Atlanta. Since she feels that the Marriott Marquis is out of her price range, Ms. O'Hara would like a room at a less-expensive hotel within 6 miles from the conference. This hotel should have the following amenities:

1. Exercise room
2. Room service
3. Internet (wired or wireless)

Ms. O'Hara will need to rent a car during her stay in Atlanta. She does not want to spend more than $75 per day, or $300 total for the car rental.

A.2.6 GOAL 6

Dr. Frank-N-Furter is presenting a paper at the *Time Warp* conference being held at the Westin Tabor Center hotel in Denver, Colorado. He is presenting his paper at 2:00

pm on Tuesday, October 7, but he must be present for the opening session at 8:00 am on Monday, October 6 and remain for the duration of the conference, which ends at 3:00 pm on Friday, October 10. He has a travel budget of $1200.

Dr. Frank-N-Furter is traveling from Columbus, Ohio and insists on a non-stop flight to Denver. Since he feels that the Westin Tabor Center is out of his price range, Dr. Frank-N-Furter would like a room at a less-expensive hotel within 12 miles from the conference. This hotel should have the following amenities:

1. Exercise room
2. Internet (wired or wireless)
3. Restaurant/dining room
4. TV with premium channels

Dr. Frank-N-Furter will need to rent a car during his stay in Denver. He does not want to spend more than $75 per day, or $350 total for the car rental

A.2.7 GOAL 7

Mr. Petty is presenting a paper at the *Stock Car Racing* conference being held at the Dunhill hotel in Charlotte, North Carolina. He is presenting his paper at 1:00 pm on Tuesday, September 23, but he must be present for the opening session at 9:00 am on Tuesday, September 23 and remain for the duration of the conference, which ends at 5:00 pm on Friday, September 26. He has a travel budget of $1000.

Mr. Petty is traveling from Detroit, Michigan and insists on a non-stop flight to Charlotte. Since he feels that the Dunhill is out of his price range, Mr. Petty would like a room at a less-expensive hotel within 12 miles from the conference. This hotel should have the following amenities:

1. Wireless Internet
2. Restaurant/dining room

Mr. Petty will need to rent a car during his stay in Charlotte. He does not want to spend more than $65 per day, or $320 total for the car rental.

A.2.8 GOAL 8

Mr. Buffett is presenting a paper at the *Reuben Sandwich* conference being held at the Hilton Garden Inn hotel in Omaha, Nebraska. He is presenting his paper at 11:00 am on Wednesday, October 22, but he must be present for the opening session at 8:00 am on Monday, October 20 and remain for the duration of the conference, which ends at 4:00 pm on Friday, October 24. He has a travel budget of $1200.

Mr. Buffett is traveling from Chicago, Illinois and insists on a non-stop flight to Omaha. Since he feels that the Hilton Garden Inn is out of his price range, Mr. Buffett would like a room at a less-expensive hotel within 8 miles from the conference. This hotel should have the following amenities:

1. Room service

2. Exercise room
3. Internet (wired or wireless)

Mr. Buffett will need to rent a car during his stay in Omaha. He does not want to spend more than $55 per day, or $325 total for the car rental.

A.2.9 GOAL 9

A.2.9.1 GOAL 9A

Ms. Kilcher is presenting a paper at the *Who Will Save Your Soul* conference being held at the Captain Cook hotel in Anchorage, Alaska. She is presenting her paper at 9:00 am on Friday, October 31, but she must be present for the opening session at 8:00 am on Tuesday, October 28 and remain for the duration of the conference, which ends at 3:00 pm on Friday, October 31. She has a travel budget of $2400.

Ms. Kilcher is traveling from Salt Lake City, Utah and insists on a non-stop flight to Anchorage. Since she feels that the Captain Cook is out of her price range, Ms. Kilcher would like a room at a less-expensive hotel within 10 miles from the conference. This hotel should have the following amenities:

1. Restaurant/dining room
2. Exercise room
3. Wireless Internet

Ms. Kilcher will need to rent a car during her stay in Anchorage. She does not want to spend more than $80 per day, or $380 total for the car rental.

A.2.9.2 GOAL 9B

Ms. Kilcher is presenting a paper at the *Who Will Save Your Soul* conference being held at the Captain Cook hotel in Spokane, Washington. She is presenting her paper at 9:00 am on Friday, October 31, but she must be present for the opening session at 8:00 am on Tuesday, October 28 and remain for the duration of the conference, which ends at 3:00 pm on Friday, October 31. She has a travel budget of $2400.

Ms. Kilcher is traveling from Salt Lake City, Utah and insists on a non-stop flight to Spokane. Since she feels that the Davenport is out of her price range, Ms. Kilcher would like a room at a less-expensive hotel within 8 miles from the conference. This hotel should have the following amenities:

1. Restaurant/dining room
2. Exercise room
3. Wireless Internet

Ms. Kilcher will need to rent a car during her stay in Spokane. She does not want to spend more than $80 per day, or $380 total for the car rental.

A.2.10 GOAL 10

Dr. Van Zant is presenting a paper at the *Lynyrd Skynyrd* conference being held at the Omni Jacksonville hotel in Jacksonville, Florida. He is presenting his paper at 11:00 am on Thursday, October 9, but he must be present for the opening session at 9:00 am on Tuesday, October 7 and remain for the duration of the conference, which ends at 2:00 pm on Friday, October 10. He has a travel budget of $1000.

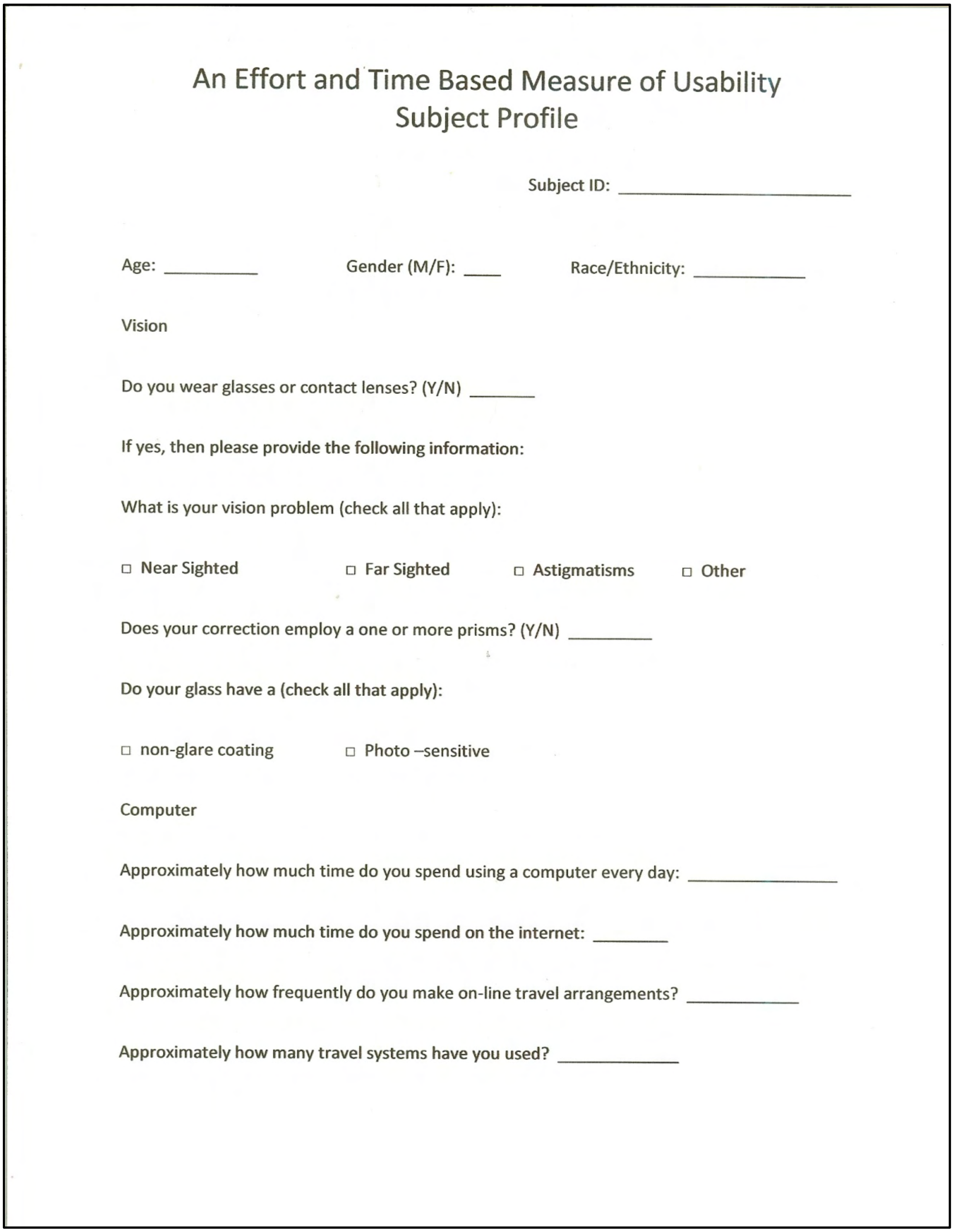
Dr. Van Zant is traveling from Boston, Massachusetts and insists on a non-stop flight to Jacksonville. Since he feels that the Omni Jacksonville is out of his price range, Dr. Van Zant would like a room at a less-expensive hotel within 10 miles from the conference. This hotel should have the following amenities:

1. Internet (wireless or wired)
2. Restaurant/dining room

Dr. Van Zant will need to rent a car during his stay in Jacksonville. He does not want to spend more than $50 per day, or $220 total for the car rental.

APPENDIX B FORMS

B.1 SUBJECT PROFILE



An Effort and Time Based Measure of Usability
Subject Profile

Subject ID: ________________________

Age: __________ Gender (M/F): ____ Race/Ethnicity: ____________

Vision

Do you wear glasses or contact lenses? (Y/N) _______

If yes, then please provide the following information:

What is your vision problem (check all that apply):

□ Near Sighted □ Far Sighted □ Astigmatisms □ Other

Does your correction employ a one or more prisms? (Y/N) _________

Do your glass have a (check all that apply):

□ non-glare coating □ Photo –sensitive

Computer

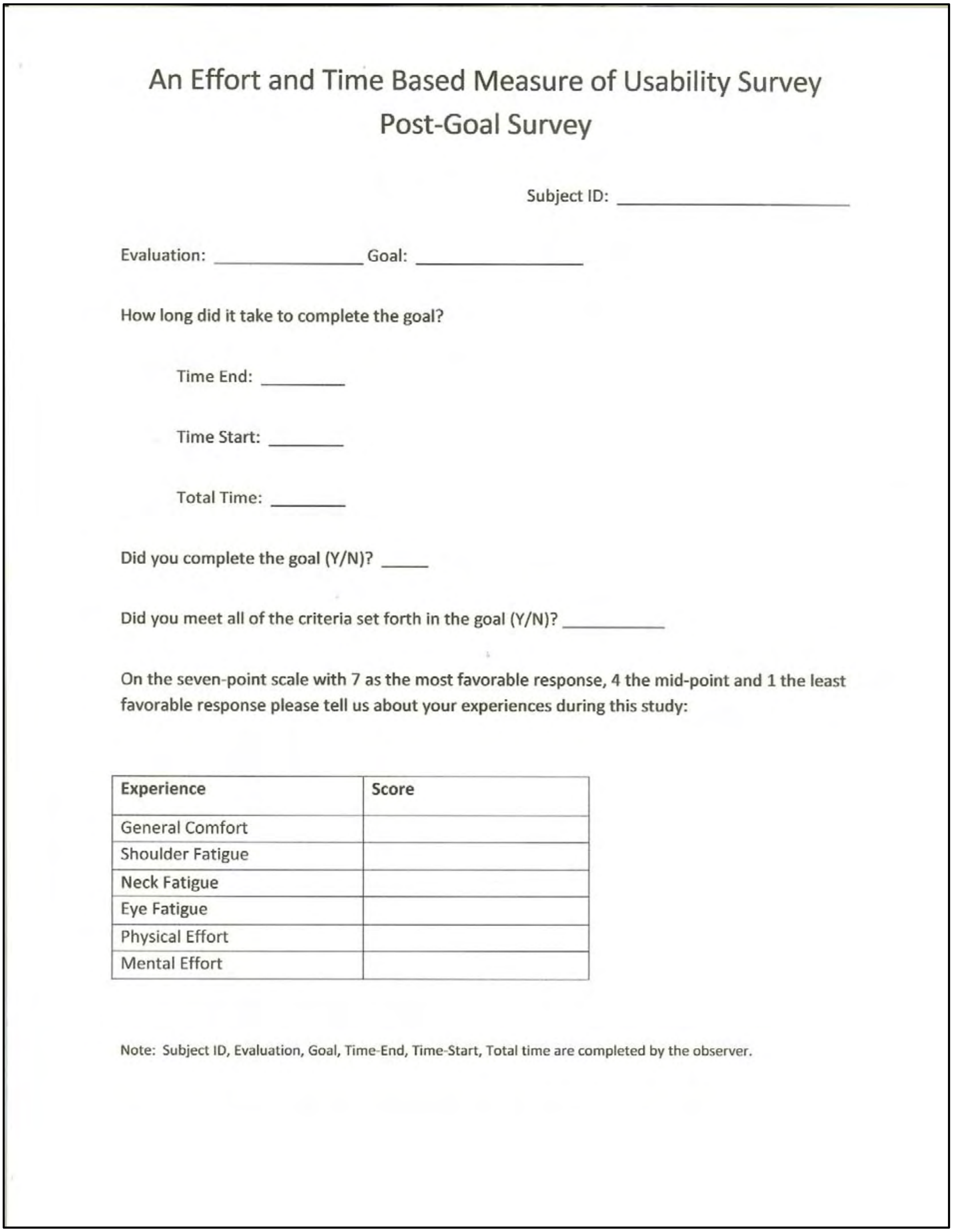
Approximately how much time do you spend using a computer every day: ________________

Approximately how much time do you spend on the internet: ________

Approximately how frequently do you make on-line travel arrangements? ____________

Approximately how many travel systems have you used? _____________

B.2 POST-GOAL SURVEY



An Effort and Time Based Measure of Usability Survey
Post-Goal Survey

Subject ID: ______________________

Evaluation: ______________ Goal: ________________

How long did it take to complete the goal?

Time End: ________

Time Start: _______

Total Time: _______

Did you complete the goal (Y/N)? _____

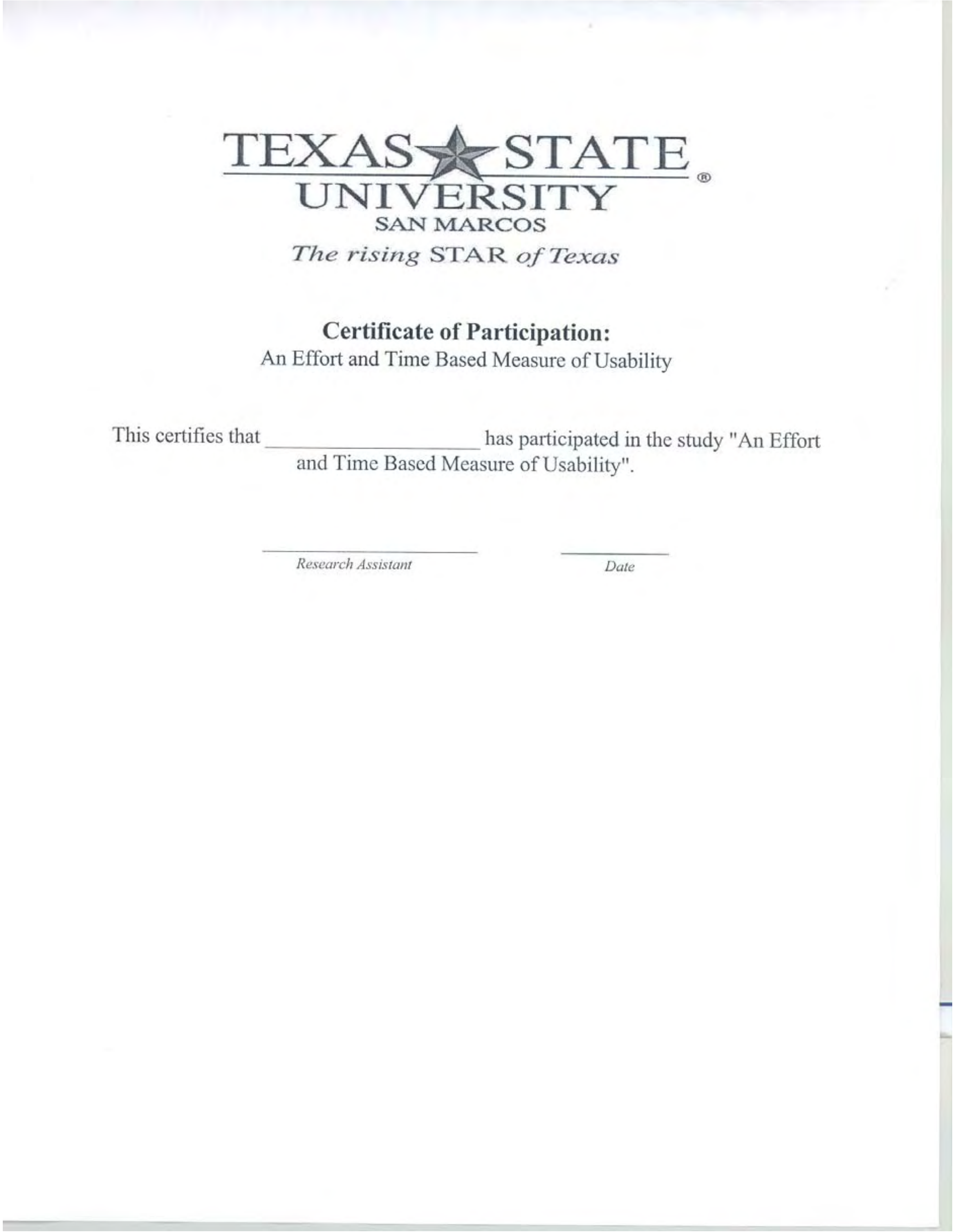
Did you meet all of the criteria set forth in the goal (Y/N)? __________

On the seven-point scale with 7 as the most favorable response, 4 the mid-point and 1 the least favorable response please tell us about your experiences during this study:

Experience	Score
General Comfort	
Shoulder Fatigue	
Neck Fatigue	
Eye Fatigue	
Physical Effort	
Mental Effort	

Note: Subject ID, Evaluation, Goal, Time-End, Time-Start, Total time are completed by the observer.

B.3 PARTICIPATION CERTIFICIATE



TEXAS STATE UNIVERSITY®
SAN MARCOS
The rising STAR *of Texas*

Certificate of Participation:
An Effort and Time Based Measure of Usability

This certifies that ____________________ has participated in the study "An Effort and Time Based Measure of Usability".

____________________ *Research Assistant*

__________ *Date*

APPENDIX C EVALUATION PROTOCOL

The following is a set of instructions for administering the eye-tracking pilot study. If you have any questions about these instructions, please ask Dr. Tamir, Dr. Komogortsev or Dr. Mueller for clarification.

Text in italics indicates directions that you are to follow. Bolded text indicates instructions that you are to provide to subjects. Please do not substantially deviate from or alter these instructions. Please adhere to these instructions as strictly as possible.

During the course of the experiment, you may be asked questions by subjects. Please do not provide any information other than what is contained in the consent form. If subjects request answers beyond the scope of the consent form, the consent form provides appropriate contact information for such requests.

Functionally blind persons and persons who are physically unable to use a mouse and keyboard while keeping their chin on a chin-rest for fifteen minutes are not eligible to participate in the study. If any ineligible persons volunteer for participation, perform only steps 1-6 and 19.

Please make sure that you read and understand the complete set of instructions before administering the study to any subjects. Do not administer the study until you have been trained to properly calibrate/recalibrate the eye-tracker and start/stop the logging utilities.

1. *Direct the subject to sit in a seat in front of the eye-tracker, then close the lab door most of the way (leaving it open just a crack), and put the "Do Not Disturb" sign on the door.*

2. *State the following:*

    Thank you for volunteering to participate in this study. Before we proceed, I'd like you to carefully review the following statement of informed consent. After reviewing the consent form, if you would like to continue, please sign and put today's date on the line labeled "Subject's Signature" and return the form to me.

3. *Give the subject one copy of "Consent Form: An Effort and Time Based Measure of Usability" that has been signed and dated on the line labeled "Researcher's Signature". After the subject signs and dates the form and returns it to you, sign your name and put today's date on the line labeled "Researcher Assistant's Signature." Place the form face-down on top of the forms in the "Consent Forms" folder.*
4. *Hand the subject one blank unsigned copy of "Consent Form: An Effort and Time Based Measure of Usability".*
5. *Open the coding spreadsheet. Put the subject's name into the next available space. Note the code next to the subject's name. This will be the subject's subject id.*
6. *State the following:*

    This copy of the consent form is yours to keep. We will now proceed with the study. Remember, you may withdraw at any time. If you wish to do so, please let me know and we will discontinue.

    Write the subject's subject ID on a "Subject Profile", hand it to the subject and ask them to complete it and return it to you. When the subject returns the form, place it in the Subject Profiles folder.

7. *If at any point the subject states a desire to discontinue, then immediately stop and skip down to step 19.*
8. *Open Tobii Studio and open the project named "Pilot study." Open a command prompt and in the logs directory, create a new subdirectory named for the subject's subject id.*
9. *On the eye-tracker computer, go to Control Panel, Internet Options, then under "Browsing history" click the "Delete" button, then click the "Delete all…" button, check the "Also delete files and settings stored by add-ons" box, then click "Yes". Next, prepare, but do not start recording, a mouse/keyboard log named* `[subject id]-[exercise #]`. *In Tobii, open a new recording session named* `[subject id]-[exercise #]`.
10. *State the following:*

    Please turn off your cell phone and any other electronic devices that you have with you at this time, and please remove any hats or non-prescription sunglasses that you are wearing.

    We are now going to take some measurements using the eye tracker. Please place your chin on the chin rest and direct your attention to the monitor. You may look at the monitor and blink your eyes as you normally would, but please do not remove your chin from the chin rest or move your head unless you wish to discontinue the experiment.

11. *Direct the subject to place their chin on the chin rest. If necessary, adjust the height of the chin rest so that the subject is looking directly at the monitor. If you have not run any experiments yet, minimize Tobii and state the following:*

> **In a few moments, you're going to see a circle with a dot in its center on the screen. Please follow the dot with your eyes. Try not to anticipate the movement of the dot. Remember, you may look at the monitor and blink your eyes like you normally would. We may repeat this process a number of times.**

Now run the accuracy calibration procedure then skip down to step 13. If the error rate for this procedure is not less than 50% or is not less than 3 degrees in one eye, skip down to step 19.

12. *State the following if necessary:*

> **In a few moments, you're going to see a circle with a dot in its center on the monitor. Please follow the dot with your eyes. Try not to anticipate the movement of the dot. Remember, you may look at the monitor and blink your eyes like you normally would. We may repeat this process two or three times.**

13. *Calibrate/recalibrate the eye-tracker. Do not make more than three calibration attempts or recalibrate more than twice. If the eye-tracker fails to gather any calibration data after three attempts, instruct the subject that they may now remove their chin from the chin-rest and skip down to step 19.*
14. *State the following:*

> **Please hold your head still and keep your chin on the chin rest while I read you some instructions.**

Until the conclusion of Step 16, make sure that the subject does not remove their chin from the chin rest unless they wish to discontinue. Make sure they do not obstruct the eye tracker with their free hand.

15. *State the following:*

> **You are now going to carry out the exercises which will be described on the sheet in front of you to the best of your ability. You will be using the keyboard and mouse in front of you, which you may adjust at this time.**
>
> **Try to follow the directions as closely as possible and as best as you can. These exercises are *not* a test of you or your skills. You are not being evaluated on your ability to complete the exercises or your ability to use a computer system.**
>
> **In these exercises, you will be given a task with certain requirements. You should try to meet the requirements as closely as possible, but you may complete the assigned task without precisely fulfilling every requirement.**
>
> **You may move your eyes from the monitor to the sheet and back, but please do not move your head or remove your chin from the chin rest unless you wish to discontinue. I cannot communicate with you in any way during the exercise. If at any point you are unsure of how to proceed, simply take whatever steps you think may be correct.**
>
> **You will be utilizing an actual travel website for these exercises, but you will not be booking any actual travel or making any actual purchases.**

Please do not enter any personal information into the system at any time (I will be monitoring as well to make sure that this doesn't happen).

You will be completing a total of ten exercises today, with periodic breaks. This will take approximately two hours in total.

Would you like me to review any of these instructions?

Review the instructions with the subject if necessary, but do not provide any information other than what is contained in these instructions and the consent form.

16. Ask the subject:

Are you ready to begin?

When the subject indicates that they are ready, place the next (or first if you have not run any exercises yet) goal sheet onto the bracket attached to the monitor. Be sure that the sheet does not obstruct the monitor.

State the following:

Please do not touch the keyboard or mouse until I tell you to begin.

Start the Tobii recording and mouse/keyboard logging. State to the subject:

You may begin.

If the subject asks for assistance, simply state: "I apologize, but I cannot help you." Do not assist the subject with the exercises in any way whatsoever, even if they request assistance. Do not let the subject enter any personal information at any point. The exercise is considered to be completed once a "login to complete this order" message is *displayed on-screen, the Web interface is non-responsive for two minutes, no progress is being made toward the goal for two minutes, or the subject states that they are finished with the exercise. Once the subject completes the exercise or ten minutes have elapsed (whichever comes sooner), stop the logging and recording, and inform the subject that the exercise is complete and they may now remove their chin from the chin-rest.*

Write the following in the appropriate fields on an "After Goal" form (please write all times in 24-hour/military format): Subject's subject ID, start time, stop time, elapsed time, website used, and goal number. Now hand the form to the subject and ask them to complete the remaining fields and return the form to you.

17. *State the following:*

    We will now continue with the next exercise.

18. *Repeat steps 9, 11-14, and 16-17 for exercises 2-10. If at any point the subject seems frustrated or upset, assure the subject that they are doing fine and remind them that they are not being personally evaluated or tested.*

19. *State the following:*

    > **Thank you very much for participating in this study. This concludes your participation. Please take your copy of the consent form with you, and thank you again.**

    If the subject desires proof of participation, sign and date a "Proof of Participation" form and give it to the subject. Inform the subject that they may show or not show this certificate to anyone completely at their discretion.

Dismiss the subject. If the subject wishes to discuss the study, you may do so with her or him at this time.

If the subject completed the experiment, then on the Coding Spreadsheet, in the "Completed experiment?" column, put "Yes."

If the subject did not meet the participation criteria, then on the Coding Spreadsheet, in the "Completed experiment?" column, put "No: Ineligible."

If the eye-tracker could not be calibrated for the subject, then on the Coding Spreadsheet in the "Completed experiment?" column, put "No: Failed calibration."

If the subject discontinued the experiment, then on the Coding Spreadsheet, in the "Completed experiment?" column, put "No: " and note the point at which the subject discontinued. If the subject completed any forms, file them in the appropriate folder.

APPENDIX D RAW DATA

D.1 SUBJECT PROFILES

D.1.1 SYSTEM A

ID	Age	Gender	Race	Glasses	Computer Usage	Internet Usage	Travel System Usage	Travel Systems
P0-101	20	M	C	N	2.5	2	2	4
P0-102	25	M	C	N	4	4	1	4
P0-103	26	M	C	Y	8	2	0	0
P0-104	22	M	H	N	1	1	0	0
P0-105	19	F	C	N	1	1	0	0
P0-106	23	F	O	N	6	4	2	4
P0-107	31	F	C	N	3	3	3	5
P0-108	29	M	C	Y	8	6	2	5
P0-110	26	M	C	N	5	3	4	3
P0-111	26	M	H	Y	7	2	0	0

D.1.2 SYSTEM B

ID	Age	Gender	Race	Glasses	Computer Usage	Internet Usage	Travel System Usage	Travel Systems
P1-113	28	M	C	Y	1	1	0	0
P1-114	34	F	H	N	3	2	0	2
P1-115	22	M	C	Y	4	3	2	3
P1-116	21	F	B	Y	4	4	2	2
P1-118	22	F	C	N	2	2	0	0
P1-119	22	M	C	Y	1	1	0	3
P1-121	34	M	O	Y	5	4	5	6
P1-122	24	M	B	Y	2	1	1	3
P1-124	34	M	H	N	6	3	2	4
P1-125	29	F	B	N	16	10	2	2

D.2 RAW DATA

D.2.1 MANUAL

D.2.2 SYSTEM A

ID	Mickeys	Clicks	Keystrokes	Transfers
p0-101-01.txt	59925	67	65	10
p0-101-02.txt	62224	75	83	8
p0-101-03.txt	34203	61	65	13
p0-101-04.txt	47073	55	82	9
p0-101-05.txt	57189	74	39	7
p0-101-06.txt	32966	32	74	9
p0-101-07.txt	49919	58	62	5
p0-101-08.txt	37287	42	36	19
p0-101-09.txt	44172	53	70	5
p0-101-10.txt	71298	69	68	11
p0-102-01.txt	37652	42	97	10
p0-102-02.txt	36324	89	45	15
p0-102-03.txt	37272	57	56	17
p0-102-04.txt	37528	56	95	23
p0-102-05.txt	27762	86	64	16
p0-102-06.txt	31408	49	52	16
p0-102-07.txt	37260	72	56	11
p0-102-08.txt	33649	64	54	15
p0-102-09.txt	21441	28	70	14
p0-102-10.txt	42831	69	49	10
p0-103-01.txt	84223	57	101	31
p0-103-02.txt	115581	115	100	17
p0-103-03.txt	114091	73	78	14
p0-103-04.txt	103254	98	112	17
p0-103-05.txt	109210	110	71	12
p0-103-06.txt	87817	92	101	16
p0-103-07.txt	111071	85	113	16
p0-103-08.txt	86968	96	101	10
p0-103-09.txt	55983	73	107	18
p0-103-10.txt	169197	134	182	26
p0-104-01.txt	26870	44	203	33
p0-104-02.txt	39259	43	55	14
p0-104-03.txt	44628	34	71	12
p0-104-04.txt	39326	54	109	15
p0-104-05.txt	33140	44	52	7
p0-104-06.txt	41238	49	62	8
p0-104-07.txt	36171	58	78	16
p0-104-08.txt	40414	63	64	7
p0-104-09.txt	28969	49	69	7
p0-104-10.txt	46400	81	82	8

ID	Mickeys	Clicks	Keystrokes	Transfers
p0-105-01.txt	36669	71	113	25
p0-105-02.txt	44376	72	73	9
p0-105-03.txt	17934	45	34	7
p0-105-04.txt	45459	71	70	9
p0-105-05.txt	26226	51	43	5
p0-105-06.txt	28374	57	40	5
p0-105-07.txt	27806	68	41	8
p0-105-08.txt	36269	64	97	13
p0-105-09.txt	32839	68	49	6
p0-105-10.txt	25350	42	38	11
p0-106-01.txt	39166	59	118	19
p0-106-02.txt	41825	56	57	12
p0-106-03.txt	36203	57	58	10
p0-106-04.txt	36121	59	127	28
p0-106-05.txt	57417	103	65	14
p0-106-06.txt	33252	51	46	13
p0-106-07.txt	49292	71	62	13
p0-106-08.txt	41580	84	86	18
p0-106-09.txt	48065	93	85	9
p0-106-10.txt	42015	80	69	10
p0-107-01.txt	55565	48	106	14
p0-107-02.txt	44944	55	58	9
p0-107-03.txt	44162	52	79	9
p0-107-04.txt	46771	45	93	24
p0-107-05.txt	41358	57	93	17
p0-107-06.txt	51211	62	87	15
p0-107-07.txt	36871	39	78	12
p0-107-08.txt	41822	58	89	13
p0-107-09.txt	35380	41	94	14
p0-107-10.txt	37572	50	67	19
p0-108-01.txt	30441	33	87	7
p0-108-02.txt	69525	112	56	7
p0-108-03.txt	37546	41	44	6
p0-108-04.txt	28377	54	54	6
p0-108-05.txt	21519	32	41	7
p0-108-06.txt	14316	26	39	17
p0-108-07.txt	21845	28	34	6
p0-108-08.txt	30621	38	34	7
p0-108-09.txt	24196	40	88	7
p0-108-10.txt	12967	25	33	4

ID	Mickeys	Clicks	Keystrokes	Transfers
p0-110-01.txt	57518	51	59	8
p0-110-02.txt	54619	53	71	10
p0-110-03.txt	54163	56	47	4
p0-110-04.txt	55121	73	100	5
p0-110-05.txt	59246	86	102	14
p0-110-06.txt	28497	36	64	10
p0-110-07.txt	17221	32	45	8
p0-110-08.txt	64736	103	141	19
p0-110-09.txt	39181	50	41	6
p0-110-10.txt	48276	84	62	10
p0-111-01.txt	18452	23	44	5
p0-111-02.txt	30461	47	18	6
p0-111-03.txt	15452	28	23	7
p0-111-04.txt	25979	39	64	8
p0-111-05.txt	9669	19	22	8
p0-111-06.txt	14508	33	25	6
p0-111-07.txt	14158	31	21	5
p0-111-08.txt	14605	34	17	5
p0-111-09.txt	10109	21	26	8
p0-111-10.txt	15890	28	21	5

D.2.3 SYSTEM B

ID	Mickeys	Clicks	Keystrokes	Transfers
p1-113-01.txt	25496	30	65	6
p1-113-02.txt	32373	57	29	6
p1-113-03.txt	24352	35	37	8
p1-113-04.txt	40990	70	48	6
p1-113-05.txt	22596	39	22	6
p1-113-06.txt	16164	31	17	6
p1-113-07.txt	14198	30	23	8
p1-113-08.txt	15771	33	15	6
p1-113-09.txt	19102	43	48	8
p1-113-10.txt	17435	41	23	9
p1-114-01.txt	24110	39	89	14
p1-114-02.txt	37744	54	53	9
p1-114-03.txt	22017	30	31	6
p1-114-04.txt	35973	71	39	5
p1-114-05.txt	32478	45	37	11
p1-114-06.txt	31579	43	26	5
p1-114-07.txt	26942	63	28	5
p1-114-08.txt	20788	47	31	6
p1-114-09.txt	33142	54	63	10
p1-114-10.txt	25213	46	30	7
p1-115-01.txt	31187	37	69	12
p1-115-02.txt	46576	42	43	8
p1-115-03.txt	30152	23	60	11
p1-115-04.txt	46615	51	50	7
p1-115-05.txt	48553	50	44	8
p1-115-06.txt	24463	31	37	8
p1-115-07.txt	36871	30	68	9
p1-115-08.txt	29570	52	38	8
p1-115-09.txt	28419	34	51	20
p1-115-10.txt	54749	46	48	8
p1-116-01.txt	58692	49	40	7
p1-116-02.txt	53099	51	34	6
p1-116-03.txt	49427	50	46	6
p1-116-04.txt	27417	36	50	10
p1-116-05.txt	63310	46	41	7
p1-116-06.txt	46449	47	35	6
p1-116-07.txt	19149	23	53	5
p1-116-08.txt	26410	35	35	6
p1-116-09.txt	16792	27	42	5
p1-116-10.txt	24773	33	49	6

ID	Mickeys	Clicks	Keystrokes	Transfers
p1-118-01.txt	60113	85	39	5
p1-118-02.txt	73785	60	20	6
p1-118-03.txt	35543	43	41	9
p1-118-04.txt	49887	60	37	16
p1-118-05.txt	32291	41	42	7
p1-118-06.txt	45429	53	22	6
p1-118-07.txt	32288	45	22	4
p1-118-08.txt	18049	34	17	5
p1-118-09.txt	34756	57	68	9
p1-118-10.txt	23404	40	30	4
p1-119-01.txt	49385	44	46	7
p1-119-02.txt	43051	40	33	11
p1-119-03.txt	30424	40	36	5
p1-119-04.txt	22290	38	48	6
p1-119-05.txt	15720	24	31	5
p1-119-06.txt	18162	23	26	6
p1-119-07.txt	26415	23	44	6
p1-119-08.txt	34465	40	40	7
p1-119-09.txt	24121	34	43	8
p1-119-10.txt	21723	35	34	4

ID	Mickeys	Clicks	Keystrokes	Transfers
p1-121-01.txt	35446	57	164	12
p1-121-02.txt	29580	55	78	21
p1-121-03.txt	52300	71	110	41
p1-121-04.txt	32250	37	107	24
p1-121-05.txt	46909	58	126	19
p1-121-06.txt	29942	49	160	26
p1-121-07.txt	32440	44	133	16
p1-121-08.txt	42678	56	49	7
p1-121-09.txt	31686	44	59	15
p1-121-10.txt	41324	59	170	23
p1-122-01.txt	76382	56	83	14
p1-122-02.txt	63457	89	39	19
p1-122-03.txt	34762	41	58	8
p1-122-04.txt	21982	36	57	7
p1-122-05.txt	40291	55	49	23
p1-122-06.txt	29434	35	45	19
p1-122-07.txt	16221	29	28	12
p1-122-08.txt	19055	38	63	9
p1-122-09.txt	33602	40	49	12
p1-122-10.txt	17918	36	33	9

ID	Mickeys	Clicks	Keystrokes	Transfers
p1-124-01.txt	48117	63	109	10
p1-124-02.txt	40552	47	27	6
p1-124-03.txt	18936	28	42	5
p1-124-04.txt	41571	61	92	16
p1-124-05.txt	41717	47	24	6
p1-124-06.txt	35731	47	45	7
p1-124-07.txt	23476	36	23	5
p1-124-08.txt	37088	39	16	7
p1-124-09.txt	36439	52	64	10
p1-124-10.txt	64927	74	49	11
p1-125-01.txt	34708	33	56	9
p1-125-02.txt	34311	48	45	9
p1-125-03.txt	22231	43	69	6
p1-125-04.txt	37765	64	102	8
p1-125-05.txt	44632	54	53	7
p1-125-06.txt	41935	51	38	10
p1-125-07.txt	38480	61	62	9
p1-125-08.txt	30865	52	82	12
p1-125-09.txt	29964	48	61	7
p1-125-10.txt	19175	31	48	6

D.3 EYE DATA

D.3.1 SYSTEM A

Name	Duration (min)	Validity LE	Average sac. amp.	Sac count	Fix counter	Fix average duration	Fix per	Sac per	Eye path travelled (deg)	Average pupil dilation (mm)
p0-101-01.tsv	599	63	3.42	944	739	443	85.22	14.78	4498	3.58
p0-101-02.tsv	496	81	3.72	1169	670	474	53.24	46.76	3838	3.55
p0-101-03.tsv	357	81	4	760	564	419	72.43	27.57	3469	3.35
p0-101-04.tsv	446	93	4.42	1096	764	371	42.82	57.18	4607	3.24
p0-101-05.tsv	392	84	4.04	807	637	366	47.26	52.74	3772	3.16
p0-101-06.tsv	289	89	3.91	669	421	478	49.63	50.37	2299	3.14
p0-101-07.tsv	331	89	3.87	769	473	478	42.93	57.07	2348	3.14
p0-101-08.tsv	244	88	4.34	597	372	353	33.03	66.97	2200	3.06
p0-101-09.tsv	319	92	4.27	669	533	330	29.31	70.69	2689	3.06
p0-101-10.tsv	440	86	3.81	947	629	518	69.54	30.46	3626	3.05
p0-102-01.tsv	589	90	4.05	1374	1023	463	80.6	19.4	6490	3.62
p0-102-02.tsv	535	91	4.34	1305	974	439	76.09	23.91	6699	3.48
p0-102-03.tsv	581	91	4.2	1380	1028	451	77.45	22.55	6368	3.42
p0-102-04.tsv	618	92	4.06	1498	1039	484	79.7	20.3	6700	3.41
p0-102-05.tsv	577	90	3.76	1268	891	523	78.12	21.88	5400	3.32
p0-102-06.tsv	464	90	4.26	992	723	515	77.08	22.92	5011	3.37
p0-102-07.tsv	521	92	4.23	1185	856	496	79.17	20.83	5679	3.4
p0-102-08.tsv	435	90	4.59	909	735	460	75.21	24.79	5066	3.38
p0-102-09.tsv	266	86	4.51	594	465	420	76.15	23.85	3415	3.39
p0-102-10.tsv	523	92	4.34	1233	923	464	79.94	20.06	6155	3.41
p0-103-01.tsv	640	82	3.62	453	888	229	21.64	78.36	4367	2.65
p0-103-02.tsv	576	89	3.93	590	884	263	23.77	76.23	4848	2.6

Name	Duration (min)	Validity LE	Average sac. amp.	Sac count	Fix counter	Fix average duration	Fix per	Sac per	Eye path travelled (deg)	Average pupil dilation (mm)
p0-103-03.tsv	578	89	3.68	648	946	260	24.55	75.45	5332	2.57
p0-103-04.tsv	560	85	3.64	401	773	225	19.81	80.19	4602	2.54
p0-103-05.tsv	402	87	3.84	425	602	261	23.4	76.6	3351	2.49
p0-103-06.tsv	483	88	3.74	434	730	265	22.46	77.54	4136	2.48
p0-103-07.tsv	467	88	3.5	444	717	261	22.83	77.17	3675	2.45
p0-103-08.tsv	463	88	3.76	529	795	279	26.67	73.33	4542	2.46
p0-103-09.tsv	334	86	4.08	306	492	246	22.47	77.53	2772	2.5
p0-103-10.tsv	595	87	4.19	519	895	242	21.34	78.66	5434	2.53
p0-104-01.tsv	615	71	3.27	677	395	932	52	48	2053	3.37
p0-104-02.tsv	596	86	3.92	1154	694	596	50.49	49.51	4102	3.33
p0-104-03.tsv	602	90	3.68	1020	650	674	52.21	47.79	3903	3.33
p0-104-04.tsv	618	88	3.86	1172	771	592	51.45	48.55	4918	3.34
p0-104-05.tsv	598	90	4.02	1081	770	597	52.29	47.71	5054	3.37
p0-104-06.tsv	625	90	4.22	1071	829	559	46.29	53.71	5262	3.33
p0-104-07.tsv	463	93	4.23	795	734	449	38.86	61.14	4265	3.41
p0-104-08.tsv	606	94	4.46	1132	933	482	42.62	57.38	5850	3.39
p0-104-09.tsv	379	88	4.61	579	514	455	36.47	63.53	3727	3.34
p0-104-10.tsv	525	90	4.69	837	752	471	38.47	61.53	4599	3.35
p0-105-01.tsv	619	67	4.1	900	679	548	84.47	15.53	4982	3.6
p0-105-02.tsv	619	84	4.19	1243	790	596	77.1	22.9	5481	3.18
p0-105-03.tsv	453	69	4.71	733	505	554	73.8	26.2	4011	3.13
p0-105-04.tsv	580	62	4.66	865	653	487	75.29	24.71	4882	3.17
p0-105-05.tsv	439	70	4.85	713	560	491	72.63	27.37	3995	3.12
p0-105-06.tsv	390	79	4.96	744	537	510	70.19	29.81	4240	3.14
p0-105-07.tsv	335	83	4.79	601	428	586	69.51	30.49	3250	3.12
p0-105-08.tsv	493	77	4.95	797	638	536	75.08	24.92	4779	3.24
p0-105-09.tsv	408	78	4.28	712	535	534	71.87	28.13	3570	3.13

Name	Duration (min)	Validity LE	Average sac. amp.	Sac count	Fix counter	Fix average duration	Fix per	Sac per	Eye path travelled (deg)	Average pupil dilation (mm)
p0-105-10.tsv	342	75	4.64	569	417	563	77.38	22.62	3084	3.24
p0-106-01.tsv	591	71	3.71	1149	717	470	57.89	42.11	4303	3.83
p0-106-02.tsv	593	70	3.71	1097	755	410	41.2	58.8	4009	3.66
p0-106-03.tsv	575	79	4.14	1154	822	418	48.47	51.53	5052	3.68
p0-106-04.tsv	597	71	4.75	1087	857	405	70.33	29.67	6323	3.58
p0-106-05.tsv	649	80	4.34	1252	931	411	41.46	58.54	5550	3.48
p0-106-06.tsv	428	78	5.19	825	604	459	68.13	31.87	4903	3.54
p0-106-07.tsv	423	66	5.13	614	498	490	78.94	21.06	3736	3.57
p0-106-08.tsv	331	79	4.76	645	520	435	75.84	24.16	3368	3.58
p0-106-09.tsv	424	81	4.54	797	649	459	76.8	23.2	4496	3.56
p0-106-10.tsv	392	83	5.3	803	674	412	72.66	27.34	4915	3.48
p0-107-01.tsv	597	96	4.94	1647	1119	447	75.69	24.31	8499	3.18
p0-107-02.tsv	600	95	5.02	1582	1149	435	76.99	23.01	8641	3.12
p0-107-03.tsv	550	93	4.79	1408	1074	414	74.5	25.5	7921	3.08
p0-107-04.tsv	603	92	4.8	1215	960	401	56.24	43.76	7248	3.07
p0-107-05.tsv	514	93	5.13	1341	992	413	70.11	29.89	8004	3.06
p0-107-06.tsv	602	83	4.92	1296	1025	404	65.43	34.57	8190	2.97
p0-107-07.tsv	387	87	5.26	780	668	418	62.27	37.73	5689	2.95
p0-107-08.tsv	557	88	5.23	883	840	354	45.11	54.89	6536	2.95
p0-107-09.tsv	397	89	5.19	601	590	388	47.96	52.04	4684	2.92
p0-107-10.tsv	430	91	5.47	984	768	431	66.41	33.59	6489	2.93
p0-108-01.tsv	391	94	3.78	961	659	496	73.5	26.5	3983	3.36
p0-108-02.tsv	637	96	3.66	1618	1044	531	70.23	29.77	5814	3.3
p0-108-03.tsv	312	95	3.98	761	512	524	74.76	25.24	3323	3.34
p0-108-04.tsv	337	95	3.66	803	521	560	71.94	28.06	3010	3.27
p0-108-05.tsv	201	94	3.66	435	316	547	69.8	30.2	1921	3.28
p0-108-06.tsv	193	95	3.57	418	282	594	71.94	28.06	1603	3.27

Name	Duration (min)	Validity LE	Average sac. amp.	Sac count	Fix counter	Fix average duration	Fix per	Sac per	Eye path travelled (deg)	Average pupil dilation (mm)
p0-108-07.tsv	238	95	4.15	544	385	533	70.77	29.23	2459	3.32
p0-108-08.tsv	282	95	4.22	629	439	555	71.57	28.43	2935	3.33
p0-108-09.tsv	302	94	3.87	615	461	555	74.82	25.18	2735	3.37
p0-108-10.tsv	145	95	4.52	321	226	538	66.63	33.37	1653	3.22
p0-110-01.tsv	659	96	4.81	1726	1235	449	77.03	22.97	8426	3.64
p0-110-02.tsv	569	92	4.81	1338	986	454	76.11	23.89	7041	3.73
p0-110-03.tsv	565	95	4.96	1299	918	524	77.07	22.93	6732	3.67
p0-110-04.tsv	605	95	4.74	1401	991	507	74.76	25.24	6888	3.62
p0-110-05.tsv	596	94	5.11	1237	1046	390	54.98	45.02	7203	3.65
p0-110-06.tsv	306	93	4.83	727	549	454	72.14	27.86	3829	3.62
p0-110-07.tsv	183	90	5	397	284	485	74.1	25.9	2173	3.68
p0-110-08.tsv	586	92	4.97	1258	932	494	71.14	28.86	6780	3.63
p0-110-09.tsv	386	96	5.21	891	677	481	72.27	27.73	4850	3.58
p0-110-10.tsv	366	93	5.14	859	650	457	72.5	27.5	4785	3.58
p0-111-01.tsv	394	89	4.15	824	700	334	37.71	62.29	4684	3.35
p0-111-02.tsv	477	89	4.09	839	864	334	36.73	63.27	5369	3.3
p0-111-03.tsv	313	89	4.14	572	539	341	37.87	62.13	3647	3.28
p0-111-04.tsv	461	90	4.03	815	790	356	37.52	62.48	5276	3.27
p0-111-05.tsv	237	88	4.44	468	406	346	39.38	60.62	3035	3.26
p0-111-06.tsv	266	86	4.19	416	441	317	33.67	66.33	2982	3.17
p0-111-07.tsv	197	85	3.78	247	318	307	31.84	68.16	2310	3.2
p0-111-08.tsv	251	87	4.02	426	403	314	32.06	67.94	2638	3.21
p0-111-09.tsv	188	91	4.16	328	317	347	34.14	65.86	2346	3.2
p0-111-10.tsv	274	86	3.97	426	450	319	34.24	65.76	3028	3.2

D.3.2 SYSTEM B

Name	Duration (min)	Validity LE	Average sac. amp.	Sac count	Fix counter	Fix average duration	Fix per	Sac per	Eye path travelled deg	Average pupil dilation (mm)
p1-113-01.tsv	304	86	4.11	871	461	296	30.82	69.18	2491	2.88
p1-113-02.tsv	298	97	5.46	120	106	157	6.97	93.03	711	2.78
p1-113-03.tsv	252	87	4.99	489	330	219	21.6	78.4	2079	2.87
p1-113-04.tsv	290	87	4.95	621	368	236	22.22	77.78	2549	2.85
p1-113-05.tsv	207	86	5.27	363	267	221	21.18	78.82	1899	2.9
p1-113-06.tsv	157	91	5.69	367	240	263	25.69	74.31	1720	3.01
p1-113-07.tsv	113	88	5.85	290	159	240	24.51	75.49	1371	3.03
p1-113-08.tsv	110	89	5.17	317	187	253	29.49	70.51	1436	3.05
p1-113-09.tsv	120	87	5.22	342	214	259	30.83	69.17	1338	3.05
p1-113-10.tsv	136	88	5.06	344	220	242	27.38	72.62	1506	3.02
p1-114-01.tsv	308	98	5.43	335	255	179	10.98	89.02	1742	2.77
p1-114-03.tsv	189	97	7.39	81	58	148	6.77	93.23	544	2.72
p1-114-04.tsv	311	95	6.04	166	127	155	7.77	92.23	1110	2.78
p1-114-05.tsv	235	92	6.1	72	55	156	6.17	93.83	426	2.68
p1-114-06.tsv	277	90	6.63	46	60	133	6.12	93.88	771	2.72
p1-114-07.tsv	235	90	6.73	33	53	141	6.33	93.67	495	2.77
p1-114-08.tsv	173	90	5.28	29	53	153	7.07	92.93	420	2.78
p1-114-09.tsv	201	92	4.06	38	66	150	7.59	92.41	510	2.74
p1-114-10.tsv	204	91	6.1	25	49	144	6.68	93.32	458	2.71
p1-115-01.tsv	294	87	4.22	944	602	285	45.48	54.52	4404	3.35
p1-115-02.tsv	215	87	4.47	546	439	245	37.95	62.05	2952	3.31
p1-115-03.tsv	147	88	4.93	499	317	278	47.14	52.86	2399	3.43
p1-115-04.tsv	158	89	4.62	518	324	310	49.77	50.23	2396	3.36
p1-115-05.tsv	197	85	4.26	509	364	316	44.8	55.2	2547	3.35
p1-115-06.tsv	95	90	4.6	317	195	310	50.15	49.85	1391	3.46

Name	Duration (min)	Validity LE	Average sac. amp.	Sac count	Fix counter	Fix average duration	Fix per	Sac per	Eye path travelled deg	Average pupil dilation (mm)
p1-115-07.tsv	132	93	4.87	448	277	333	47.94	52.06	2022	3.45
p1-115-08.tsv	141	88	4.68	448	315	285	47.86	52.14	2176	3.48
p1-115-09.tsv	111	88	5.46	339	222	318	50.07	49.93	2042	3.54
p1-115-10.tsv	200	89	4.16	543	391	330	49.52	50.48	2611	3.55
p1-116-01.tsv	315	82	6.55	185	177	183	11.29	88.71	1513	2.45
p1-116-02.tsv	293	77	5.71	158	151	187	12.26	87.74	1457	2.48
p1-116-03.tsv	264	80	4.8	130	148	206	11.47	88.53	1337	2.48
p1-116-04.tsv	154	78	8.57	47	53	169	8.12	91.88	569	2.46
p1-116-05.tsv	290	77	6.15	132	133	176	9.65	90.35	928	2.49
p1-116-06.tsv	189	80	5.33	82	106	188	11.22	88.78	858	2.51
p1-116-07.tsv	118	81	7.02	63	69	167	11.6	88.4	405	2.52
p1-116-08.tsv	123	83	6.21	76	86	183	12.25	87.75	853	2.47
p1-116-09.tsv	111	76	6.62	72	91	178	16.08	83.92	756	2.62
p1-116-10.tsv	129	85	6.47	58	79	170	11.18	88.82	837	2.54
p1-118-01.tsv	306	89	3.94	722	510	258	28.09	71.91	3030	2.86
p1-118-02.tsv	304	87	4.38	490	442	222	23.68	76.32	3666	2.84
p1-118-03.tsv	234	86	5.29	434	347	269	29.98	70.02	3270	2.93
p1-118-04.tsv	230	88	4.78	432	365	273	29.56	70.44	2694	2.85
p1-118-05.tsv	186	87	5.37	322	274	224	24.12	75.88	2355	2.85
p1-118-06.tsv	260	88	4.72	364	355	240	23.01	76.99	2746	2.82
p1-118-07.tsv	160	88	4.84	274	248	239	26.45	73.55	2133	2.87
p1-118-08.tsv	94	86	4.61	143	153	219	24.33	75.67	1365	2.88
p1-118-09.tsv	254	86	4.76	453	385	242	26.86	73.14	3247	2.85
p1-118-10.tsv	129	88	4.36	169	167	236	22.13	77.87	1433	2.84
p1-119-01.tsv	308	91	3.35	961	558	299	36.67	63.33	3475	3.01
p1-119-02.tsv	238	90	3.79	907	498	285	42.61	57.39	3626	2.91

Name	Duration (min)	Validity LE	Average sac. amp.	Sac count	Fix counter	Fix average duration	Fix per	Sac per	Eye path travelled deg	Average pupil dilation (mm)
p1-119-03.tsv	183	89	4.25	300	210	268	23.47	76.53	1402	2.91
p1-119-04.tsv	129	84	4.62	152	114	225	17.82	82.18	919	2.94
p1-119-05.tsv	84	83	4.18	223	139	352	46	54	1068	3.08
p1-119-06.tsv	121	85	3.94	133	96	273	19.06	80.94	1046	2.85
p1-119-07.tsv	95	85	4.11	157	107	311	26.66	73.34	854	2.79
p1-119-08.tsv	170	86	4.57	292	217	273	24.8	75.2	1793	2.87
p1-119-09.tsv	162	87	4.02	344	241	285	30.26	69.74	1621	2.96
p1-119-10.tsv	122	87	3.89	327	221	334	41.99	58.01	1611	2.91
p1-121-01.tsv	247	93	4.72	715	503	221	29.94	70.06	3818	2.98
p1-121-02.tsv	168	92	5.89	332	285	229	24.25	75.75	2446	2.94
p1-121-03.tsv	234	92	5.09	518	408	221	24.49	75.51	3036	2.87
p1-121-04.tsv	134	89	4.95	299	241	243	27.3	72.7	1801	2.86
p1-121-05.tsv	189	91	6.04	392	324	213	24.08	75.92	2866	2.84
p1-121-06.tsv	178	89	5.1	344	282	242	24.76	75.24	2130	2.82
p1-121-07.tsv	176	91	5.03	303	284	242	24.36	75.64	2081	2.75
p1-121-08.tsv	158	92	5.44	316	286	230	25.83	74.17	1960	2.73
p1-121-09.tsv	138	91	5.64	270	231	228	24.17	75.83	1603	2.77
p1-121-10.tsv	151	90	5.76	289	248	221	23.55	76.45	2093	2.83
p1-122-01.tsv	307	93	4.99	932	574	263	31.55	68.45	4744	3.33
p1-122-02.tsv	312	83	5.12	795	537	278	32.7	67.3	4188	3.17
p1-122-03.tsv	265	91	5.45	582	450	237	25.51	74.49	3883	3.28
p1-122-04.tsv	147	94	4.99	402	245	238	26.27	73.73	1976	3.17
p1-122-05.tsv	160	89	5.22	285	211	248	21.92	78.08	1740	3.25
p1-122-06.tsv	140	86	6.47	326	238	230	27.41	72.59	2459	3.24
p1-122-07.tsv	102	40	5.6	95	75	235	26.11	73.89	801	3.25
p1-122-08.tsv	128	78	6.59	187	177	237	24.81	75.19	1619	3.29

Name	Duration (min)	Validity LE	Average sac. amp.	Sac count	Fix counter	Fix average duration	Fix per	Sac per	Eye path travelled deg	Average pupil dilation (mm)
p1-122-09.tsv	269	71	5.51	386	336	254	26.72	73.28	3439	3.26
p1-122-10.tsv	123	84	5.98	259	201	245	28.94	71.06	1938	3.26
p1-124-01.tsv	305	84	4.48	336	409	210	19.74	80.26	2697	3.31
p1-124-02.tsv	288	88	4.07	331	382	209	18.84	81.16	2700	3.17
p1-124-03.tsv	172	88	3.95	189	212	205	17.26	82.74	1676	3.11
p1-124-04.tsv	306	86	4.26	304	372	208	18.03	81.97	2708	3.1
p1-124-05.tsv	304	86	4.02	274	333	201	16.65	83.35	2645	3.06
p1-124-06.tsv	303	78	4.08	326	392	205	20.54	79.46	2883	3.07
p1-124-07.tsv	189	81	3.72	226	237	226	21.15	78.85	1663	3.06
p1-124-08.tsv	263	80	3.93	232	290	214	18.06	81.94	2201	3.02
p1-124-09.tsv	257	82	4.36	260	298	209	18.68	81.32	2229	3.03
p1-124-10.tsv	312	84	4.43	261	371	204	17.68	82.32	2597	3.02
p1-125-01.tsv	306	71	4.79	422	416	242	29.12	70.88	2638	3.82
p1-125-02.tsv	288	82	4.45	476	403	281	30.95	69.05	2801	3.89
p1-125-03.tsv	248	66	3.96	379	363	248	35.68	64.32	2586	3.71
p1-125-04.tsv	304	64	4.68	484	429	261	35.69	64.31	3573	3.71
p1-125-05.tsv	295	75	5.11	361	375	260	28.35	71.65	2812	3.71
p1-125-06.tsv	229	77	5.22	299	286	246	26.6	73.4	2438	3.58
p1-125-07.tsv	263	77	5.3	261	275	263	24.65	75.35	2154	3.62
p1-125-08.tsv	206	76	4.64	187	229	255	24.65	75.35	1537	3.52
p1-125-09.tsv	237	76	4.82	274	264	283	26.71	73.29	2015	3.6
p1-125-10.tsv	137	74	5.57	198	177	229	25.79	74.21	1565	3.46

D.4 SUMMARIZED BY SYSTEM

D.4.1 SYSTEM A

D.4.2 EYE DATA SUMMARY

Goal	Duration	Validity LE	Average. sac. amp.	Sac count	Fix count	Average Fix duration	Fix per	Sac per	Eye path travelled (deg)	Average pupil dilation (mm)
1	569.40	81.90	3.99	1065.50	815.40	481.10	64.58	35.43	5228.50	3.42
2	569.80	87.30	4.14	1193.50	881.00	453.20	58.20	41.81	5584.20	3.33
3	488.60	87.10	4.23	973.50	755.80	457.90	61.31	38.69	4975.80	3.29
4	542.50	86.30	4.26	1035.30	811.90	438.80	57.99	42.01	5445.40	3.25
5	460.50	87.00	4.32	902.70	715.10	434.50	54.94	45.06	4728.50	3.22
6	404.60	87.10	4.38	759.20	614.10	455.50	57.70	42.30	4245.50	3.20
7	354.50	86.80	4.39	637.60	536.10	450.30	57.12	42.88	3558.40	3.22
8	424.80	87.80	4.53	780.50	660.70	426.20	54.83	45.17	4469.40	3.22
9	340.30	88.10	4.47	609.20	523.30	421.50	54.23	45.77	3528.40	3.21
10	403.20	87.80	4.61	749.80	638.40	441.50	59.91	40.09	4376.80	3.20

D.4.3 MANUAL DATA SUMMARY

Goal	Mickeys	Clicks	Keystrokes	Transfers
1	44648.1	49.5	99.3	16.2
2	53913.8	71.7	61.6	10.7
3	43565.4	50.4	55.5	9.9
4	46500.9	60.4	90.6	14.4
5	44273.6	66.2	59.2	10.7
6	36358.7	48.7	59.0	11.5
7	40161.4	54.2	59.0	10.0
8	42795.1	64.6	71.9	12.6
9	34033.5	51.6	69.9	9.4
10	51179.6	66.2	67.1	11.4

D.4.4 SYSTEM B

D.4.5 EYE DATA SUMMARY

Goal	Duration	Validity LE	Average. sac. amp.	Sac count	Fix counter	Fix average duration	Fix per	Sac per	Eye path travelled (deg)	Average pupil dilation (mm)
1	298.80	87.40	4.66	642.30	457.00	243.60	28.55	72.63	3055.20	3.08
2	258.40	88.00	5.07	423.60	349.91	224.10	24.18	76.30	2509.10	3.02
3	218.40	86.20	4.88	368.60	288.73	230.60	25.38	75.56	2277.80	3.04
4	209.60	85.10	5.25	333.10	260.73	231.90	25.09	75.91	1961.10	3.00
5	215.20	84.90	5.23	290.70	244.09	234.40	25.21	75.71	1963.10	3.03
6	188.10	85.40	5.19	259.10	222.00	233.80	24.49	76.52	1816.60	3.01
7	154.30	81.40	5.16	214.60	180.55	240.90	25.30	75.95	1390.40	3.01
8	176.00	85.00	4.99	223.60	226.55	229.90	24.19	76.03	1545.00	3.01
9	167.30	83.50	5.25	276.50	224.27	240.00	26.39	74.29	1874.80	3.04
10	169.50	85.60	4.99	339.20	257.73	249.60	29.10	70.64	2059.50	3.08

D.4.6 MANUAL DATA SUMMARY

Goal	Mickeys	Clicks	Keystrokes	Transfers
1	44363.6	49.3	76.0	9.6
2	45452.8	54.3	40.1	10.1
3	32014.4	40.4	53.0	10.5
4	35674.0	52.4	63.0	10.5
5	38849.7	45.9	46.9	9.9
6	31928.8	41.0	45.1	9.9
7	26648.0	38.4	48.4	7.9
8	27473.9	42.6	38.6	7.3
9	28802.3	43.3	54.8	10.4
10	31064.1	44.1	51.4	8.7

BIBILOGRAPY

[1] Software Engineering-Product Quality-Part 1: Quality Model. Geneva Switzerland: International Standards Organization, 2001.

[2] Software Engineering-Product Quality-Part 2: External Metrics. Geneva Switzerland: International Standards Organization, 2001.

[3] Experience Curve Effects. 2008. <http://en.wikipedia.org/wiki/Experience_curve_effects>.

[4] Anonymous. IEEE Std 610.12-1990 IEEE Standard Glossary of Software Engineering Terminology. New York, NY: Institute of Electrical and Electronic Engineers, 1990.

[5] Anonymous. IEEE Std 829-1998 IEEE Standard for Software Test Documentation. New York, NY: Institute of Electrical and Electronic Engineers, Inc., 1998.

[6] Anonymous. Software Engineering-Product Quality-Part 1: Quality Model. Geneva Switzerland: International Standards Organization, 2001.

[7] Anonymous. Software Engineering-Product Quality-Part 2: External Metrics. Geneva Switzerland: International Standards Organization, 2001.

[8] Anonymous. Merriam-Webster Online Dictionary. 2008. Merriam-Webster, Inc..

[9] Boehm, B., et al. *Characteristics of Software Quality*. American Elsevier, New York, 1978.

[10] Caulton, D. A. "Relaxing the homogeneity assumption in usability testing." Behavior & Information Technology 20.1 (2001): 7.

[11] Duchowski, A. *Eye Tracking Methodology: Theory and Practice*. 2nd ed. Springer, 2007.

[12] Dumas, J. S., and Redish, J. C. *A Practical Guide to Usability Testing*. 1993. Intellect Books, Portland, OR, USA, 1999. 1994.

[13] Ebbinghaus, H. "Memory: A Contribution to Experimental Psychology." 1885. <http://psy.ed.asu.edu/~classics/Ebbinghaus/index.htm>.

[14] Grady, R. *Practical Software Metrics for Project Management and Process Improvement*. Prentice-Hall, 1992.

[15] Kit, E. *Software Testing in the Real World*. Addison-Wesley, Reading, MA, 1995.

[16] Komogortsev, O. V., and Khan, J. "Eye Movement Prediction by Oculomotor Plant Kalman Filter with Brainstem Control." Journal of Control Theory and Applications 7.1 (2009).

[17] Komogortsev, O. V., and Khan, J. I. **Eye Movement Prediction by Kalman Filter with Integrated Linear Horizontal Oculomotor Plant Mechanical Model**. Proc. Eye Tracking Research & Applications Symposium (ETRA 2008), ACM Press (March 2008), 229-36.

[18] Leigh, R. J., and Zee, D. S. "The Neuology of Eye Movements." Oxford University Press, 2006.

[19] McCall, J. A., Richards, P. K., and Walters, G. F. Factors in Software Quality. : Nat'l Tech. Information Service, 1977.

[20] Myers, G. *The Art of Software Testing*. John Wiley & Sons, New York, NY, 1979.

[21] Nielsen, J. *Usability Engineering*. Academic Press, San Francisco, CA, USA, 1993.

[22] Poole, A., and Ball, L. J. "Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects." Encyclopedia of Human Computer Interaction. 2004.

[23] Pressman, R. *Software Engineering: A Practitioner's Approach*. 6th ed. McGraw-Hill, New York, NY., 2005.

[24] Rubin, J., and Chisnell, D. *Handbook of Usability Testing: How to Plan , Design, and Conduct Effective Tests*. Wiley Publishing, Inc., Indianapolis, IN, USA, 2008.

[25] Tamir, D., Komogortsev, O. V., and Mueller, C. J. "An Effort and Time Based Measure of Usability." Proceedings of the 6th international Workshop on Software Quality (May 10, 2008).

[26] Tullis, T., and Albert, B. *Measuring The User Experience*. Morgan Kaufmann, Burlington, MA, 2008.