TOWARD A LAYOUT-BASED PREDICTOR OF USER EFFORT

REQUIRED TO ACHIEVE SOFTWARE SYSTEM USABILITY GOALS

THESIS

Presented to the Graduate Council of
Texas State University-San Marcos
in Partial Fulfillment
of the Requirements

for the Degree

Master of SCIENCE

by

Liam Feldman, B.A.

San Marcos, Texas
August 2009

TOWARD A LAYOUT-BASED PREDICTOR OF USER EFFORT REQUIRED TO ACHIEVE SOFTWARE SYSTEM USABILITY GOALS

Committee Members Approved:

____
Carl J. Mueller, Chair

____
Khosrow Kaikhah

____
Oleg V. Komogortsev

____
Dan Tamir

Approved:

____
J. Michael Willoughby
Dean of the Graduate College

Dedicated to my goofball, Amanda Dunagin, with love from the goober.

ACKNOWLEDGMENTS

A Texas-sized “shout-out” goes firstly to all of my family members and friends. Thank “all y’all” so much for supporting me, encouraging me, and just plain putting up with me during the time that I dwelled in the deep, dark jungles of Thesisland.

Khosrow Kaikhah, the senior-most member of my thesis committee, provided valuable perspective, insight, and guidance which shaped the work of my thesis from its inception. It has been a privilege to be his student and his colleague.

The Usability Studies Research Group at Texas State University-San Marcos generously allowed me to participate and collaborate in their research. The work of this thesis would have been impossible if they had not kindly welcomed me into their ranks.

Dan Tamir, thesis committee member, provided the conceptual bedrock of time and effort based metrics, and also helped me a great deal with the structure and flow of my writing. I am thankful to him for including me in his work and letting me contribute to a number of his publications.

Oleg Komogortsev also granted me the privilege of being his co-author on a number of papers, and provided a good deal of encouragement for my work. He also very generously allowed me use of his MATLAB eye-movement analysis scripts, without which a good deal of the data analysis of this work would have been impossible.

Finally, my utmost thanks go to my thesis advisor and chair of my thesis committee, Carl Mueller, for being my mentor, encouraging me throughout this process,

believing in my capabilities, pushing me, not letting me settle for mediocrity, and reminding me that the “perfection” asymptote is never reached, but the “good enough” margin is attainable. I sincerely hope that this work turned out to be good enough.

This manuscript was submitted on July 9, 2009.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

ABSTRACT

TOWARD A LAYOUT-BASED PREDICTOR OF USER EFFORT

REQUIRED TO ACHIEVE SOFTWARE SYSTEM USABILITY GOALS

by

Liam Feldman, B.A.

Texas State University-San Marcos

August 2009

SUPERVISING PROFESSOR: CARL J. MUELLER

In order to learn how to operate unfamiliar software systems, users must expend mental and physical energy, which may be objectively and quantitatively measured as effort. This thesis hypothesizes that the amount of effort needed by users to achieve operability goals is intrinsic to aspects of system interface layout. To test this hypothesis, two experiments are conducted wherein effort expenditure by users is measured during interaction with varying software systems with differing interface layout properties. The findings of the experiments demonstrate a correlation between the intrinsic effort of an interface and its usability as predicted by extant interface layout guidelines. Based on empirical results, a widget-based predictor of user effort required for goal achievement is derived.

CHAPTER I

USABILITY EVALUATION AND SOFTWARE TESTING

In 2001, the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) adopted the following position regarding software quality: "Computers are being used in an increasingly wide variety of application areas, and their correct operation is often critical for business success and/or human safety. Developing or selecting high quality software products is therefore of prime importance (ISO, 2001)." The ISO/IEC Quality Model identifies usability as one of six fundamental attributes of software quality, and lists several usability sub-characteristics including understandability, learnability, and operability.

Fundamental software quality attributes are assessed by means of evaluation activities or testing. The Institute of Electrical and Electronics Engineers (IEEE) defines *testing* as: "An activity in which a system or component is executed under specified conditions, the results are observed or recorded, and an evaluation is made of some aspect of the system or component (IEEE, 2008)." Comprehensive testing is an activity crucial to the success of any given software engineering process. Testing of usability characteristics is particularly important, given that usability strongly correlates with a product's salability, reputation, supportability, training and documentation-related expenses, and potential for adverse legal action (Pressman, 2005).

In spite of the fact that usability testing is a critical endeavor, research conducted in 2007 determined that software engineers routinely neglect to perform usability testing as part of their development process. The majority of developers surveyed regarded usability evaluations as being useless. A minority found them to be valuable—yet let them fall by the wayside anyway (Vukelja, Müller, & Opwis, 2007).

Perhaps this is because usability evaluation today seems to be more informal art than formal science. There are a multitude of cognitive evaluation methods for testing the usability of software systems, including heuristic evaluations, walkthroughs, predictive modeling, and logging actual use. While techniques such as these can yield valuable insight into the usability-related issues which might be plaguing a software system, these methods are geared toward yielding qualitative information based on subjective data, and require that usability experts analyze and interpret results.

Software engineering expert Pressman (2005) asserts, “Although there is significant literature on the design of human/computer interfaces, relatively little information has been published on metrics that would provide insight into the quality and usability of the interface.” There is currently a shortage of techniques which yield quantitative information based on objectively measured data. As noted software engineering author Myers (2004) notes, the typical software engineer reacts to the subjective nature of contemporary usability testing with frustration and skepticism.

According to the ISO/IEC 9126-1:2001(E) standard, “The level of quality in the users' environment may be different from that in the developers' environment, because of differences between the needs and capabilities of different users and differences between different hardware and support environments. The user evaluates only those attributes of software, which are used for his tasks." This same standard delineates three complementary but separate views of software quality: External, internal, and quality-in-use, i.e. quality from the point-of-view of the user (ISO, 2001). Despite the fact that usability falls under the external and internal quality views in the standard, published material on usability design by-and-large is concerned with usability from the user's perspective. While notions such as: "Usability means focusing on users," or, "Users' needs [should] drive design decisions (Dumas & Redish, 1999)," are noble and ring true, they are not necessarily helpful from an engineering viewpoint.

Given the strong need for high quality in software, and given that usability is a fundamental quality attribute, the current disconnect between usability testing and conventional software engineering practice is worrisome. DeMarco (1982) penned the often repeated observation that, "You can neither predict nor control what you can't measure." The author of this thesis in turn asserts that if usability is not quantified, then it cannot be consistently engineered.

A comprehensive framework of quantitative usability measures and metrics, one which is capable of providing objective information useful to software developers, appears to be forthcoming. It seems likely that such a paradigm would benefit not only the practice of usability evaluation, but also the process of engineering usability into software systems. As noted usability design and evaluation practitioner Nielsen (1993) points out, "A Holy Grail for many usability scientists is the invention of analytic methods that would allow designers to predict the usability of a user interface before it has even been tested."

Viewing usability from an effort-based standpoint may very well provide such a predictive methodology. Effort appears to be the atomic quantity by which usability, operability and learnability of a system may be measured. As will be discussed in Chapter III, the notion of effort as a driver of usability is found in a number of places in the literature, but an effort-based approach to measurement has not yet been tried and appears to be novel.

The research in this thesis examines and tests the hypothesis that effort expended in usability testing is driven by the interface characteristics of the software system under test. The layout characteristics of interfaces in particular will be examined and tested, and the initial components of a predictive model of effort intrinsic to software system usability will be formulated. The crux of the research is an investigation of the time and effort required by users to achieve usability goals in software systems that have varying interface properties.

Chapter II of this work further elaborates on the various ways in which software system usability is currently defined, designed for, inspected and evaluated, and provides further discussion on how certain conventional usability evaluations do not conform well to traditional software testing traits. An overview of various standards, guidelines and best practices pertaining to interface design, including a discussion of Fitts' law and its applicability to graphical interfaces, is provided in Chapter III. Chapter III also details a measurement framework derived from effort and time based measures, and discusses the implications of such a framework. The experimental protocol for this thesis' research and analysis of results is given in Chapter IV. Chapter V discusses the usage of interface characteristics to model the intrinsic effort and time associated with software system usability. Direction for future research is provided in Chapter VI.

CHAPTER II

USABILITY DEFINITIONS AND QUALITATIVE EVALUATION METHODS

Fundamental notions of how to define interface usability differ among design and evaluation practitioners. A spectrum of direct and indirect definitions of usability and its components may be found in the literature. These range from the simple three-part breakdown found in ISO 9241-11 to the extensive multi-view hierarchy found in ISO 9126-1. This chapter examines some of the principle ways in which usability is defined, and also breaks down some usability evaluation methods which are commonly practiced but are not directly utilized in this thesis' research.

Definitions and Models of Usability

ISO 9241-11 is one of the oft-cited standards among the usability community (Jokela, Iivari, Matero, & Karukka, 2003). This standard frames usability in terms of the key characteristics *effectiveness*, *efficiency*, and *satisfaction*. A critical mandate of the ISO 9241-11 standard is that usability must be looked at in terms of specific users accomplishing specified tasks in a specific usage context (ISO, 1998).

Usability evaluation experts Dumas and Redish (1999) draw heavily upon ISO 9241-11 in defining usability. Their definition essentially paraphrases ISO 9241-11's notions of *effectiveness* and *efficiency*, although they prefer the terms *productivity* and *ease-of-use*. Nielsen (1993), on the other hand, is more expansive in delineating usability characteristics. He asserts that there are five dimensions, which comprise usability:

Learnability, efficiency, memorability, errors, and *satisfaction*. These same five fundamental characteristics, expressed as the attributes *facilitates learning*, *helps learners remember*, *reduces likelihood of errors*, *enables efficiency*, and *makes users satisfied*, may be found in Pressman (2005).

As mentioned in Chapter I, the ISO 9126-1 standard stands in contrast to ISO 9241-11 in that it details three separate but complementary perspectives on software quality: *External, internal,* and *quality-in-use*. Usability falls under the purview of the external and internal views. The ISO-9126 definition of usability brings several characteristics into play (emphasis added): "[Usability is] the capability of the software product to be *understood*, *learned*, *used* and *attractive* to the user, when used *under specified conditions.* Some aspects of *functionality*, *reliability* and *efficiency* will also affect usability." Other important usability attributes listed elsewhere in the standard include *operability*, *memorability*, *recoverability,* and *suitability* (note that a characteristic of the same name may be found in the internal quality view under functionality, but in the context of usability, *suitability* has to do with the presence of appropriate functions to accomplish goals and not deviation from specifications or stated objectives) (ISO, 2001).

Mueller (2009) provides an Ishikawa or "Fishbone" diagram which lends further hierarchical structure to the several characteristics of usability by framing these attributes within the context of productivity issues (Figure 1). Tamir, Komogortsev and Mueller propose the notion that all usability characteristics can and should be examined in terms of the underlying dimension of effort. They assert that, "usability relates to the physical effort that is required in order to use software in the accomplishment of interactive tasks."

Usability traits and associated sub-traits, according to these researchers, should be defined based upon measures of effort and time (Komogortsev, Mueller, Tamir, & Feldman 2009; Tamir, Komogortsev, & Mueller 2008).

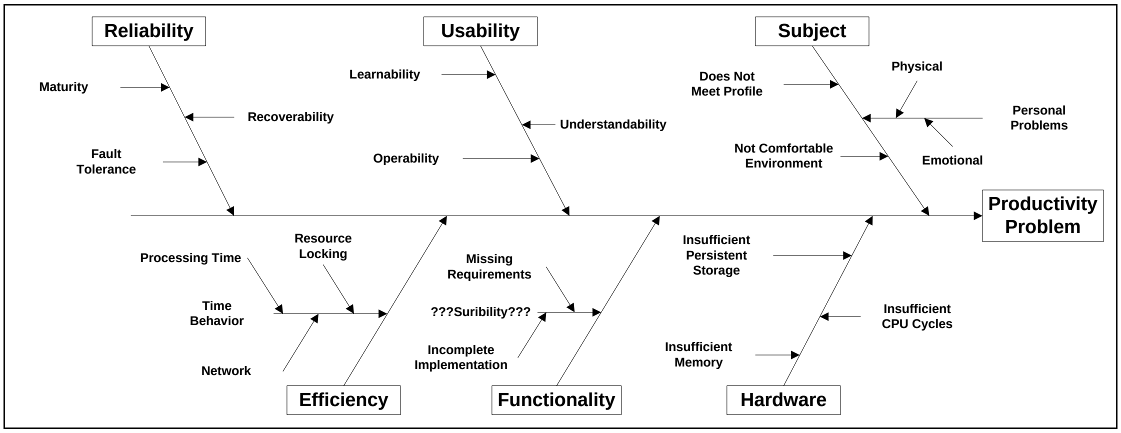


Figure 1. Mueller's Productivity Hierarchy.

The relationship between usability and effort is supported by Bevan (2001), who cites ISO/IEC 9126 (a predecessor to ISO/IEC 9126-1) in defining usability as, "a set of attributes that bear on the effort needed for use, and on the individual assessment of such use, by a stated or implied set of users." Jones (1997) concurs with a similar definition of usability: "Usability is the total effort required to learn, operate, and use software or hardware." This thesis thus adopts the Tamir et al. (2008) and Komogortsev et al. (2009) operative definition of usability.

Usability Testing and Evaluation – Qualitative Methodologies

Despite the fact that axiomatic definitions of usability vary widely among evaluators, several works have been written on how to conduct usability testing, including seminal publications by Dumas and Redish (1999) and usability evaluation expert practitioners Tullis and Albert (2008). Additional guidance on evaluating usability includes Dieli's (1988) problem-solving approach to test planning and a system for examining how much effort software developers put into usability design by Granollers and Lorés (2006).

Heuristic evaluation is a commonly employed technique in which the judgments of experts are drawn upon and aggregated. According to Human-Computer Interface experts Sharp, Rogers and Preece (2007), a heuristic evaluation is, “A usability inspection technique… in which experts, guided by a set of usability principles known as *heuristics*, evaluate whether user-interface elements… conform to [these] principles.” This technique is advantageous in that a collection of multiple perspectives will often discern issues that would have been missed by only one evaluator. However, it has been empirically determined that this method exhibits rapidly diminishing returns once the number of evaluators exceeds five (Sharp et al., 2007). Another drawback to this technique is expense: The cost of hiring one expert evaluator, much less several, can be substantial. It should also be noted that an evaluation based upon approximations of user behavior is probably going to be less precise than testing of actual users would be.

The *walkthrough* is a method in which the evaluator, “Walk[s] through a task with the system and note[s] problematic usability features (Sharp et al., 2007).” There are two common varieties of walkthroughs: The *cognitive walkthrough,* which consists of, “Simulating a user’s problem-solving process at each step in the human-computer dialog, checking to see if the user’s goals and memory for actions can be assumed to lead to the next correct action (Sharp et al., 2007),” and the *pluralistic walkthrough,* which is where, “Users, developers and usability experts work together to step through a scenario, discussing usability issues associated with dialog elements involved in the scenario steps (Sharp et al., 2007).” Cognitive walkthroughs may be performed using in-house personnel, making them less costly than heuristic evaluations, but the process can be time-consuming, labor-intensive, and lacking in objectivity (Sharp et al., 2007). Pluralistic walkthroughs often yield more objective information than cognitive walkthroughs, but they are more expensive and have a greater chance of veering off-track (Sharp et al., 2007).

Prototypes can be utilized as an evaluation technique earlier on in the development process than can other methods. Sharp et al. (2007) state that, “A prototype can be anything from a paper-based storyboard through to a complex piece of software.” Prototypes may be tested both internally and externally, may be as simple and cheap or complex and expensive as the design team desires, and are compatible with other evaluation techniques. It is unclear, however, how close prototypes must be to the final product design in order to be a valid tool. Also, a given prototype may not accurately simulate the actual performance of the system being designed.

Focus groups are a study of user interactions with a given product in a controlled setting. *Field studies* are similar to focus groups except that the study takes place “in the field,” that is to say in the environment where the user will actually use the product. In both field studies and focus groups, a facilitator moderates and guides the study process by prompting and questioning the user. Dumas and Redish (1999) caution that a focus group or field study by itself is not sufficient enough to provide valid usability data. These authors further warn that there is currently a “lively debate” within the usability evaluation community as to whether or not field studies have any utility whatsoever.

The methods detailed above are substantially subjective and yield results which are qualitative in nature. Even a cursory glance at works such as Myers (2004), Kaner, Falk and Nguyen (1999), and Fenton and Pfleeger (1997) reveals that there are substantial differences between the respective theories of software testing and usability evaluation. The differences are particularly striking when it comes to measurements.

Measurements or metrics are the foundation of any valid scientific or engineering process, according to Fenton and Pfleeger (1997). These authors take the point-of-view that software metrics should increase understanding of, permit control over, and foster improvement of processes and products, and should yield data which is correct, accurate, precise, and consistent (ideally, data should also be associated with a particular activity as well as be replicable). Pressman (2005) lists a slightly different enumeration of desirable attributes of software measurements: Simplicity, computability, consistency, objectivity, and consistent use of units.

Due to various issues such as lack of objectivity, compromised independence-of-review, inappropriate perspective, approximation, assumption, and an inability to control for numerous variables, the author of this thesis concludes that techniques in the vein of heuristic evaluation, walkthroughs, prototyping, focus groups and field studies will not and can not produce data which conform well to all or even most of the desirable attributes of metrics listed above. Methods such as the ones discussed above yield results grounded in a basis of opinion. As such, these results may only be evaluated in terms of a nominal scale.

Adherents to the above-listed techniques might counter that normalization devices such as Likert or semantic differential scales are sufficient to map qualitative data to an ordinal scale. While this assertion is true, one must take utmost care to ensure that such mapping is performed in a valid manner. Even if the mapping in question is performed correctly, one still cannot correctly draw the same conclusions as one could when utilizing a ratio or an absolute scale. For example, one might legitimately assert that a survey response of “5: Strongly Agree” is three rating points higher than a response of “2: Disagree,” but it would be specious to argue that the former is 2.5 times as positive a response as the latter. The author also notes that there are several methodologies for standardizing, normalizing and aggregating usability study data into uniform measures or scoring systems – notable examples of these include works by Sauro and Kindlund (2005), Hornbæk and Law (2007), Hazenzahl and Sandweg (2004), and Bertoa, Troya, and Vallecillo (2006) – but in the author’s opinion, these techniques are flawed in that they either intermingle objectively and subjectively derived data, or treat measures with a subjective basis as if they were objective in nature.

The methods summarized above are not suitable for isolating quantitative information on the physical and mental effort required to interact with a given software or hardware interface. A technique such as a pluralistic walkthrough might be useful for qualitatively enhancing other methods, but as has been discussed, the nature of techniques such as these does not fit well with the study methodology described by and utilized in this thesis.

CHAPTER III

QUANTITATIVE EVALUATION METHODS AND RESEARCH HYPOTHESES

There are instances where usability test results are objectively quantifiable using absolute time or counting scales. Examples of such include Nielsen's "typical quantifiable usability measurements," such as task completion time, successful functional executions versus error occurrences, and functions used or unused (Nielsen, 1993); Tullis and Albert's thorough discussion of performance-based metrics in *Measuring the User Experience* (2008); and Pressman's true/false usability indicators, e.g., "Are navigation mechanisms, content, and functions placed in a manner that allows the user to find them quickly (Pressman, 2005)?"

This chapter discusses time and effort measurements of usability, with a focus on the methods of *logging actual use* and *predictive modeling* and a particular emphasis on the use of *Fitts' Law* as a predictor for efficient user performance. The chapter concludes with a delineation of the research hypotheses for this thesis.

Logging Actual Use and Effort and Time Based Measures

Logging actual use employs instrumentation to record usage of a given interface by users in that interface's probable usage context, i.e. "in the field." This method logs "real life" interaction with an interface by "real users" without reliance on a test mediator or facilitator. As Nielsen (1993) points out, a principal advantage to logging actual use techniques is that they "…show how users perform their actual work." Logging actual use is not difficult to implement and is perhaps the most objective source of usability data possible. However, there are disadvantages to logging actual use: Privacy and ethical concerns must be addressed, data processing and reduction can be labor-intensive, and qualitative information cannot be obtained by this method alone.

Tullis and Albert's (2008) *Measuring the User Experience* provides an overview of various physiological instruments which can be employed to quantitatively log usability-related information. These instruments are well-suited for capturing quantitative data, but can be difficult to implement in a testing environment. The use of devices such as facial electromyogram sensors, galvanic skin response meters and heart-rate monitors may be impractical to utilize or prove to be overly intrusive for test subjects. Data on pupillary response, verbalization, and non-verbal behavior may be easier to capture, but require a specially trained observer for valid interpretation. Among the various physiological measurement devices, the one that appears to strike the best balance between objectivity of captured data, ease of implementation, practicality and subject comfort is the infrared light eye-tracking camera system, or eye-tracker.

Komogortsev et al. propose a framework of effort and time based measures of usability (Komogortsev, Mueller, Tamir, & Feldman, 2009) which employs logging actual use techniques and eye movement data from an eye-tracker. Effort and time based measures quantify the effort users must exert in order to use a given software product. Overall usage effort is conceptualized as being a combination of mental and physical effort. Mental effort is quantified as the cognitive effort indicated by eye movement measures, plus a static estimation of other contributing mental factors. Physical effort is derived from measures of manual effort, measures of eye movements, and a static estimation of other contributing physical factors.

Manual effort measures may include (but are not limited to) the keystrokes, mouse clicks, mouse movements, and switches between input devices (i.e. changing from mouse to keyboard or vice-versa) logged while subjects interact with a given interface. Eye-movement metrics may incorporate (but are not limited to) fixations (the stabilization of the gaze-point within a limited visual range), saccades (a rapid transition from one fixation point to another), pupil dilation, total eye-path traversal distance, and computations of extraocular muscular force recorded by an eye-tracking device.

Mueller, Tamir, Komogortsev, and Feldman (2009) have conducted a between-subjects usability study of two different online travel systems, “System A” and “System B”. Twenty subjects, ten using System A and ten using System B, each executed a set of ten homologous travel-booking tasks. An eye-tracker, coupled with a utility to log keyboard and mouse activity, recorded data. The results of that study showed a significant variation in the amount of user effort required to complete exercises using System A versus System B. Mueller et al. conducted their testing under tight budgetary and resource constraints, yet still managed to generate significant findings. This implies that usability testing using effort and time based measurements holds promise as an easy-to-implement, cost-effective methodology, producing results that are more objective, quantifiable and reproducible than other evaluation methods.

Predictive Modeling

Predictive models, like heuristic evaluations, are tools for evaluating a system without direct input from a user base. Unlike heuristic evaluations, predictive models rely upon empirically-derived formulas, functions and frameworks rather than expert opinions. Frequently utilized models include GOMS, the Keystroke-Level Model, and Fitts' Law (Sharp et al., 2007). Other examples include Sears' formulation of a "Layout Appropriateness" usability metric based upon widget layout and interaction sequences (Sears, 1993), and an examination by Brinkman, Haakma and Bouwhuis (2007) of "the physical interaction effort to operate components in a single device."

Fitts' Law is a simple predictive model which dictates that the time needed for acquisition of a stationary target by a subject using a moving object will vary depending upon the distance to the target and the size of the target. Specifically, Fitts' Law may be formulated as:

$$T = a + b \log_2 (D/W + 1)$$

Where T equals mean time to acquire a fixed target, a is a constant value representing time necessary to move a device plus time necessary to halt a device, b is a measure of the inherent movement rate of a device, D is distance from target, and W is target size (Shneiderman & Plaisant, 2005).

Technical design researchers Moore and Fitz (1993) have applied six principles or *Laws of Gestalt Psychology* to a predictive model of design best practices:

1) *Proximity*, proximally placed objects tend to be perceived as a grouping.
2) *Closure*, a closed shape correlates with perceptions of completeness or wholeness.
3) *Symmetry*, symmetrically arranged text or graphics conveys a sense of balance and aesthetic appeal.
4) *Figure-ground segregation*, a shape must be distinguishable from its surroundings to be perceivable.

5) *Good continuation*, observers tend to follow natural extensions of a shape e.g. the focal point of an arrow.
6) *Similarity*, objects with comparable styles, colors, shapes, or other aspects will tend to be associated together.

Research Hypotheses

Usability literature contains a wealth of guidelines, which informally could be looked at as the “What should one do?” portion of the collective body of design knowledge. A great deal has been written on the subject of precisely which standards, guidelines and best practices a good design practitioner should adhere to. What seems to be lacking, however, are objectively verifiable and quantifiable measurements of the effects of either holding to or deviating from best practices. The consequences of poor software usability are well-known and are discussed in Chapter I, but precisely why, one might ask, is a lack of usability associated with such dire consequences? For every “What should one do?” how does one go about quantifying and formulating the answer to the corollary, “…and why should one do it?”

Answering this question requires the use of an appropriate set of metrics such as those provided by Komogortsev et al. (2009). These researchers have provided a novel but promising means of re-examining and scientifically validating traditional usability standards. Grounding research within a framework of time and effort metrics means that informal, “tried-and-true” standards may now be formally and scientifically scrutinized and verified, or perhaps even refuted.

Fitts’ Law has been selected as an initial jumping-off point for research grounded in effort and time-based measurements. Strictly speaking, Fitts’ Law deals only with a relationship between time, distance, target size and target acquisition. However, time and effort, according to the research of Mueller et al. (2009), are strongly correlated. Given such a strong interconnection between effort and time, Fitts' Law takes on implications beyond what its formula literally expresses. This research posits that Fitts' Law may be extendable and applicable to measures of effort as well. Time, mouse movements, mouse clicks, eye movements, and eye fixations could conceivably all be functionally related to the distance one must travel to acquire a fixed target with the moving target of a mouse pointer.

If Fitts' Law mathematically dictates that effort and time to accomplish tasks (and therefore usability) increases as proximity decreases, then the Gestalt Law of Proximity might provide insight into the reasons behind this phenomenon. According to the Law of Proximity, "proximate objects appear to be distinguished from other groups of objects, even if their individual members are of radically different shapes and functions (Moore & Fitz, 1993)." The implication here is that decreasing the distance between two functionally connected but physically incongruent classes of elements, such as text labels and text-entry fields, strengthens the perceived relationship between these element classes. As with Fitts' Law, there is again an implication that measures of effort and time will correlate with the distance between functionally interconnected elements.

Element proximity in interface layout is a key independent variable for each of the two experiments conducted. In experiment one, described in Chapter IV, subjects are asked to acquire and click on fixed-size targets placed at three different radii from a central point. Fitts' Law predicts that the time to complete this task will increase logarithmically with distance from target. The initial research hypothesis tested was that mouse-movement effort as measured in "mickeys" (mouse-pointer pixels traversed on screen) would grow logarithmically as well. It was further hypothesized and tested whether or not eye-movement effort, as measured in mean *saccade amplitude* (distance traveled in a jump between gaze-points of interest), and *gaze-path traversal* (total distance traveled by the eyes throughout the course of a task) would exhibit similar logarithmic growth.

In experiment two, also described in Chapter IV, the Gestalt Law of Proximity is evaluated in terms of effort and time based metrics. Subjects in this experiment interacted with three different interfaces. Interface one ("Form A") places the functionally connected yet visually dissimilar elements of text to be entered and text-entry fields at a maximal distance from each other, interface two ("Form B") reduces the distance between entry fields and entry data, while interface three ("Form C") interleaves text and entry fields. The hypothesis of this experiment is that in accordance with the Law of Proximity, time and effort will exhibit a demonstrable functional dependency on layout proximity and element distance.

The research of this thesis thus sets out to test the following:

1) Time and effort are key dimensions of a given interface's usability.
2) Effort and time are quantifiable with various metrics.
3) Certain quantities of effort and time are intrinsic to achieving usability tasks when interacting with a given interface.
4) The layout and placement characteristics of an interface will influence its underlying usability dimensions, including time and effort. Altering the aspects of an interface will correlate with changes in intrinsic usage effort and time.

5) Furthermore, because interface characteristics are drivers of intrinsic usage time and effort, empirical observations may be used to formulate predictors based on interface characteristics. These formulas would predict the time and effort necessary to achieve usability tasks within particular usage contexts for particular interfaces.

Fitts' Law predicts a logarithmic increase in time to acquire a target as distance to target increases. If the measures of time and effort inherent to interacting with a given software interface also conform to such a growth rate, then the following formulas for time and effort metrics are proposed for acquiring a target set at a fixed distance from a starting point when subjects use a pointing device such as a mouse:

$$tot = A \log_2 (D / W + 1) + a$$

$$msa = E \log_2 (D / W + 1) + e$$

$$gp = F \log_2 (D / W + 1) + f$$

$$mi = G \log_2 (D / W + 1) + g$$

Where *A*, *E*, *F*, *G*, *a*, *e*, *f*, and *g* are constant scaling factors, *tot* is time-on-task, *msa* is mean saccade amplitude over the course of a task, *gp* is gaze-path traversed over the course of a task, *mi* is mickeys (mouse-pixels) traversed over the course of a task, *D* is distance from target, and *W* is target size.

One may further hypothesize the following formulas for interfaces with two principle element groups and varying element group proximities:

$$tot = A \log_2 (| c_1 - c_2 |) + k$$

$$kc + mc = B \log_2 (| c_1 - c_2 |) + l$$

$$msa = E \log_2 (| c_1 - c_2 |) + m$$

$$gp = F \log_2 (| c_1 - c_2 |) + n$$

Where *A*, *B*, *E*, and *F* are constant scaling factors, *tot* is time-on-task, *kc* is key-clicks necessary to achieve a task, *mc* is mouse-clicks necessary to achieve a task, *msa* is mean saccade amplitude over the course of a task, *gp* is gaze-path traversed over the course of a task, *k, l, m,* and *n* are respectively the baseline minimal time, clicks (both keyboard and mouse), saccade amplitude, and gaze-path needed in order to accomplish a given task using a given interface, and c_1 and c_2 are the mean center points (centroids) of the two principle element groupings or areas-of-interest of the interface.

CHAPTER IV

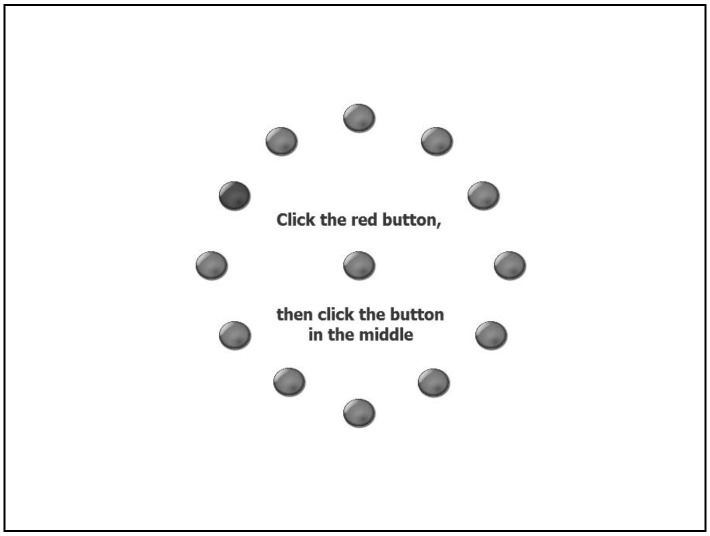
EXPERIMENTAL PROTOCOLS AND RESULTS

In order to lay the groundwork for a broad predictor of user effort intrinsic to a given user interface, two experiments were conducted for this thesis' research. The first experiment is a simple time and effort metric verification of Fitts' law involving acquisition of targets placed at varying radii from a fixed center-point. The second and more complex experiment validates the Law of Proximity using measurements of effort and time. A study methodology is utilized similar to the one employed in Mueller et al. (2009). See Appendix A for technical details on the applications used in testing, and Appendix B for Institutional Review Board approval information.

First Experiment: Fitts' Law

In this thesis' first experiment, subjects interacted with an interface consisting of a simple set of directions, a center-target, and a circle of targets surrounding the center-target at a fixed radius. All targets were identically sized. Subjects were asked to perform the task of alternately clicking on the center-target followed by a randomly selected target in the surrounding circle. After 30 clicks, the radius of the surrounding circle increased by 65 pixels, and the subject repeated the task. See Figures 2, 3 and 4 for illustrations of the test interface.

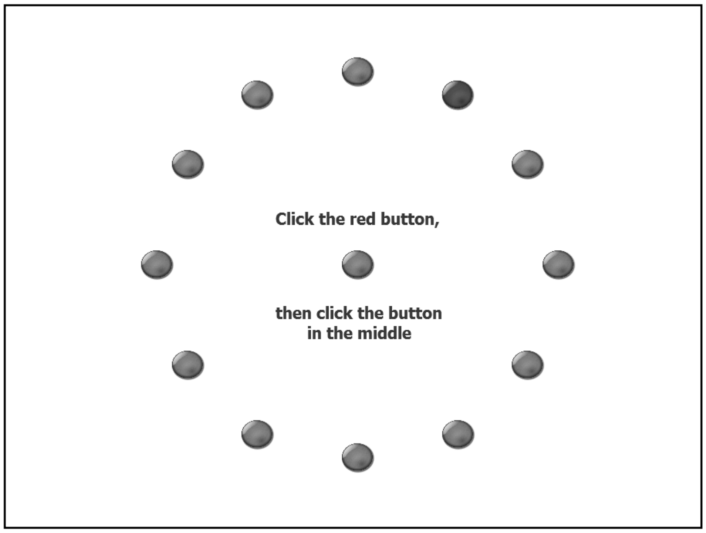
Subjects performed three tasks in total through the course of the experiment: One with a surrounding radius of 195 pixels, one with a radius of 260 pixels, and one with a



Click the red button,

then click the button in the middle

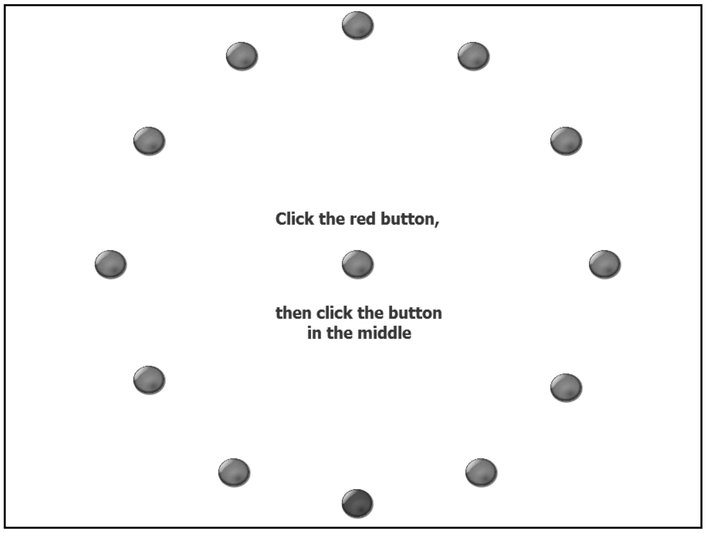
Figure 2. Experiment 1, Task #1.



Click the red button,

then click the button in the middle

Figure 3. Experiment 1, Task #2.



Click the red button,

then click the button in the middle

Figure 4. Experiment 1, Task #3.

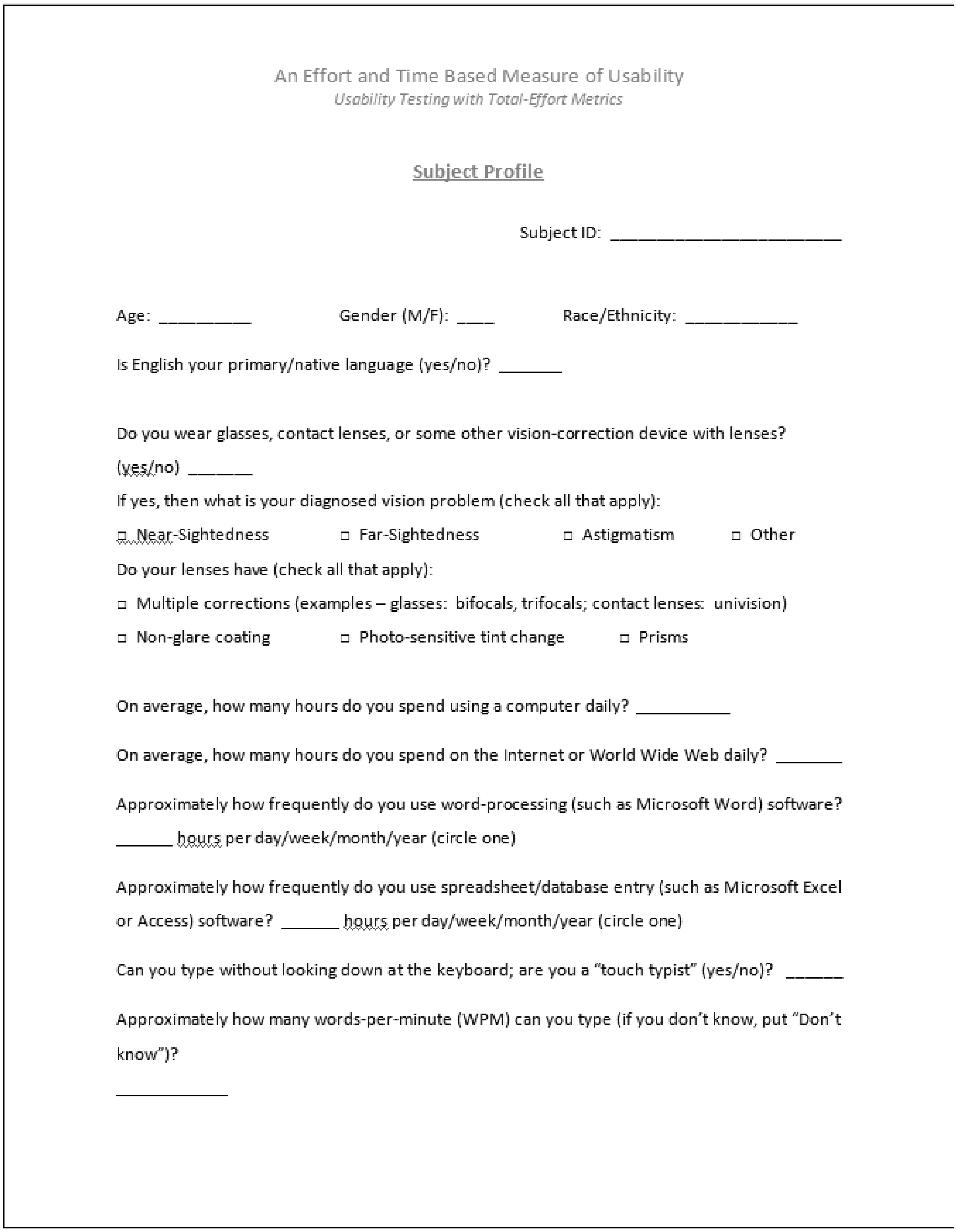
radius of 325 pixels. The test application logged mouse movement, mouse click, and time-on-task data for each subject and set of tasks. A Tobii X120 eye-tracking camera logged eye-movements. The time and effort measures of time-on-task, total “mickeys” (mouse-pixels) traversed, saccade amplitude (point-to-point eye movement) and gaze-path traversal were recorded and analyzed.

Subjects for this study were volunteers recruited from a population of undergraduate and graduate students in the Computer Science/Software Engineering program at Texas State University–San Marcos. At the start of every test session, subjects completed a written questionnaire regarding vision correction, overall computer and specific application usage habits, and demographic information relevant to the study (Figure 5). The eye-tracker was then calibrated, and if the calibration was successful, subjects were advised that they would be completing a series of short exercises guided only by on-screen directions. The test facilitator did not prompt or assist subjects in any way.

Nine subjects in total completed test sessions: Seven men and two women ranging in age from 19 to 31 years old, with an average age of 25.2 years old, standard deviation ±4.3 years. Test subjects as a whole reported weekly computer usage averaging 52.1 ±26 hours and mean weekly Internet/WWW usage of 39.3 ±28.5 hours. Stated word-processor usage averaged 8 ±7.4 hours per week, while database and spreadsheet usage averaged 2.6 +2.7/-2.6 hours per week.

Results and Analysis, Experiment One

As discussed in Chapter 3, Fitts' Law predicts a logarithmic increase in time to acquire a stationary target as either target size decreases or distance-to-target increases. Based on Fitts' Law, this experiment hypothesized logarithmic increases in mean time-



An Effort and Time Based Measure of Usability

Usability Testing with Total-Effort Metrics

Subject Profile

Subject ID: ________________________

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

Is English your primary/native language (yes/no)? _______

Do you wear glasses, contact lenses, or some other vision-correction device with lenses? (yes/no) _______

If yes, then what is your diagnosed vision problem (check all that apply):

□ Near-Sightedness □ Far-Sightedness □ Astigmatism □ Other

Do your lenses have (check all that apply):

□ Multiple corrections (examples – glasses: bifocals, trifocals; contact lenses: univision)

□ Non-glare coating □ Photo-sensitive tint change □ Prisms

On average, how many hours do you spend using a computer daily? __________

On average, how many hours do you spend on the Internet or World Wide Web daily? _______

Approximately how frequently do you use word-processing (such as Microsoft Word) software? ______ hours per day/week/month/year (circle one)

Approximately how frequently do you use spreadsheet/database entry (such as Microsoft Excel or Access) software? ______ hours per day/week/month/year (circle one)

Can you type without looking down at the keyboard; are you a "touch typist" (yes/no)? ______

Approximately how many words-per-minute (WPM) can you type (if you don't know, put "Don't know")?

____

Figure 5. Experiments 1 and 2, Pre-Test Survey.

on-task, gaze path traversal (i.e. visual degrees traversed by the gaze of the subject over the course of a task), saccade amplitude (i.e. visual degrees traversed per point-to-point gaze transition), and mickeys traversed (i.e. mouse-pixels corresponding to pointing-device movements).

With the exception of mean saccade amplitude, actual results did not conform to the research hypotheses. Time-on-task remained essentially flat for all three tasks, perhaps due to the fact that time-on-task measures will tend to decrease as subject familiarity with identical or similar task scenarios increases (Ritter & Schooler, 2002). Two key measures – mickeys traversed and gaze-path traversal – demonstrated linear growth rates as distance-to-target uniformly increased. Mean saccade amplitude did demonstrate a logarithmic growth rate as expected. See figures 6, 7, 8 and 9 for a summary of the research results.

Second Experiment: Law of Proximity

In this thesis' second experiment, subjects were asked to complete simple form fill-in/data-entry tasks. These tasks consisted of copying various pieces of information for fictitious customers displayed on-screen into corresponding textbox fields. The test application logged keystroke, mouse movement, mouse click, and time-on-task data for each subject and set of tasks. A Tobii X120 eye-tracking camera logged eye-movements. The time and effort measures of time-on-task, total keystrokes, correctional keystrokes, saccade amplitude (point-to-point eye movement) and gaze-path traversal were recorded and analyzed.

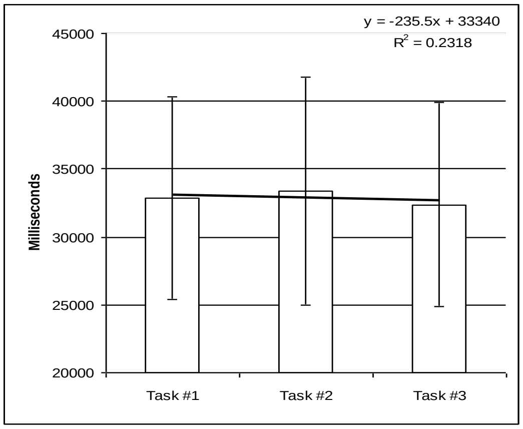


Figure 6. Mean Time-On-Task.

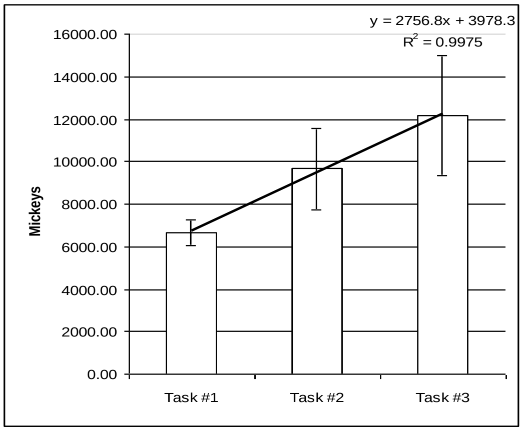


Figure 7. Mean Mickeys Traversed.

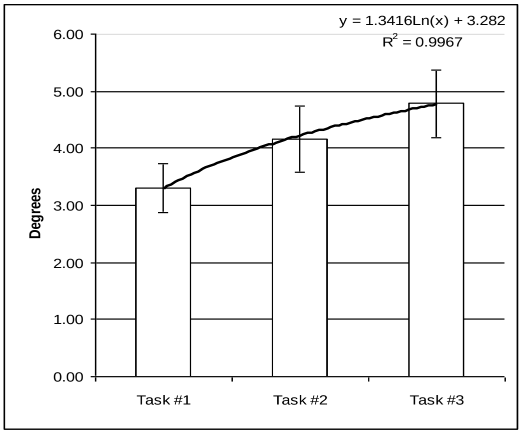


Figure 8. Mean Saccade Amplitude.

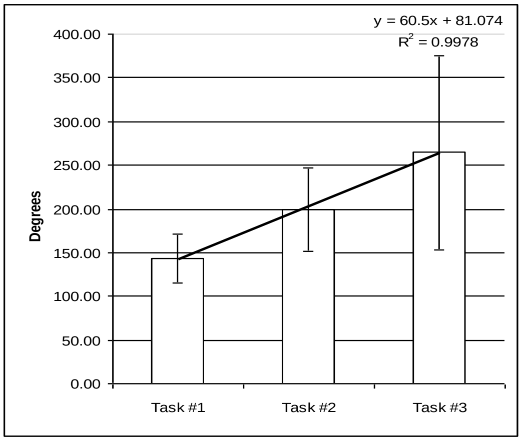
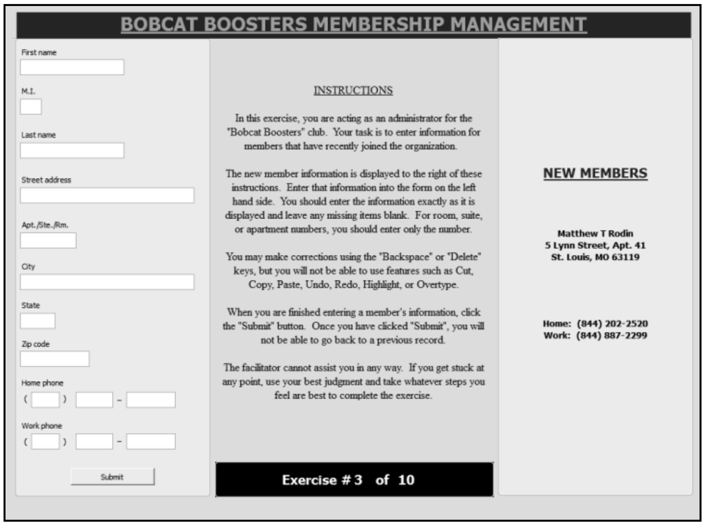


Figure 9. Mean Gaze-Path Traversal.

Every subject interacted with three different interface form factors, each with varying distance between groups of related elements. Elements in the “Form A” interface, as shown in Figure 10, were placed so as to maximize the distance between the display of data to be entered and the actual data-entry fields. “Form B,” shown in Figure 11, placed the data-entry display a short distance away from the data-entry fields. “Form C,” shown in Figure 12, interleaved the display of each data element with its corresponding entry field. The order of form factors presented to Group I was reversed from Group II so that each group served as a control for the other, particularly with regard to factors of fatigue, motivation, and learning.

Subjects for this study were volunteers recruited from a population of undergraduate and graduate students in the Computer Science/Software Engineering program at Texas State University–San Marcos, divided arbitrarily into two groups, Group I and Group II. At the start of every test session, subjects completed a written



INSTRUCTIONS

In this exercise, you are acting as an administrator for the "Bobcat Boosters" club. Your task is to enter information for members that have recently joined the organization.

The new member information is displayed to the right of these instructions. Enter that information into the form on the left hand side. You should enter the information exactly as it is displayed and leave any missing items blank. For room, suite, or apartment numbers, you should enter only the number.

You may make corrections using the "Backspace" or "Delete" keys, but you will not be able to use features such as Cut, Copy, Paste, Undo, Redo, Highlight, or Overtype.

When you are finished entering a member's information, click the "Submit" button. Once you have clicked "Submit", you will not be able to go back to a previous record.

The facilitator cannot assist you in any way. If you get stuck at any point, use your best judgment and take whatever steps you feel are best to complete the exercise.

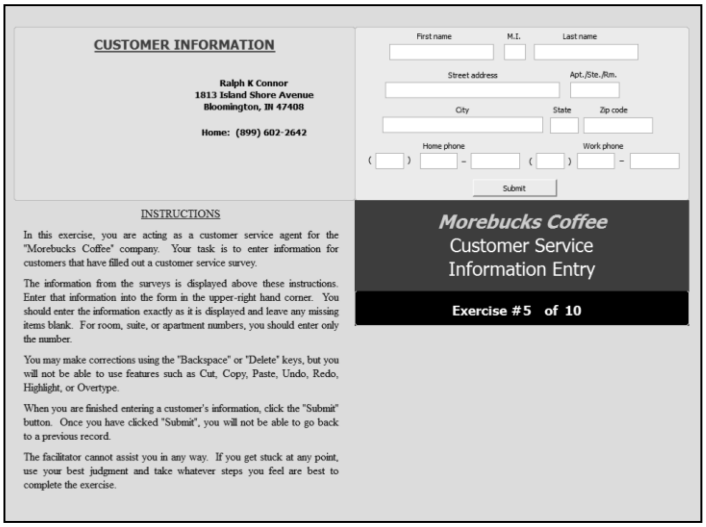
Exercise # 3 of 10

NEW MEMBERS

Matthew T Rodin
5 Lynn Street, Apt. 41
St. Louis, MO 63119

Home: (844) 202-2520
Work: (844) 887-2299

Figure 10. Experiment 2, Form A.



CUSTOMER INFORMATION

Ralph K Connor
1813 Island Shore Avenue
Bloomington, IN 47408

Home: (899) 602-2642

INSTRUCTIONS

In this exercise, you are acting as a customer service agent for the "Morebucks Coffee" company. Your task is to enter information for customers that have filled out a customer service survey.

The information from the surveys is displayed above these instructions. Enter that information into the form in the upper-right hand corner. You should enter the information exactly as it is displayed and leave any missing items blank. For room, suite, or apartment numbers, you should enter only the number.

You may make corrections using the "Backspace" or "Delete" keys, but you will not be able to use features such as Cut, Copy, Paste, Undo, Redo, Highlight, or Overtype.

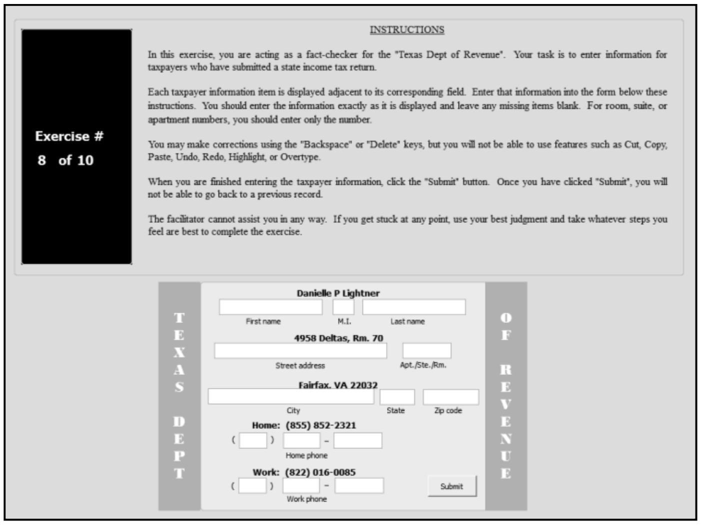
When you are finished entering a customer's information, click the "Submit" button. Once you have clicked "Submit", you will not be able to go back to a previous record.

The facilitator cannot assist you in any way. If you get stuck at any point, use your best judgment and take whatever steps you feel are best to complete the exercise.

Morebucks Coffee
Customer Service
Information Entry

Exercise #5 of 10

Figure 11. Experiment 2, Form B.



INSTRUCTIONS

In this exercise, you are acting as a fact-checker for the "Texas Dept of Revenue". Your task is to enter information for taxpayers who have submitted a state income tax return.

Each taxpayer information item is displayed adjacent to its corresponding field. Enter that information into the form below these instructions. You should enter the information exactly as it is displayed and leave any missing items blank. For room, suite, or apartment numbers, you should enter only the number.

You may make corrections using the "Backspace" or "Delete" keys, but you will not be able to use features such as Cut, Copy, Paste, Undo, Redo, Highlight, or Overtype.

When you are finished entering the taxpayer information, click the "Submit" button. Once you have clicked "Submit", you will not be able to go back to a previous record.

The facilitator cannot assist you in any way. If you get stuck at any point, use your best judgment and take whatever steps you feel are best to complete the exercise.

Exercise #
8 of 10

Danielle P Lightner

4958 Deltas, Rm. 70

Fairfax. VA 22032

Home: (855) 852-2321

Work: (822) 016-0085

Figure 12. Experiment 2, Form C.

questionnaire regarding vision correction, overall computer and specific application usage habits, and demographic information relevant to the study (Figure 5). The eye-tracker was then calibrated, and if the calibration was successful, subjects were advised that they would be completing a series of short exercises guided only by on-screen directions. The test facilitator did not prompt or assist subjects in any way.

After stating that they were ready to proceed, subjects were presented with a form fill-in interface: Form A was presented to Group I, while Form C was presented to Group II. The interface instructed subjects to complete ten exercises on behalf of a fictional organization. It then displayed a collection of fictional data-entry records to be entered by subjects. Once the first ten exercises had been completed, all subjects completed ten additional similar exercises using Form B. Finally, subjects carried out ten last exercises: Group I used Form C and Group II used Form A.

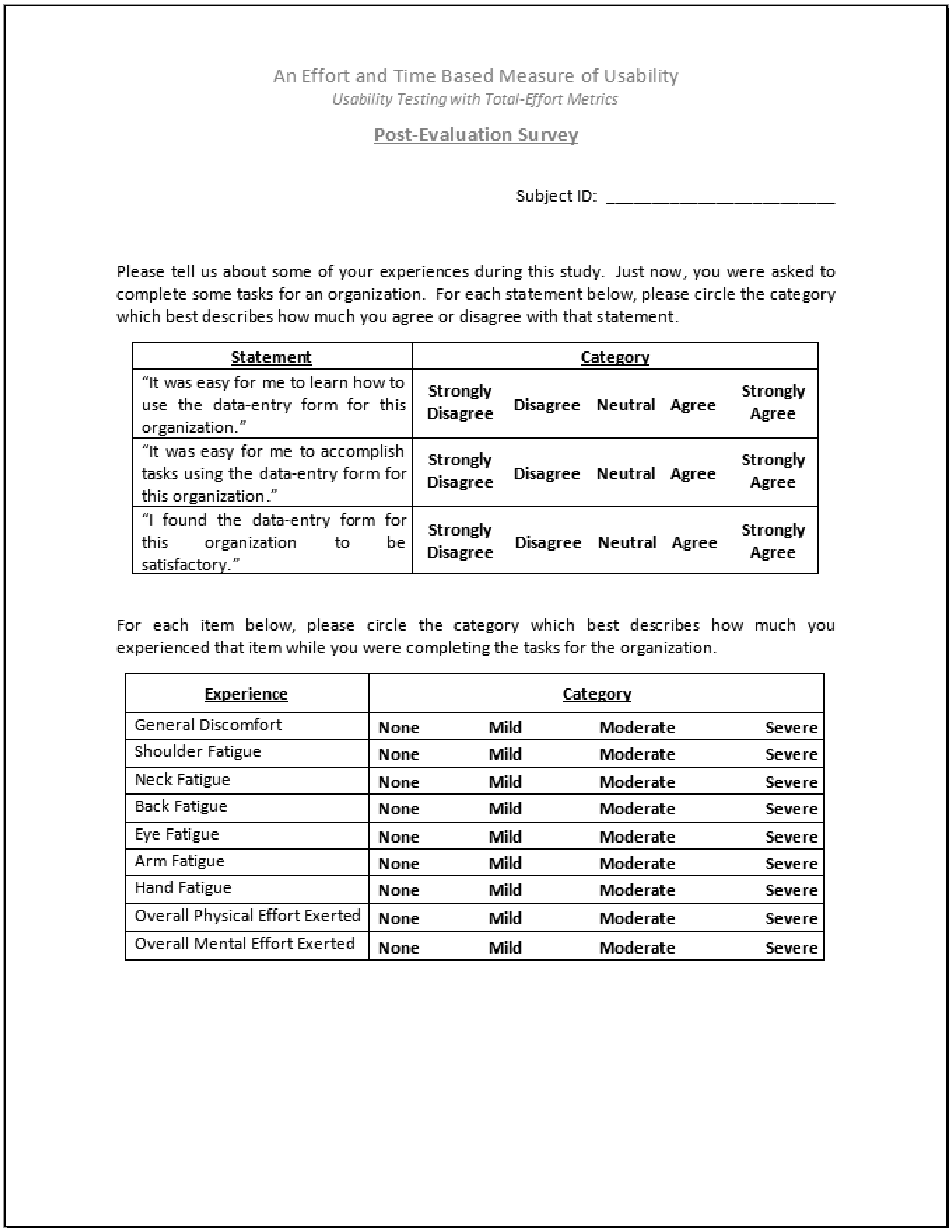
Subjects were given a short break after exercises ten and twenty, after which a brief eye-tracker recalibration procedure was performed. At the conclusion of exercises ten, twenty and thirty, subjects completed a written survey (Figure 13) rating the usability, operability, and satisfaction level for each of three interfaces used throughout the session. On the same form, subjects also rated their experiences of discomfort, fatigue, and effort-exertion while utilizing each interface.

11 subjects in total completed test sessions: Nine men and two women ranging in age from 20 to 29 years old, with an average age of 24.6 years old, standard deviation ±2.8 years. Test subjects as a whole reported weekly computer usage averaging 45.2 ±17.3 hours and mean weekly Internet/WWW usage of 28 ±17.1 hours. Stated word-processor usage averaged 11.8 +12.3/-11.8 hours per week, while database and spreadsheet usage had a mean of 5.1 +10.1/-5.1 hours per week. Eight subjects indicated that they are "touch typists" (i.e. able to type without looking down at the keyboard). Four subjects reported having learned English as a secondary language.

Results and Analysis, Experiment Two

Looked at in isolation, each category of data captured by this study – qualitative, time-on-task, keystroke count, correctional keystrokes, and eye movements – provides useful but limited insight into the usability aspects of the interfaces tested. The captured timing data, in combination with the qualitative information gathered, indicate that Form

A has some sort of efficiency issue while Form C is superior in terms of efficiency-of-use. Keystroke data further indicate that Form A inhibits usage effectiveness whereas Form C allows tasks to be accomplished more effectively. Keystroke and time-on-task



An Effort and Time Based Measure of Usability

Usability Testing with Total-Effort Metrics

Post-Evaluation Survey

Subject ID: ________________________

Please tell us about some of your experiences during this study. Just now, you were asked to complete some tasks for an organization. For each statement below, please circle the category which best describes how much you agree or disagree with that statement.

Statement	Category				
"It was easy for me to learn how to use the data-entry form for this organization."	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
"It was easy for me to accomplish tasks using the data-entry form for this organization."	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
"I found the data-entry form for this organization to be satisfactory."	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**

For each item below, please circle the category which best describes how much you experienced that item while you were completing the tasks for the organization.

Experience	Category			
General Discomfort	**None**	**Mild**	**Moderate**	**Severe**
Shoulder Fatigue	**None**	**Mild**	**Moderate**	**Severe**
Neck Fatigue	**None**	**Mild**	**Moderate**	**Severe**
Back Fatigue	**None**	**Mild**	**Moderate**	**Severe**
Eye Fatigue	**None**	**Mild**	**Moderate**	**Severe**
Arm Fatigue	**None**	**Mild**	**Moderate**	**Severe**
Hand Fatigue	**None**	**Mild**	**Moderate**	**Severe**
Overall Physical Effort Exerted	**None**	**Mild**	**Moderate**	**Severe**
Overall Mental Effort Exerted	**None**	**Mild**	**Moderate**	**Severe**

Figure 13. Experiment 2, Post-Test Survey.

(i.e. time required by subject to complete a given task) data indicate the presence of a usability issue, but provide no indication as to precisely what the nature of the issue is. On the other hand, when the eye-tracker data is added into the picture, an explanation for the underlying usability issue of Form A becomes clearer.

On balance, the qualitative data alone do not provide a clear picture as to which of the three evaluated interfaces are most usable, much less why one is more or less usable than the other. Qualitative ratings for the three interfaces mostly conformed to the research hypothesis that user perceptions of usability, learnability and satisfaction would increase as element layout proximity decreased. Subjects rated Form C, the form with interleaved data to be entered and data entry fields, as being the most usable and satisfying to use. Form A, the form which maximized the distance between data to be entered and data entry fields, was rated as being the least usable and satisfying to use. Subject ratings of learnability did not conform to expectations; subjects rated Form B, the intermediate-distance form, as being the most learnable.

It was expected that subjects would rate Form A as involving the most discomfort and exertion to use, using Form B would be rated more comfortable and less effort-intensive to use than Form A, and Form C would be rated with a perception of the least amount of discomfort and exertion. Subjects did rate Form A as inducing the most discomfort, but also involving the least amount of physical exertion. Form B was rated as involving the least amount of mental exertion.

Time-on-task data, i.e. “stopwatch” data, are a staple of conventional usability evaluation methods (Dumas & Redish, 1993; Nielsen, 1993). The information captured regarding time-on-task provides a somewhat better indicator of which interfaces exhibit usability issues. A time-on-task graph for Groups I and II combined is provided in Figure 14. Time-on-task measures, it should be noted, tend to exhibit a decreasing slope as subject familiarity with identical or similar task scenarios increases (Ritter & Schooler, 2002). Thus a null hypothesis for comparing the time-on-task results for Groups I and II is that time-on-task for each group will decrease at a uniform rate.

As shown in Figures 15 and 16, the null hypothesis did not hold. The times-on-task for Group I decreased at a sharper rate than those in Group II. This is as expected given that the three interfaces which Group I interacted with were presented in decreasing order of element layout proximity, whereas the three interfaces which Group II interacted with were presented in increasing order of element layout proximity. The time-on-task data imply that task efficiency increases as interface element proximity decreases.

Keystroke data for Groups I and II combined is graphed in Figure 17. The segregated keyboard logging data shown in Figures 18 and 19 indicate that task effectiveness, as measured by keystrokes necessary to accomplish a task, also tends to increase as interface element proximity decreases. As with time-on-task, when interfaces are presented in decreasing order of element closeness, task completion keystrokes decrease as expected. When interfaces are presented in increasing order of element closeness, the same flattening of the curve is observed as was seen with the time-on-task charts. The curve-flattening illustrates that an increase in effectiveness due to learning over time is in effect colliding with the ineffectiveness burden imposed by the wide spacing between the interface’s elements.

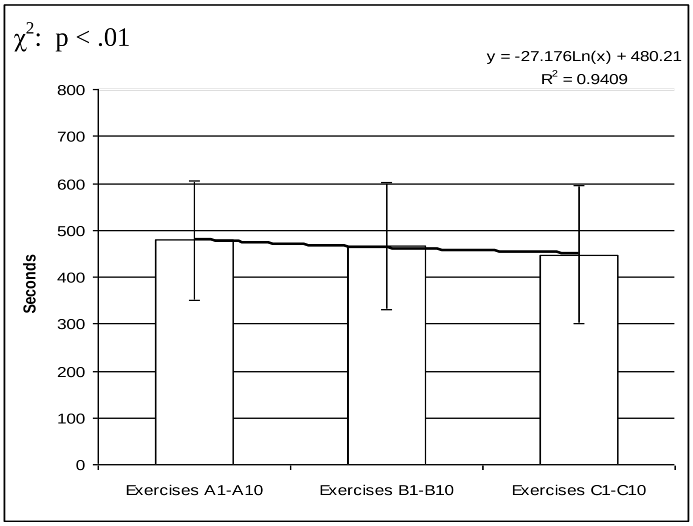


Figure 14. Mean Time-On-Task-Set, By Form Factor.

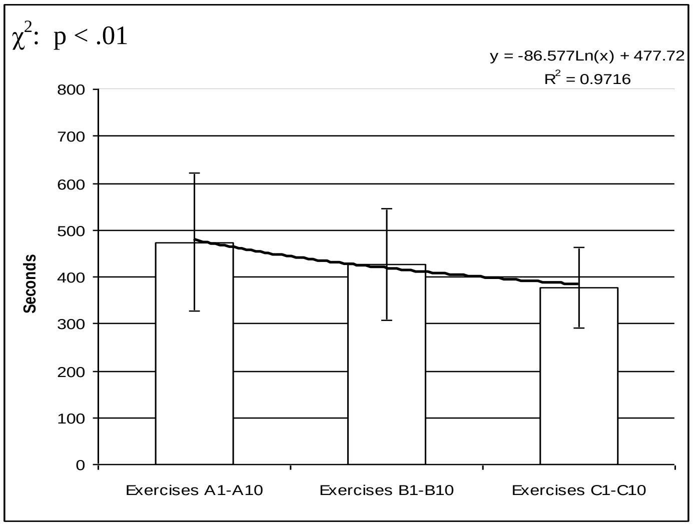


Figure 15. Mean Time-On-Task-Set, Group I, By Presentation Order.

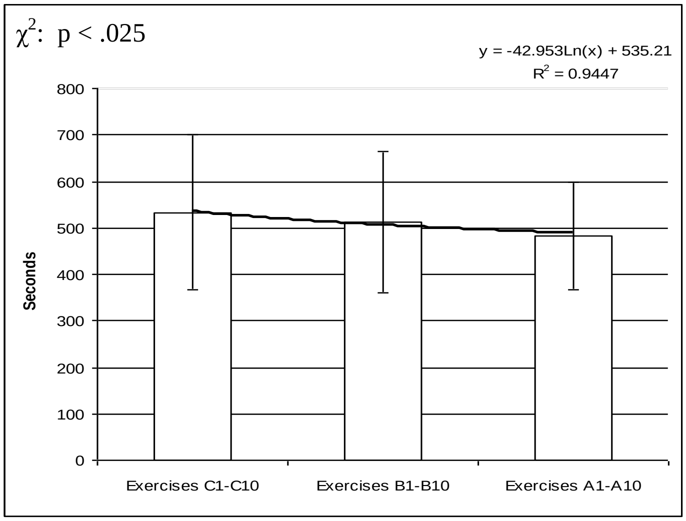


Figure 16. Mean Time-On-Task-Set, Group II, By Presentation Order.

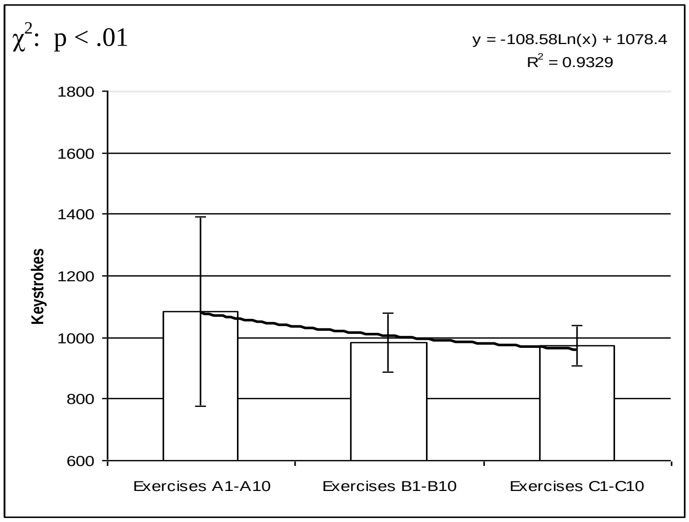


Figure 17. Mean Keystrokes, By Form Factor.

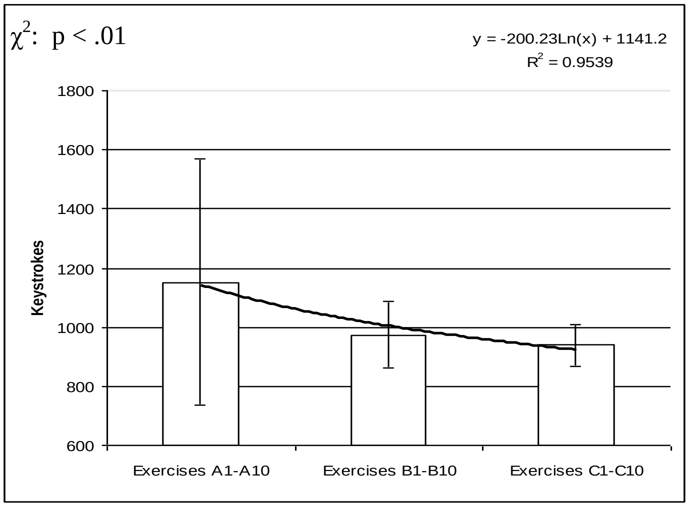


Figure 18. Mean Keystrokes, Group I, By Presentation Order.

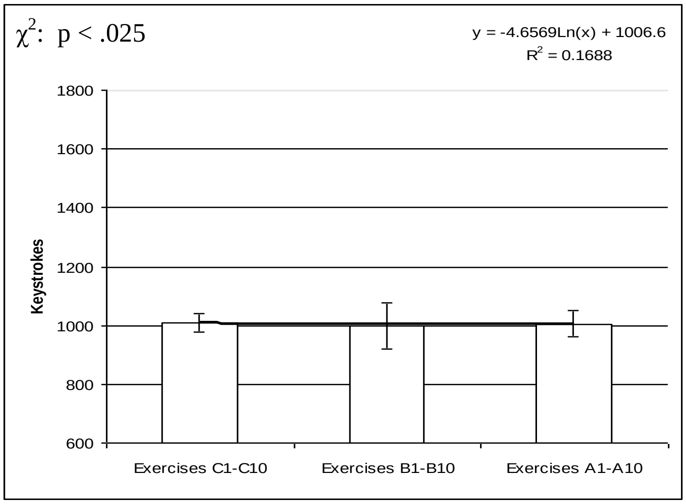


Figure 19. Mean Keystrokes, Group II, By Presentation Order.

Similar trends are seen in data for the total number of correction-keystrokes, i.e. the number of keypresses necessary to undo a mistake. Correction-keystrokes is defined as 2 × "Backspace" keypresses + 2 × "Delete" keypresses + any arrow-key presses (note that the experiment disabled cut-and-paste and highlight/delete input features). Figure 20 depicts correction-keystrokes for Groups I and II combined, and Figures 21 and 22 show the data separated out by group.

The keystroke and time-on-task data indicate fairly definitively that there is some sort of underlying usability issue with Form A which is inhibiting user effectiveness and efficiency. The eye-tracker data confirm this finding and furthermore show the underlying cause of the usability issues. As can be seen in Figures 23 and 26, there is a marked difference in the eye-movement distances required for Forms A, B and C. An increase in related element proximity strongly correlates with shorter gaze-path traversal as well as shorter jumps between points-of-interest within the interface. In the case of eye movements, order of presentation did not induce any "learning effect" i.e. decrease in required eye movement as time progressed. Figures 24, 25, 27 and 28 clearly show that change in eye-movement effort is similar regardless of whether the forms are presented in decreasing or increasing order of element proximity.

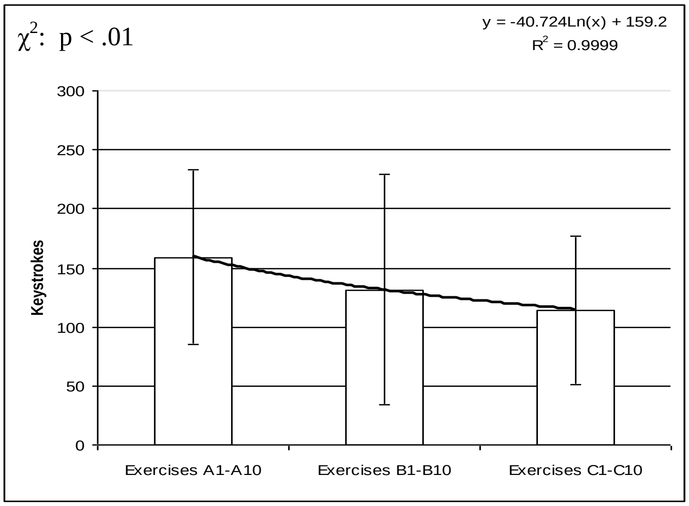


Figure 20. Mean Correction-Keystrokes, By Form Factor.

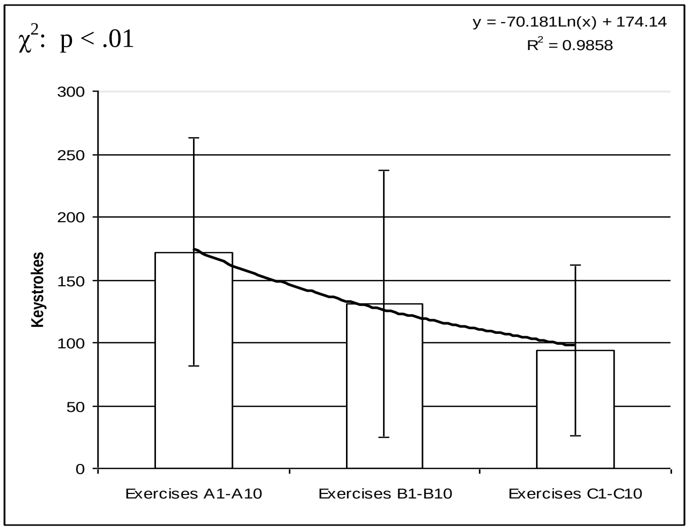


Figure 21. Mean Correction-Keystrokes, Group I, By Presentation Order.

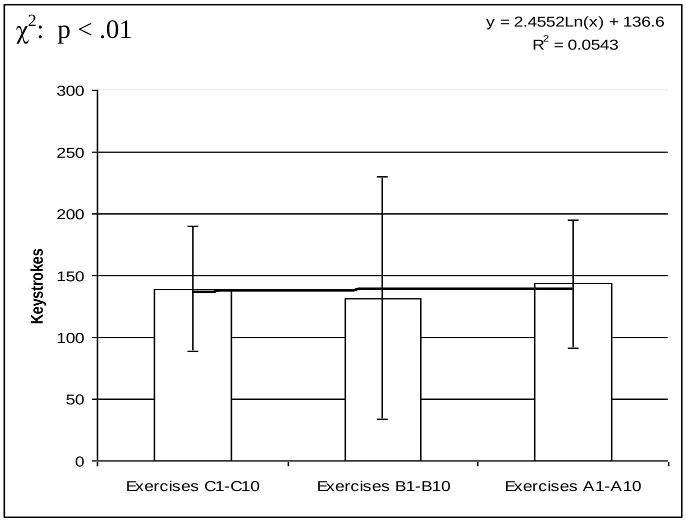


Figure 22. Mean Correction-Keystrokes, Group II, By Presentation Order.

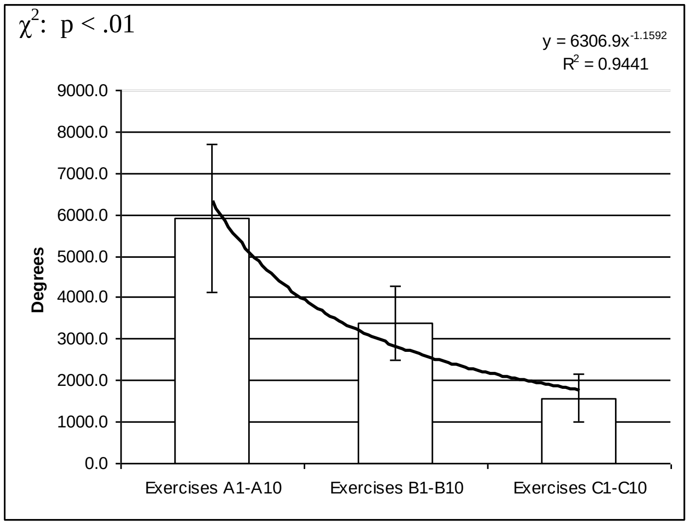


Figure 23. Mean Gaze-Path Traversal, By Form Factor.

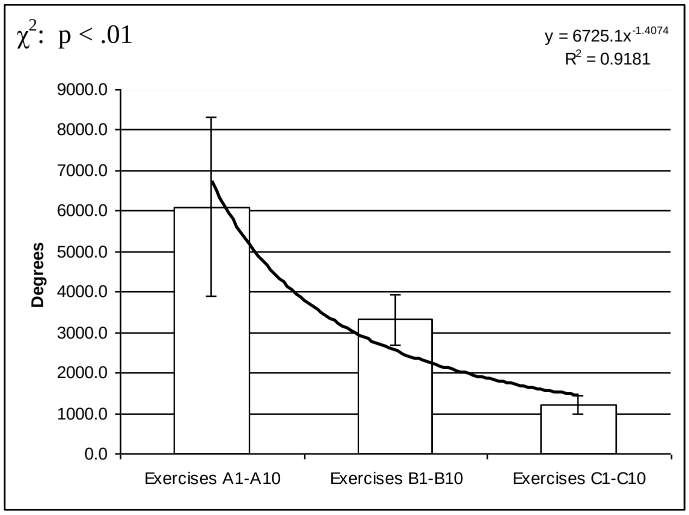


Figure 24. Mean Gaze-Path Traversal,
Group I, By Presentation Order.

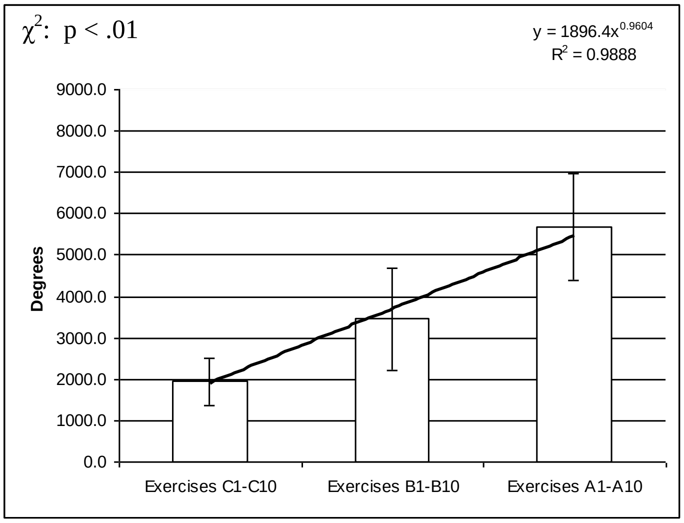


Figure 25. Mean Gaze-Path Traversal,
Group II, By Presentation Order.

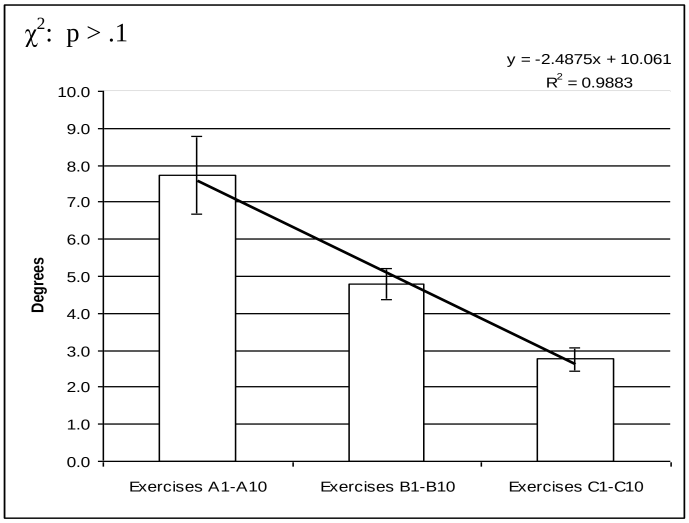


Figure 26. Mean Saccade Amplitude, By Form Factor.

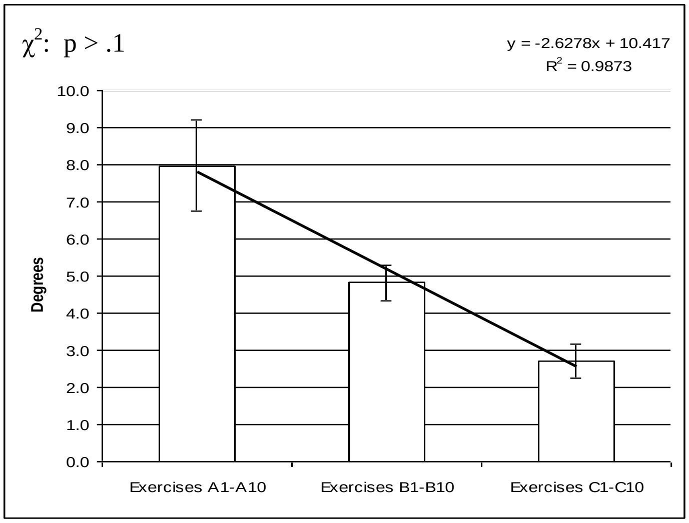


Figure 27. Mean Saccade Amplitude, Group I, By Presentation Order.

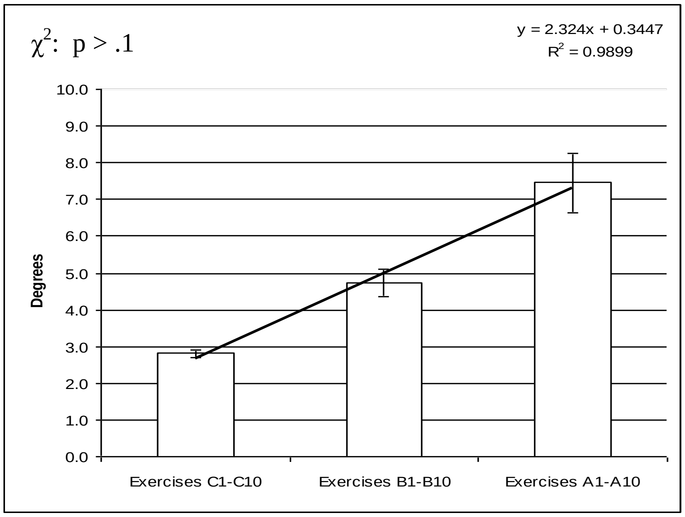


Figure 28. Mean Saccade Amplitude, Group II, By Presentation Order.

CHAPTER V

PREDICTORS OF EFFORT INTRINSIC TO A GIVEN INTERFACE

The brain carries out cognitive processing in a manner which can be analogized to a massively-parallel machine (Miyata & Norman, 1986). Physical activities, by necessity, are mostly carried out in a serial fashion. Thus it is important to keep in mind that the component metrics of an effort-based framework are in effect a serialization of parallel processing activities. It is also critical to note that for a given subject, effort measurements recorded of repetitions of identical or similar tasks will demonstrate a "learning effect" or "learning curve". In other words, any set of homologous tasks performed in series will become easier over time to accomplish (Ritter & Schooler, 2002).

Accurately subtracting the effects of the "learning curve" from effort-based metrics is a topic of active research at this time. Nonetheless, it is believed that the data presented in Chapter IV demonstrate that other factors besides learning can and will have effects on measurements of user effort. The specific weighting factors needed in order to formulate a singular "effort score" are not known at this time, but it is still of value to examine each effort component individually.

Fitts' Law as a Predictor of User Effort

Chapter IV illustrated that in this thesis' first experiment, time-on-task did not increase logarithmically as predicted. Time-on-task remained essentially unchanged for each of the three tasks presented to subjects. It seems unrealistic to predict that this trend would hold as target distance continued to increase, therefore this thesis does not formulate a predictor of a time-on-task growth trend as distance from target grows larger.

The results from the effort measures of mickeys traversed and gaze-path traversal demonstrate a definite linear growth trend as distance from target uniformly increases, while saccade amplitude does appear to conform to a logarithmic growth rate. Further research involving varying task lengths and subtask sizes is called for in order to verify the empirical conclusions reached by this thesis. Enough data has now been gathered, however, to create the following formulations of effort predictors for target acquisition tasks:

$$msa = E \log_2 (D / W + 1) + e$$

$$gp = F (D / W) + f$$

$$mi = G (D / W) + g$$

Where *E*, *F*, *G*, *e*, *f*, and *g* are constant scaling factors, *msa* is mean saccade amplitude over the course of a task, *gp* is gaze-path traversed over the course of a task, *mi* is mickeys (mouse-pixels) traversed over the course of a task, *D* is distance from target, and *W* is target size.

Table 1 compares hypothesized logarithmic growth, predicted linear growth, and observed values for experiment 1. See Chapter III for the formulas used for the logarithmic growth predictions in Table 1.

Table 1. Predicted vs. Observed Metrics of User Effort, Experiment 1.[1]

Metric	Distance To Target	Logarithmic Predicted Mean Value	Linear Predicted Mean Value	Observed Mean Value
Mean saccade amplitude	195 px	3.60[2]	n/a	3.3
	260 px	4.22[2]	n/a	4.16
	325 px	4.72[2]	n/a	4.79
Gaze-path traversal	195 px	144.02[3]	148.97[4]	143.22
	260 px	168.63[3]	198.62[4]	198.78
	325 px	188.90[3]	248.28[4]	264.22
Mickeys traversed	195 px	6300.88[5]	6827.73[6]	6654.85
	260 px	7377.39[5]	9103.64[6]	9652.47
	325 px	8264.24[5]	11379.55[6]	12168.41

[1] Target size $W = \pi(5^2) \approx 78.54$
[2] $E = 2$, $e = 0$
[3] $F = 80$, $f = 0$
[4] $F = 60$, $f = 0$
[5] $G = 3500$, $g = 0$
[6] $G = 2750$, $g = 0$

A Predictor of Effort-Based Metrics Based on Layout Proximity

As mentioned in Chapter III, the “Law of Proximity” from Gestalt psychology is an important usability design notion which is conceptually related to Fitts’ Law. Experiment two, described in Chapter IV, is a validation of the Law of Proximity through use of effort-based metrics. Table 2 compares the hypothesized effort-metrics growth rates to the empirical results for Group I. The constants for the minimum values used in Table 2 are the minimum observed values from Experiment 2, with the exception of correction-keystrokes, where the minimum is zero. See Chapter III for the formulas used for the hypothesized predictions in Table 2.

Table 2. Predicted vs. Observed Metrics of User Effort, Experiment 2.

Metric	**Distance Between Centroids**	**Predicted Mean Value**	**Observed Mean Value**	**Standard Deviation**
Time-on-task	10 px	340.446[1]	376.684	±85.2
	350 px	453.291[1]	427.09	±118.681
	650 px	472.939[1]	474.255	±146.347
Keystrokes	10 px	937[2]	939	±71
	350 px	1091[2]	975	±113
	650 px	1117[2]	1152	±418
Correctional-Keystrokes	10 px	66[3]	94	±68
	350 px	169[3]	131	±107
	650 px	187[3]	172	±91
Mean Saccade Amplitude	10 px	3.73[4]	2.7	±0.4
	350 px	6.04[4]	4.8	±0.5
	650 px	6.44[4]	8	±1.2
Gaze-Path Traversal	10 px	1186.2[5]	1211	±244.5
	350 px	1699.1[5]	3309.7	±617.3
	650 px	1788.4[5]	6095	±2216.3

[1] $A = 22$, $k = 267.364$
[2] $B = 30$, $l = 837$
[3] $B = 20$, $l = 0$
[4] $E = 0.45$, $m = 2.24$
[5] $F = 100$, $n = 854$

Based on the results obtained, for interfaces with two separate but related areas-of-interest, it appears that Fitts' Law is a reliable model for time-on-task and keystroke measures. This does not appear to be the case for the eye-movement measures of mean saccade amplitude and gaze-path traversal, however, and alternative predictive formulas are proposed below. On the basis of the data observed in experiment two, the following predictors are proposed which relate time and effort measures, layout characteristics and task achievement for tasks of a sufficient length:

$$tot = A \log_2 (| c_1 - c_2 |) + k$$

Where *tot* is time-on-task, *A* is a constant scaling factor, c_1 and c_2 are the respective centroids (center points as measured by horizontal and vertical mean calculation) of two related element groupings, and k is a constant representing the minimum achievable time-on-task.

$$kc + mc = B \log_2 (| c_1 - c_2 |) + l$$

Where kc is key-clicks necessary for task accomplishment, mc is mouse-clicks necessary for task accomplishment, B is a constant scaling factor, c_1 and c_2 are the respective centroids of two related element groupings, and l is a constant representing the minimum number of key-clicks and mouse-clicks necessary to accomplish a task.

$$corr = B \log_2 (| c_1 - c_2 |) + l$$

Where $corr$ is actions necessary to correct input errors for a given task, C is a constant scaling factor, and c_1 and c_2 are the respective centroids of two related element groupings. Note that this is an unexpected finding: Corrective keystrokes is more a measure of user efficiency than of effectiveness in accomplishing tasks, yet based on the data gathered this metric appears to conform to a similar pattern as does overall keystroke rate.

The following non-logarithmic growth formulas are proposed as predictors of eye-movement metrics:

$$msa = E (| c_1 - c_2 |) + m$$

Where msa is mean saccade amplitude over the course of accomplishing a task, E is a constant scaling factor, c_1 and c_2 are the respective centroids of two related element groupings, and m is a constant representing the minimum saccade amplitude necessary to accomplish a task.

$$gp = F \sqrt[2]{(| c_1 - c_2 |)^3} + n$$

Where *gp* is gaze path traversal over the course of accomplishing a task, *E* is a constant scaling factor, c_1 and c_2 are the respective centroids of two related element groupings, and *n* is a constant representing the minimum gaze-path traversal necessary to accomplish a task. As shown in Table 3, these two proposed formulas for predicting eye-movement metrics yield a tighter conformance to empirically observed values than the Fitts' Law based logarithmic growth formulas.

Note that the above equations are not universally applicable. Usability must be evaluated in terms of certain users with comparable aptitude using particular applications in a given usage environment (ISO, 2001). Therefore, the formulas given above should only be applied to a comparable set of users using a given application.

Table 3. Fitts' Law Based (Logarithmic) vs. Alternative Predictions of User Effort, Experiment 2.

Metric	Distance Between Centroids	Logarithmic Predicted Mean Value	Alternative Predicted Mean Value	Observed Mean Value
Mean Saccade Amplitude	10 px	3.73[1]	2.34[3]	2.7
	350 px	6.04[1]	5.74[3]	4.8
	650 px	6.44[1]	8.74[3]	8
Gaze-Path Traversal (*F* = 100, *n* = 854)	10 px	1186.2[2]	863.5[4]	1211
	350 px	1699.1[2]	2818.4[4]	3309.7
	650 px	1788.4[2]	5825.5[4]	6095

[1] $E = 0.45$, $m = 2.24$

[2] $F = 100$, $n = 854$

[3] $E = 0.01$, $m = 2.24$

[4] $F = 0.3$, $n = 854$

CHAPTER VI

CONCLUSION AND FUTURE RESEARCH

An effort-based metric methodology yields elegant, quantitatively-expressed insights into the usability characteristics of software systems. It is an analysis framework that integrates commonplace methods with measurements that are not yet universally utilized, but have high applicability to usability testing. The approach is easy to employ, less expensive to implement than other techniques, and tolerant of noise factors like physical environment anomalies, changes in user motivation, or unexpected issues that arise during the course of testing. As a standalone tool or a complement to a traditional evaluation, it can be used to validate existing design guidelines or discover unanticipated areas of concern.

This thesis serves as a proof-of-concept for measurement of user interactions with software interfaces utilizing metrics with a basis in effort and time. It is also a preliminary step toward a holistic predictor of the effort and time intrinsic to using various interfaces. The methods employed in this work draw from traditional usability testing, but also innovate in novel ways.

“Stopwatch” measures and qualitative evaluation are valid and important tools. However, this thesis demonstrates that there are additional quantitative dimensions that can and should be explored in order to realize a full picture of a product’s usability. It is the hope of the author that a testing framework grounded in effort and time based measures of usability will help bridge the gap between usability evaluation and conventional software testing.

Due to time and resource constraints, the experiments conducted for this thesis were limited in terms of number of subjects and number of different tasks tested. There are numerous variations on the experiments which could be created and executed in order to derive greater breadth and depth of data. Such experiments could be used to validate or revise the suggested predictors of time and effort metrics.

This thesis has concerned itself with the relationship between layout characteristics and measurements of user effort. Layout is but one of several interface design concerns. Best practices have been proposed for several other areas of design, including widget characteristics, element interaction, functional sequencing, dialog phrasing, online or inline documentation, colors, fonts, frame sizing and placement, and numerous additional items (Fowler, 1998; Nielsen, 1993). It would be valuable to conduct effort-based metric verifications of standards and guidelines for each of these areas. A time and effort-based measurement validation of the five Gestalt laws not tested in this thesis (Moore & Fitz, 1993) (see Chapter III), particularly the Laws of Symmetry and of Similarity, would also be useful.

As discussed in Chapter V, further research is required to determine a set or sets of weighting factors which would properly scale and interrelate various effort-based metrics. In addition, as discussed in Chapter III, effort-based measurements are not limited to the ones discussed in this thesis. It would be of value to research additional effort-based metrics. Also of use would be research into additional sub-metrics which might be derived from the base set of metrics that have been discussed in this thesis.

The pointing device used in the research for this thesis was a typical two-button optical wheel mouse. Davis (2009) has suggested a comparative study involving multiple pointing devices such as pressure tablets, smaller sized “mini” optical mouse devices, or optical trackball devices. More exotic input methods such as a pure gaze-based interface using an eye-tracker (Komogortsev et al., 2009a) might be evaluated using effort-based metrics as well.

In the short term, it would be valuable to conduct an examination of the variations in effort necessary to conduct tasks using horizontal mouse movement versus vertical mouse movement. As Davis (2009) points out, non-horizontal mouse movements involve differing musculature and greater physical effort than purely horizontal mouse movements. A repetition of experiment one involving target acquisitions over a horizontal horizon, a vertical horizon and a diagonal horizon would be of value.

Over the longer term, it would be useful to gain greater insight into the biomechanical bases of physical effort and the cognitive components of mental effort. This would likely require additional instrumentation, but care must be taken to ensure that any physiometric device used in testing is not overly invasive or discomfiting to subjects. Eventually, experiments involving instruments such as a functional magnetic resonance imaging might be conducted which draw upon the methodologies of neuroeconomics (Glimcher, Camerer, Fehr, & Poldrack, 2008).

APPENDIX A

TESTING APPLICATION TECHNICAL DETAILS

Applications used in testing for Experiment 1 and Experiment 2 of this thesis were designed, programmed and tested by the author using the programming language C++ in conjunction with the non-commercial, open source software development toolkit Qt, which is copyright © 2008-2009 Nokia Corporation and/or its subsidiaries. Nokia, Qt and their respective logos are trademarks of Nokia Corporation in Finland and/or other countries worldwide. Usage of Qt was licensed under the GNU General Public License Version 3, dated June 29, 2007.

Certain images contained within the applications used in testing for Experiments 1 and 2 are believed by the author to be within the public domain and were downloaded from various "clip art" sites on the World Wide Web.

The technical computing suite MATLAB 2009, commercially/academically licensed to Texas State University-San Marcos, was used for analysis of eye-movement data.

APPENDIX B

INSTITUTIONAL REVIEW BOARD NOTICE

All human subject testing conducted during the work of this thesis was carried out with the approval of the Texas State University-San Marcos Institutional Review Board (IRB), IRB approval #2008-70391. Before the commencement of any testing, all test subjects signed an informed consent form which advised subjects of the minimal risks pertinent to the study, data to be collected from subjects, usage and anonymousness of data collected, and contact information for the researcher conducting the study (i.e. the author of this thesis), the research supervisor, the IRB chair, and the IRB OSP administrator.

BIBLIOGRAPHY

Bertoa, M. F., Troya, J. M., & Vallecillo, A. (2006). Measuring the usability of software components. *Journal of Systems and Software* 79(3), 427-439. Elsevier, ISSN 0164-1212.

Bevan, N. (2001). International standards for HCI and usability. *International Journal of Human-Computer Studies* 55(4), 533-552. Oxford, UK: Academic Press.

Brinkman, W., Haakma, R., & Bouwhuis, D. G. (2007). Towards an empirical method of efficiency testing of system parts: A methodological study. *Interacting with Computers* 19(3), 342-356. Elsevier Science Direct.

Davis, W. (2009). Unpublished manuscript. Texas State University–San Marcos.

DeMarco, T. (1982). *Controlling software projects: Management, measurement & estimation*. Englewood Cliffs, NJ: Yourdon Press.

Dieli, M. (1988). A problem-solving approach to usability test planning. In *Professional Communications Conference, 1988. On the Edge: A Pacific Rim Conference on Professional Technical Communication, International*. 265-267. IPCC ’88 Conference Record.

Dumas, J. S., & Redish, J. (1999). *A practical guide to usability testing*. Exeter, England: Intellect Books.

Fenton, N. E., & Pfleeger, S. L. (1997). *Software metrics: A rigorous and practical approach*. Boston: PWS Pub.

Fowler, S. (1998). *GUI Design Handbook*. New York, NY: McGraw-Hill.

Glimcher, P., Camerer, C. F., Fehr, E., & Poldrack, R. (eds.) (2008). *Neuroeconomics: Decision Making and the Brain*. London, UK: Academic Press.

Granollers, T. & Lorés, J. (2006). Usability Effort: A new concept to measure the usability of an interactive system based on UCD. In *HCI Related Papers of Interacción 2004, R. Navarro-Prieto and J.L. Vidal (eds.)*. 103-117. The Netherlands: Springer.

Hassenzahl, M., & Sandweg, N. (2004). From mental effort to perceived usability: transforming experiences into summary assessments. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems* (Vienna, Austria, April 24 - 29, 2004), 1283-1286. CHI '04. New York, NY: ACM.

Hax, A. C., & Majluf, N. S. (1982). Competitive cost dynamics: The experience curve. *Interfaces* 12(5), 50-61. Hanover, MD: INFORMS.

Hornbæk, K., & Law, E. L. (2007). Meta-analysis of correlations among usability measures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, April 28-May 3, 2007), 617-626. CHI '07. New York, NY: ACM.

Institute of Electrical and Electronics Engineers (2008). *IEEE Standard for Software and System Test Documentation*. IEEE 829-2008. New York, NY: Institute of Electrical and Electronics Engineers.

International Standards Organization (1998). *Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability*. ISO 9241-11:1998(E). Geneva, Switzerland: International Standards Organization.

International Standards Organization (2001). *Software engineering — product quality — part 1: Quality model*. ISO/IEC 9126-1:2001(E). Geneva, Switzerland: International Standards Organization.

Jokela, T., Iivari, N., Matero, J., & Karukka, M. (2003). The standard of user-centered design and the standard definition of usability: Analyzing ISO 13407 against ISO 9241-11. In *Proceedings of the Latin American Conference on Human-Computer interaction* (Rio de Janeiro, Brazil, August 17 - 20, 2003). CLIHC '03, 46, 53-60. New York, NY: ACM.

Jones, C. (1997). *Software Quality – Analysis and Guidelines for Success*. Boston: International Thomson Computer Press.

Kaner, C., Falk, J., & Nguyen, H. Q. (1999). *Testing Computer Software*.

Komogortsev, O. V., Mueller, C. J., Tamir, D., & Feldman, L. (2009). An effort-based model of software usability. In *Proceedings of the International Conference on Software Engineering Theory and Practice* (Orlando, Florida, July 13-16, 2009). SETP-09.

Komogortsev, O. V. (2009a). Unpublished research. Retrieved on June 2, 2009 from http://www.cs.txstate.edu/~ok11/igaze.html.

Miyata, Y., & Norman, D. A. (1986). Psychological issues in support of multiple activities. In *User-Centered System Design*, D. A. Norman & S. W. Draper, eds. Hillsdale, NJ: Erlbaum.

Moore, P., & Fitz, C. (1993). Gestalt theory and instructional design. *Journal of Technical Writing and Communication* 23(2), 137-157. Baywood, ISSN 0047-2816.

Mueller, C. J. (2009). Unpublished manuscript. Texas State University–San Marcos.

Mueller, C. J., Tamir, D., Komogortsev, O. V., & Feldman, L. (2009). An economical approach to usability testing. In *Proceedings of the 33rd Annual IEE International Computer Software and Applications Conference* (Seattle, Washington, July 20-24, 2009). COMPSAC ’09.

Myers, G. (2004). *The Art of Software Testing*. Hoboken, NJ: John Wiley & Sons, Inc.

Nielsen, J. (1993). *Usability engineering*. Boston: Academic Press.

Pressman, R. S. (2005). *Software engineering: A practitioner's approach*. Boston: McGraw-Hill.

Ritter, F. E., & Schooler, L. J. (2002). The learning curve. In *International Encyclopedia of the Social and Behavioral Sciences*. 8602-8605. Amsterdam: Pergamon.

Sauro, J., & Kindlund, E. (2005). A method to standardize usability metrics into a single score. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, April 2-7, 2005). CHI ‘05. New York, NY: ACM.

Sears, A. (1993). Layout appropriateness: A metric for evaluating user interface widget layout. *IEEE Transactions on Software Engineering* 19(7), 707-719.

Sharp, H., Rogers, Y., & Preece, J. (2007). *Interaction Design: Beyond Human-Computer Interaction*. West Sussex, UK: John Wiley & Sons, Ltd.

Shneiderman, B., & Plaisant, C. (2005). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Boston, MA: Pearson/Addison-Wesley.

Tamir, D., Komogortsev, O. V., & Mueller, C. J. (2008). An effort and time based measure of usability. In *Proceedings of the 6th Workshop on Software Quality* (Leipzig, Germany, May 10-18, 2008). ICSE ’08.

Tullis, T. A., & Albert, W. (2008). *Measuring the user experience: Collecting, analyzing, and presenting usability metrics*. Morgan Kaufmann.

Vukelja, N., Müller, L., & Opwis, K. (2007). Are engineers condemned to design? A survey on software engineering and UI design in Switzerland. In *Lecture Notes in Computer Science* 4663/2008. Heidelberg, Germany: Springer Berlin.