LEARNING USABILITY ASSESSMENT MODELS FOR WEB SITES

A Dissertation

by

PAUL ARNOLD DAVIS

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2010

Major Subject: Computer Science

LEARNING USABILITY ASSESSMENT MODELS FOR WEB SITES

A Dissertation

by

PAUL ARNOLD DAVIS

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Frank M. Shipman, III
	Dick B. Simmons
Committee Members,	William M. Lively
	Marietta J. Tretter
Head of Department,	Valerie E. Taylor

December 2010

Major Subject: Computer Science

ABSTRACT

Learning Usability Assessment Models for Web Sites.

(December 2010)

Paul Arnold Davis, B.S., The University of Texas at Austin;

B.S., The University of Texas at Arlington;

B.S., Texas A&M University;

M.S., Texas A&M University

Co-Chairs of Advisory Committee: Dr. Frank M. Shipman,
Dr. Dick B. Simmons

This research explores an approach to learning types of usability concerns considered useful for the management of Web sites and to identifying usability concerns based on these learned models. By having one or more Web site managers rate a subset of pages in a site based on a number of usability criteria, the approach builds models that determine what automatically measurable characteristics are correlated to issues identified. To test this, the approach collected usability assessments from twelve students pursuing advanced degrees in the area of computer-human interaction. These students were divided into two groups and given different scenarios of use of a Web site. They assessed the usability of Web pages from the site, and their data was divided into a training set, used to find models, and a prediction set, used to evaluate the relative quality of models. Results show that the learned models predicted remaining data for one scenario in more categories of usability than did the single model found under the alternate scenario. Results also show how systems may prioritize usability problems for Web site managers by probability of occurrence under context rather than by merely listing pages that break specific rules, as provided by some current tools.

DEDICATION

This dissertation is dedicated to my family. It is given in thanks and appreciation to my mother, Suzanne Pflug Davis, and brother, Craig G. Davis. It is also given in memory of my good father, Arnold G. Davis, my beloved sister, Julie Sue Davis Wright ’84, and my grandparents. It is dedicated with affection to my nieces Elizabeth Wright ’14, Jacquelyn Wright ‘11, and Laura Davis as well as to the family of Don Davis, my father’s brother, and his wife Pat and their two daughters, Stephanie and Joanna.

ACKNOWLEDGEMENTS

I would like to thank my committee chairs, Dr. Frank Shipman and Dr. Dick Simmons, and my committee members, Dr. William Lively and Dr. Marietta Tretter, for their guidance throughout the course of this research.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

I wish to acknowledge and thank the Space Telescope Institute (STScI) and NASA for use of HubbleSite (hubblesite.org). My gratitude and acknowledgement go to Usablenet, Inc. for the generous use of the LIFT Machine in this study. My thanks go to graduate students at the Center for the Study of Digital Libraries (CSDL) at Texas A&M University and the Department of Computer Science and Engineering, Texas A&M University for their participation in the study.

TABLE OF CONTENTS

CHAPTER Page

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

The quality of experience that people have when visiting a Web site is a concern, and one challenge is to ensure the usability of the site. Tools supporting identification of usability issues, such as the LIFT Machine by Usablenet, Inc. [NNG 2007], find exceptions to general rules about good usability. These tools provide lists of potential problems to Web site managers who must then decide whether to change content or properties of pages. These approaches do not differentiate between the usability concerns of different sites – the list of usability issues is the same regardless of whether the site is used continuously by experts or occasionally by novices. While helpful, listing exceptions to rules may not convey their importance to those who manage or use the site.

This paper describes a novel approach to identifying usability concerns for Web sites that takes into account the expected use of the site. The approach applies a machine learning technique to identify what characteristics of a site are important or unimportant for the expected context of use. The approach extracts properties of Web pages that are indicative of usability issues. The process begins by asking Web site administrators to rate the usability of a sample of pages according to a few criteria. The sample of pages

This dissertation follows the style of *ACM Transactions on Computer-Human Interaction*.

is used as a training set, and the assessed ratings and quantitative measurements of properties are used to develop context-specific usability models. Web site administrators may then apply a model to the remaining pages of the Web site. The method identifies pages that are likely to have ratings for severity of issues with usability within ranges bounded by threshold values calculated with the prediction set. The approach categorizes Web pages on the basis of their calculated severity of usability issues. Predicted ratings for overall quality of usability are another kind of result. The method predicts ratings in five specific categories of usability.

The system described in this study finds models by means of linear regression analysis with assessed ratings and measurements of properties. Similar to the approach used by Ivory [Ivory and Hearst 2002; Ivory et al. 2000], the system quantitatively measures properties. Examples of properties are number of words, number of hypertext links, and number of images. The system architecture allows a wide range of properties to be measured. The current system, as described in this paper, applies the LIFT Machine and Ivory's techniques [Ivory and Hearst 2002; Ivory, Sinha and Hearst 2000; Ivory et al. 2001] to extract and measure different properties.

The method built into the system to learn usability models is independent of the quantitative properties used for prediction. Web site managers may apply different quantitative tools to the analysis of Web pages of a site. This flexibility allows comparing probabilities of models resulting from properties measured by different tools. Our expectation is that different, expected contexts of use (e.g. middle school students vs. use domain experts) will be predicted by different sets of page properties.

Evaluation of this approach is based on the acquisition of Web page properties by applying multiple software tools to the analysis of pages from a single Web site. The system applies the method with the objective of generating a list of pages ordered by estimated importance of usability problems. The list serves to focus a Web site administrator's attention first on those pages that most need it.

This study contributes to previous work on techniques for automatically assessing the usability of Web sites [Ivory and Hearst 2002; Ivory, Sinha and Hearst 2000; Ivory, Sinha and Hearst 2001] by considering examples of why people might use a Web site. Participants in our study were given one of two scenarios of use. Each scenario is about a group of people who use the site to accomplish different goals. The two groups of subjects then assess the usability of a set of Web pages. Thus one research question is whether models generated based on a subset of the usability ratings under a specific scenario are better at predicting usability ratings for the remainder of the ratings collected for that scenario.

The next section provides additional details concerning our overall approach. This is followed by an overview of related work in the area of automated usability analysis of Web pages. Chapter IV describes the system and its use. Chapters V and VI cover guideline and experiments respectively. Chapter VII presents results, and Chapters VIII and IX have the discussion and conclusions.

CHAPTER II

APPROACH

As the number and size of sites grow [Gulli and Signorini 2005], so will the need for resources to assess the usability of sites, and automated tools have a role in meeting this need [Brajnik 2000a]. The approach for learning usability assessment models explores the potential for automating context-specific usability assessment at page level. Figure 1 is a diagram of the model generation process.

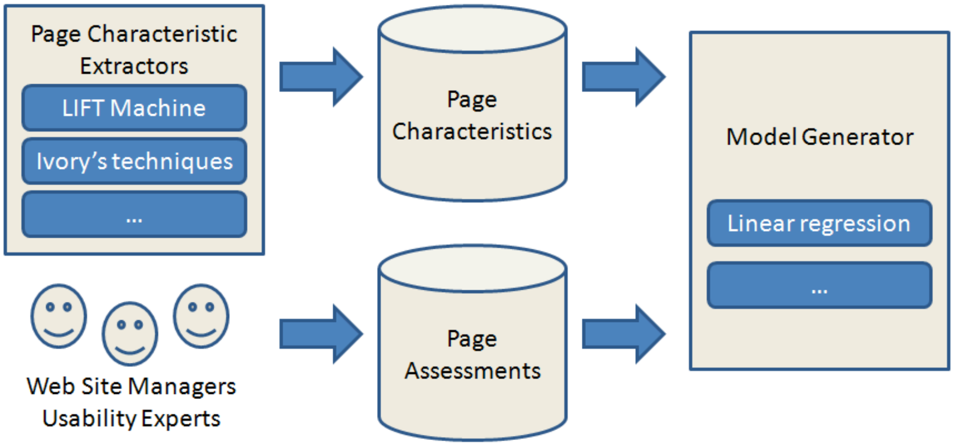


Figure 1. Model of development process.

The current system uses two software tools to collect properties of Web pages. The first is developed for this study and computes properties described in Ivory et al. [Ivory, Sinha and Hearst 2000] to be correlated with rankings of quality. The second software tool, the LIFT Machine by Usablenet, Inc. [NNG 2007], is commercially available. Among many capabilities, the LIFT Machine identifies properties associated with general usability [NNG 2007]. In practice, Web site managers would choose either of these tools, or both, in analysis of a site. Future versions of the system would incorporate additional tools.

A second data set consists of a limited number of human-generated assessments about the usability of Web pages from the Web site. Human assessments are assumed to take into account expected tasks performed when accessing the Web site and anticipated communities of users. To generate a model of value across the Web site, a sample of pages should include those that represent variety of activities, communities, and content of the site. Currently, the approach assumes confirmation of variety by human assessors.

Given these two data sets, linear regressions produce functions that map values of each property to effects on assessed usability. This process determines which values of properties provide evidence of issues with usability and which have no evidential value when assessing usability. While the study uses techniques of linear regression, future work may explore other machine learning approaches [Bishop 2006].

This study builds models by testing for likelihoods that both predicted and assessed ratings, called a “pair,” are either both above or below a threshold or cut-off value. This approach is a calculation of “precision.” Precision as used in this study is simply a ratio and meant only for this study. Although similar to other fields, such as information retrieval, the term as defined applies to this study.

This study measures two kinds of precision. Overall precision, or simply “precision,” is all relevant documents found among all retrieved documents in relation to all retrieved documents [Baeza-Yates and Ribeiro 1999; Manning et al. 2008]. If a pair falls on the side (of a threshold) not under consideration, the pair is not counted as a successful match. The second measurement of precision, called "precision at threshold" in this study, is the ratio of pairs on a side of a threshold to all retrieved documents that are also on that same side of the threshold. As with overall precision, if the cut-off value comes between the predicted and assessed ratings of a pair, the pair is not counted as a successful match. Predicted ratings are ordered sequentially. The system builds models based on processed pages but does not adapt models as more pages are processed [Finlay and Dix 2002; Russell and Norvig 2003]. Figure 2 illustrates possible combinations of predicted and assessed ratings.

Both kinds of precision are in terms of pairs that are either 1) above and at or 2) below a threshold. A pair can be at a threshold as long as the cutoff does not separate the pair. A page processed by the system will have measurements taken of properties as well as ratings assigned by heuristic assessment ("assessed rating") and model ("predicted rating".) The approach may return all processed pages. Alternatively, the approach returns a subset of all processed pages whose ordered, predicted ratings have either a floor or ceiling that is the threshold.

As an example, in Figure 2, overall precision above the threshold is 2 pairs found to 5 pages retrieved, or 40%. Precision at threshold above the cut-off is 2 pairs found to 3 pages retrieved, or 67%. For Web site managers, a precision of 40% means that 2 out of 5 pages may have issues with usability. For developers fixing problems, precision at threshold of 67% means that 2 out of 3 pages may have issues. The approach that

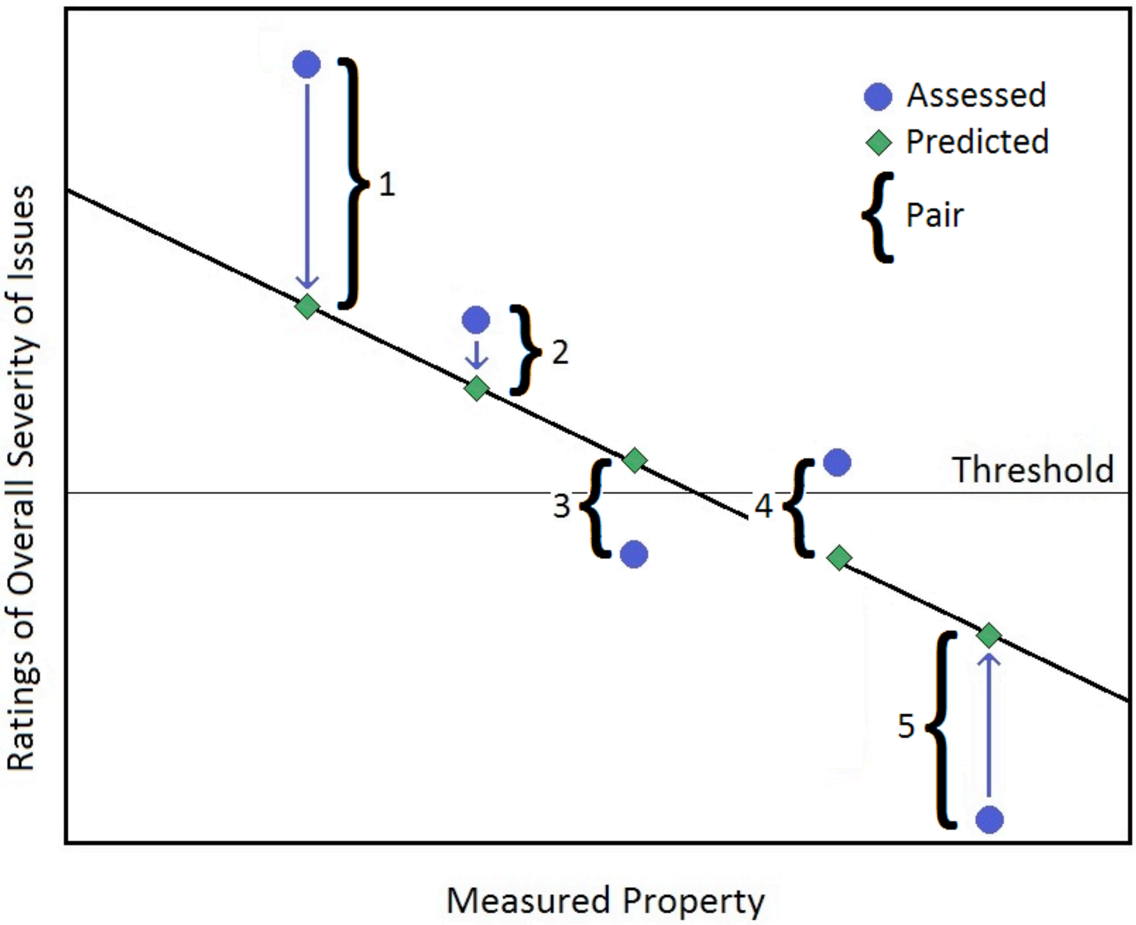


Figure 2. Model of linear relationship and paired ratings.

increases the number of successful pairs may reduce the time required to inspect and fix the designs of pages. The approach seeks models offering highest precisions.

The graph shows errors as Type I or Type II [Pallant 2007]. Type I errors are false positives. A pair with predicted rating at or above threshold but assessed rating below is a false positive. Type II errors are false negatives. A pair with predicted rating below threshold but assessed rating at or above threshold is a false negative.

CHAPTER III

RELATED WORK

A variety of studies concern the creation and improvement of the usability of interfaces, and the literature offers a wealth of information about heuristics, guidelines, frameworks, critiques, and recommendations. Published books in the field include work by Cooper [Cooper et al. 2007], Dix [Dix 1998], Nielsen [Nielsen 2000], Norman [Norman 2002], Shneiderman [Shneiderman and Plaisant 2010], and Tidwell [Tidwell 2006].

Heuristic Evaluation

When evaluating a Web site for usability issues, usability inspection is an approach for evaluating the usability of user interfaces without those for whom the interface is designed being present [Nielsen 1994]. Usability inspection applies methods to assess usability in an inexpensive manner. In general, there are four approaches to software inspection when assessing user interfaces: automatic, empirical, formal, and informal. Automatic assessment applies a program to a specification for an interface. Empirical methods have humans assess interfaces. Two empirical approaches are heuristic evaluation and user testing. Formal methods create or apply models to measure usability. Informal methods rely on knowledge and skill of subjects to find issues by “rules of thumb.” [Nielsen 1992; Nielsen 1994]

Ideally, to test the usability of a product, people who actually use the product should participate in assessments [Ivory and Hearst 2001; Nielsen 1993]. If this is not feasible, others may stand in and perform usability inspections [Nielsen 1994]. Those who do participate may evaluate for conformance to a guideline, the contents of which may range from prescriptive statements to broad principles [Ivory and Hearst 2001]. If a goal is to generalize findings, then many measurements under different contexts of use with representatives of different types of users is a useful approach [ISO 1998].

Heuristic evaluation is "a systematic inspection of a user interface design for usability" (p. 155) [Nielsen 1993]. With heuristic assessment, participants look for issues in design, development, production, and maintenance phases of the development process [Brajnik 2000b; Nielsen 1992; Nielsen 1994]. This kind of usability inspection has several characteristics. A participant (called an "evaluator" by Nielsen [Nielsen 1993]) works alone while evaluating a user interface. Only after all participants have concluded assessments may they meet to review and consolidate their findings. Participants may record their observations in writing during assessments. Another method is for an observer to write down their comments. Capture of information by audio or video recordings is also possible [Nielsen 1993].

Participants may explore as they wish and ideally revisit interfaces several times. This allows them to become familiar with the software and their interaction with elements of the interface. They evaluate their experiences with regards to a list of heuristics, described as "general rules that seem to describe common properties of usable interfaces" (p. 158) [Nielsen 1993]. Other examples of heuristics related to usability and accessibility are found those in Section 508 [ITAW 2010] and at W3C WAI [W3C 1999; W3C 2009; W3C 2010]. Nielsen described ten heuristics for usability for Web sites [Nielsen 2005].

During a heuristic evaluation, an observer should assist participants who encounter difficulties while using an interface. Observers should answer questions about content as presented through user interfaces. In contrast to user testing, observers do not have to record and interpret actions of those performing assessments. Typically, each participant should have between one to two hours to complete a session [Nielsen 1993].

The products of heuristic evaluations are lists of issues found. Participants should mark all heuristics for which violations were found. Heuristic evaluation does not explain what must be done to eliminate issues. However, by using the guideline in effect during assessments and list of issues discovered, it is possible to review the design of interfaces and apply improvements. Heuristic evaluation after changes may reveal if issues are reduced or resolved [Nielsen 1993].

Usability and Context of Use

Usability of software may differ under different contextual uses [ISO 1998; Nielsen 1993]. In the document ISO 9241-11:1998 regarding software metrics and quality [ISO 1998], a definition of context of use is: "Users, tasks, equipment (hardware, software and materials), and the physical and social environments in which a product is used …" (p. 6) which contributes to a definition of usability, which is the "… extent to which a product may be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (p. 6.) Effectiveness, efficiency, and satisfaction are called components of usability [ISO 1998]. This research includes flexibility and satisfaction among measurable attributes of usability.

If testing usability by measuring attributes, the effect of context of use is meaningful. Measurements of usability under one context of use may not apply in a different context. If a goal is to generalize findings, taking many measurements under different contexts of use with representatives of different types of users is reasonable [ISO 1998].

Requirements for usability of a product may vary [Brajnik 2004], and the expectations for usability may apply only under certain contexts of use [ISO 1998]. It is also possible that requirements are meant to apply in general and to all people [Brajnik 2004]. During assessments, it is a matter of finding attributes of components of usability that contribute or detract from usability with regards to the guideline applied [ISO 1998].

Automated Usability Testing With Web Sites

The life cycle of a Web site should include usability testing [Brajnik 2000b]. A number of authors have described practices used when designing Web sites, such as sketching designs, creating layouts of pages, and describing task completion steps for human interaction with software products [Ford and Boyarski 1997; Lynch and Horton 2008; Nielsen 2000; Shneiderman 1997; Troyer and Leune 1998]. Software development processes [Ivory 2003; Pressman 2005; Sharp et al. 2007] may include automated testing for usability of Web sites during phases after design. Testing by automated methods may take place during production, quality assurance, and maintenance phases [Ivory 2003]. The intent of testing Web sites is to "exercise software with the intent of finding (and ultimately correcting) errors." (p. 563) [Pressman 2005].

Automated Assessment of Usability of Web Sites

Automated assessment of the usability of Web sites is an underexplored area of research [Brajnik 2000a; Ivory and Chevalier 2002]. There are tools that test compliance with guidelines, such as Section 508 [ITAW 2010] and the W3C Web Accessibility Initiative [W3C 1999]. The LIFT Machine is an example of a tool with the capability for testing compliance with just Section 508 [ITAW 2010] and W3C WAI guidelines [NNG 2007]. To differentiate usability from accessibility, usability concerns interactions by people with Web sites whether or not they have disabilities. Accessibility addresses needs of those with disabilities and for whom assistive technologies may prove beneficial [W3C 2005]. If a Web site is to be usable by all, it should provide for those who require assistance in a manner that addresses limitations imposed by their disabilities [ITAW 2010; W3C 1999].

If usability criteria are measurable, automation may allow analysis of usability. Decomposition of components of usability (effectiveness, efficiency, and satisfaction) allows counts of constituent attributes. However, there may be no overall rule about what these attributes are or what they measure. If values of an attribute contribute or detract from use of a product in terms of a component, then measurements of those values may reveal potential issues.

If testing usability by measuring attributes, the effect of context of use is meaningful [ISO 1998]. Measurements of usability under one context of use may not apply in a different context of use. If a goal is to generalize findings, many measurements taken under different contexts of use with representatives of different types of users is a reasonable approach [ISO 1998].

For example, the Web Content Accessibility Guidelines (WCAG) are useful to those who evaluate whether Web sites provide accessible content [W3C 2008]. The guidelines give guidance about implementation of requirements, such as use of alternative text for images in Web pages [W3C 2005]. Implementation of this requirement follows technical specifications, such as correct use of the ALT attribute in the HTML protocol [W3C 2005]. If the requirement is not implemented, it is possible to detect the exception. For example, if all images are to have values for ALT, an automated method may detect those that do not. However, even if the requirement is fulfilled, implementation may not produce a usable product. An image may have a textual value for ALT, but the value may not be relevant to the content of the image. Consequently, automation may not correctly detect all issues.

If a Web manager is familiar with the Web site, it is possible to place a boundary around intended use of the site and choice of automation. Use of automated tools is not indicated if analysis “requires interpretation (e.g. usage of natural and concise language)” of content or “… assessment of relevance (e.g. ALT text of an image is equivalent to the image itself)" (p. 229) [Brajnik 2004]. In this case, a technology might not detect discrepancies in the planned use of a site. In some situations, the use of automation may give confusing results, and human intervention may be needed to interpret what was the planned use of interfaces. Even then, erroneous conclusions may result [Suchman 1987].

There is concern that automated methods cannot analyze Web sites as well as humans [Brajnik 2004; Nielsen 1993]. This concern would extend to assessments of Web sites under different contexts of use. A tool like the LIFT Machine may process thousands of Web pages, analyzing each for properties associated with usability and doing so tirelessly. However, it is possible that, "everything that has to do with human interpretation and context of use is likely to be poorly machine testable." (p. 230) [Brajnik 2004]. This study explores how well a semi-automated process might combine human interpretations of usability with automated analysis under different contexts of use.

The approach of this study tests interfaces of existing Web sites to build models. This differentiates it from methodologies that build contextual adaptations into Web-based applications. These methodologies apply structural models about entities and relationships with the purpose of enabling interfaces to change in anticipation of contextual needs and preferences of users [Kapitsaki et al. 2009]. Approaches have used UML in studies such as WebML [Ceri et al. 2000] and a model-driven approach for mobile systems [Kapitsaki, Kateros, Prezerakos and Venieris 2009]. Object-oriented methods, such as OO-HDM [Schwabe and Rossi 1998], facilitate authoring of navigation and interface elements that map onto various environments. Instead, this study concerns analysis and evaluation of usability of interfaces to Web sites under contexts of use.

Ivory's Categories of Automated Evaluation

Ivory [Ivory 2003; Ivory and Hearst 2001] proposed a taxonomy of six automated evaluation methods grouped into Method Classes. This study continues with this taxonomy and contributes a new approach.

The Method Classes are testing, inspection, inquiry, analytical modeling, and simulation. The sixth Method Class was named for the Web TANGO project. Each Method Class contains a Method Type, Automation Type, and Effort Level. Within each Method Class is a Method Type that describes methods of evaluation, or "how an evaluation is performed" (p. 18) [Ivory 2003]. An Automation Type describes a level of automation by an evaluation method. In Ivory's classification, automation types are none, capture, analysis, and critique. None means no automation performed. Capture indicates automated recording of data, such as log data. Analysis indicates an approach that predicts issues. Critique describes automated suggestion of improvements. By Ivory's schema, Effort Level describes how much human intervention is a part of an evaluation method. The levels are minimal (no human use of interface), model development (creation of interface by human is necessary), informal use (interaction to complete tasks is unrestricted), and formal use (interaction constrained to complete certain tasks.) [Ivory 2003]

The contribution of the approach and system described by this study is based on the Model Class called Web TANGO, which is an approach that applies statistical analysis to find models ("profiles") [Ivory 2003] to predict rankings for a site or pages. In contribution, this study builds models at page-level only and under context of use of how people might actually use a site. This study contributes the following:

- Contribution to the Method Class, Web TANGO.
- Examination at page-level by heuristic assessment [Ivory, Sinha and Hearst 2001].
- Test if usability differs if different groups of people interact with the same site under different scenarios of contextual use.
- Test if usability predicted by models differs if the models incorporate evaluations of interfaces by humans under different contexts of use.

The next subsections list software tools and describe examples of tools by Model Class. Following this, there is an examination of differences between the approach of the Web TANGO project and the contribution of this study.

Usability Testing

In usability testing, participants examine and collect data about how people use Web sites while completing tasks and then critique the usability of the sites [Ivory 2003]. An automated approach may analyze log files, collect data generated during use, or measure activity at Web servers. Examples of tools used to analyze log files are QUIP [Helfrich and Landay 1999], KALDI [Al-Qaimari and McRostie 1999], and WebQuilt

[Hong and Landay 2001]. Tools that measure usage are AWUSA [Tiedtke et al. 2002] and LumberJack [Chi et al. 2002]. There are commercial tools that monitor activity at Web servers. Examples are SiteAngel (part of Performance Manager Express) from BMC Software [BMC 2010; Hulme 2000] and Resource Analyzer from IBM [IBM ; IBM 2010].

Usability Inspection

Inspection techniques carried out by participants include heuristic assessment [Nielsen 1994] and cognitive walkthroughs [Blackmon et al. 2002]. Automated methods may apply a number of tools to count features associated with usability of Web sites.

A-Prompt is an example of such a tool. It tests the conformance of Web sites to W3C WAI Web Content Accessibility Guidelines 1.0 [W3C 1999] and Section 508 [ITAW 2010]. The tool also attempts to correct properties of selected exceptions [ATRC 2010]. This tool is only one of many that test compliance with the W3C WAI and Section 508 guidelines. The W3C WAI Web site lists the A-Prompt tool plus many others [W3C 2006].

If a tool identifies Web pages as having attributes (properties) whose values are associated with usability problems, it is possible to inspect those pages more closely. It is not always necessary to inspect all such Web pages. For example, it is possible for a tool to disallow certain values for properties [Brajnik 2004]. Once disallowed, only corrections of other exceptions are required. An example is an ALT attribute of an image that is blank instead of containing text that appropriately describes the image.

The same property might have values that reduce usability, for example, when content incorrectly describes an image. Because the values are not disallowed, no attention is brought to correct the issue.

Knowledge-based Web Automatic Reconfigurable evaluation with guidelines optimization (KWARESMI) is an automated tool that uses encoded usability guidelines to test sites for issues. It applies guidelines encoded as HTML elements in the analysis of Web pages, and it allows reconfiguration for different guidelines by testing different sets of HTML elements. Encoding of the guideline is done by means of a Guideline Definition Language (GDL). The approach allows selective application of guidelines, and this in turn expands the flexibility of the usefulness of the tool when analyzing different kinds of sites. KWARESMI relies on humans to choose which guideline to apply. The tool can generate reports and identify severity of issues [Beirekdar et al. 2005; Beirekdar et al. 2002; Beirekdar et al. 2003].

Usability Inquiry

Questionnaires and surveys are tools used in Usability Inquiry. Evaluation of results requires human intervention, and this is a field of automated usability analysis that requires more research. An example of a tool that is useful for making questionnaires is NetRaker [NetRaker 2010]. It allows remote observation by a person of a display screen through which another person interacts by means of a computer. The tool allows markup by both parties and is capable of recording test sessions.

Analytical Modeling

Analytical modeling allows prediction of usability issues by using models of how users interact with interfaces while completing tasks. Examples of approaches in analytical modeling are GOMS [Card et al. 1983], GLEAN [Kieras et al. 1995], WUSAB [Atterer 2008; Atterer et al. 2006], and USAGE [Byrne et al. 1994].

GOMS (Goals, Operators, Methods, and Selection Rules) is a model that describes essential interactions that users have with a user interface while completing tasks to reach a goal [Card, Moran and Newell 1983]. A model consists of one or more methods that consist of operators (procedural acts) that users perform (such as pressing a key on a keyboard.) Selection rules determine the method to be followed as a user interacts with an interface to complete steps [Card, Moran and Newell 1983]. There are a variety of GOMS models [John and Kieras 1996]. For example, the Keystroke-Level Model predicts time needed to complete a sequence of keystrokes to accomplish a task [Card et al. 1980]. GOMS requires knowing sequences of tasks and the resulting goals possible by task analysis. Strength of the modeling approach is that it allows predictions of numbers and efficiencies of procedures to follow as a result of constraints placed on what users must do to reach goals. USAGE was an approach to automate construction of formal GOMS models through use of interfaces [Byrne, Wood, Foley, Kieras and Sukaviriya 1994]. These models typically have limitations when users interact with interfaces that are not in a manner described by models [Kieras, Wood, Abotel and Hornof 1995].

GLEAN is a system that automates application of GOMS models [Kieras, Wood, Abotel and Hornof 1995]. The system reads a Model as input. The Model is written in the Natural GOMS Language (NGOMSL) [Kieras 1997] and predicts how long it will take users to perform steps and reach goals. The study seeks to reduce the time needed to predict usability, provide an easy way to process scripts written with NGOMSL, and allow reuse of models. However, a limitation of GOMS constrains the use of GLEAN. GOMS does not address the effects of context of use, and it does not address well any mistakes that humans make [John and Kieras 1996; Kieras 1997].

There are approaches that use languages to describe interaction. WUSAB is an approach that tests conformance of interfaces to Web pages with requirements [Atterer 2008; Atterer, Schmidt and Hussman 2006]. It does so by adding models of context and users to integrated development environments. Models written with UML describe page layouts for navigation and presentation as well as constraints, such as response time. The models describe users, the environment, and browser. As development proceeds, a validator detects exceptions to models, and developers address exceptions found, such as corrections with presentation or navigation. Models of users come from information collected about the group of people for whom the application is intended [Atterer 2008] [Atterer, Schmidt and Hussman 2006].

A method by Vanderdonckt and Beirekdar [Vanderdonckt and Beirekdar 2005] applies a Guideline Definition Language (GDL.) A GDL describes a formal guideline. An evaluation engine parses Web pages and compares measurements of properties to the guideline. The design of the system proposed in this study may also test different guidelines but it does so through heuristic assessment. Although use of GDL allows flexibility with the guideline applied, it does not evaluate people under context of use.

Analytical approaches have also tested whether features of Web sites correlate with measures of those sites in order to compute ratings of comprehensibility of content [Ma et al. 2007; Yan et al. 2007]. The approach used by Yan [Yan, Zhang and Garcia 2007] collected data for 191 metrics from 800 sites, and four librarians evaluated 25 to 50 sites for "Information Value, Information Credibility, Media Instructional Value, Affective Attention, Organization and Usability" (p. 194) [Yan, Zhang and Garcia 2007]. After multiple regression, the resulting model had 81 features with an adjusted R^2 of 0.437. The study proposed an analytical method of computing ratings of Web site comprehensibility as might benefit those who must quickly decide whether to read content returned from searches of the Web.

Simulation

Simulation is an approach that simulates how users interact with the Web by means of models. Bloodhound [Chi et al. 2003] is an approach that models users by finding similarities about content across Web pages. It associates content with log data and attempts to predict whether users will find information while navigating among pages. Models represent users who move through the site and attempt to complete tasks and reach target pages.

Bloodhound applies the concept of "Information Scent" [Chi et al. 2000]. Content and other clues at a hypertext link may give information about content at the distal end of the link. Users decide whether to follow the hypertext link based on the content at the proximal end. The approach simulates interaction and tests whether changes to information scent may improve overall experience.

Web TANGO Method

In a series of studies, Ivory et al. [Ivory and Hearst 2002; Ivory, Sinha and Hearst 2000; Ivory, Sinha and Hearst 2001] proposed and tested automated methods that statistically predicted whether or not sites would rank highly for quality. They found that several measurements of properties of pages correlated with rankings of quality. Fundamental to Web TANGO was measurement of properties of elements that comprise interfaces to Web pages. Later work by Ivory proposed design patterns arising from a longitudinal study of properties at sites receiving high rankings for quality [Ivory and Megraw 2005].

Web TANGO is an approach that applies statistical analysis to assess whether a site has characteristics that would rank it among those with “good” or “poor” quality. Web TANGO measures formats, compositions, and properties (such as count of words in body of page, count of hypertext links, size in bytes of Web page and images, number of images in a page, and so forth) of Web pages from hundreds of sites [Ivory and Hearst 2002; Ivory, Sinha and Hearst 2001].

Ivory et al. described quality as that which gives “value to users, content and design” in contribution to “popularity” of a product (p. 4) [Ivory, Sinha and Hearst 2000]. They selected 463 informational sites and sources of review, one of which was the Webby Awards [IADAS 2010]. They separated sites into categories of those receiving favorable rankings from those that were unranked. With Web TANGO, they measured properties of pages in both categories. As a result, they produced a linear regression model with multiple predictors that categorized sites through the use of un-standardized coefficients of properties (Link Count, Reading Complexity, Text Positioning Count, Color Count, Page Size, and % Body Text.) Their multivariate linear model had an *F* statistic of 4.369 ($p < 0.001$) [Ivory, Sinha and Hearst 2000].

Ivory et al. [Ivory and Hearst 2002] continued their work into automated analysis of Web sites by testing thresholds with ratings. They arbitrarily chose cut-offs of an upper 33% for top-ranked and a lower 33% for bottom ranked sites. This partitioned the distribution of ratings into categories of good, average, and poor. Their software tool, an expanded version of Web TANGO, collected 157 properties. Using 333 sites evaluated in the 2000 Webby Awards [IADAS 2010], they processed sites into six different topical categories (education, community, living, health, services, and finance.) They then applied a decision tree classifier and generated 144 rules based on properties that correlated significantly with rankings. The classifier obtained an accuracy of 94% overall for correctly classifying sites [Ivory and Hearst 2002].

To test the usefulness of their model to find issues so that they could be corrected, Ivory et al. [Ivory and Hearst 2002] applied the classifier against a small Web site to do page-level analysis. It measured several properties and identified the site as having a low quality ranking. They edited the properties identified by the classifier and retested the pages. This time the classifier ranked the site higher. The site they had tested was small (9 pages) and in the contextual category of Yahoo! Education/Health.

Context of Use

A goal of this study is to support Web site managers. Analysis of a site is at page-level only. This requires finding models that rate pages for their potential for issues by category of usability. The study includes Overall Usability in analyses. Second, the models are built under a scenario of contextual use. This supports an expected use that Web site managers may have of their site. This is in contrast to the approach used with Web TANGO, which built general models that ranked sites by non-usability categorical topic [Ivory and Hearst 2002]. In this study, the approach is to analyze stylistic elements used in the design of pages with regards to intended audiences by heuristic assessments. The approach is then to build models to predict ratings that might be given in heuristic assessments.

Other approaches have included context of use in analysis of Web sites. USIXML is an approach that encodes the design of Web pages into XML. Likewise, USIXML stores information about context that includes properties of the environment, the intended users of the site, and computational equipment [Limbourg and Vanderdonckt 2004]. WebML (Web Modeling Language) also allows encoding of context of use. Similar to USIXML, contextual information includes descriptions of the user, place of use, and device used. WebML was described as not being useful for capturing information about goals of users [Ceri et al. 2007].

CHAPTER IV

DESCRIPTION OF SYSTEM

As described previously, building context-specific models relies on extracting page characteristics (properties) and collecting human assessments (ratings) for a set of Web pages (training set). The section has descriptions of the details of this process and how results of the learned model are compared to human assessments by testing an additional set of Web pages (prediction set). Figure 2 shows a design of the system.

The approach is only partially automated. The design of the system is modeled after a generic search engine [Croft et al. 2009]. The system contains modules that parse Web pages, and each module is called a Metric Collection Tool (MCT.) In this study, the system applies two different MCTs in Metrics Processing. Each collects data for non-intersecting sets of properties. By statistical analysis, we find measurements and ratings that correlate ($p < 0.05$) and compute likelihoods and accuracy. In the design, Rating & Ordering rates and orders pages for managers.

The Web Page Metric Analyzer, or WPMA, is a MCT that parses Web pages for fifteen properties. The fifteen include eleven selected for use by Ivory et al. [Ivory, Sinha and Hearst 2000] in their studies. Inspection of pages found these fifteen properties amenable to quantitative analysis. Implementation of the WPMA was specific to the Web site chosen; a more general parser would be needed to capture the same properties from other Web sites. The second MCT used is the commercial product called the LIFT Machine [NNG 2007]. Usablenet, Inc. had associated properties with a guideline about good practices for general usability [NNG 2007]. The LIFT Machine tests for use of these properties.

In this study, statistical analysis requires intervention by a practitioner. SPSS v. 15 is used in Statistical Analysis [SPSS 2006]. Although the models are simple, a practitioner applies models, built with a training set, to a prediction set and evaluates accuracy and precision. Analysis of results referenced several sources for methods to test for correlations, t tests, normal distributions, and other statistical measures [De Veaux et al. 2008; George and Mallery 2008; Neter 1996; Neter and Wasserman 1974; Pallant 2007; Shannon and Davenport 2001; Tabachnick and Fidell 2007].

Overview of the System

Figure 3 is a data-centric view of the architecture of the system [Fowler 2004]. This study used HTTrack [77]. All pages would be visually inspected, and those that were advertisements or simply files of Flash animations would not be used. The study would only use files with an extension of .html. With regards to the organization of the site, relationships found between assessed ratings and measurements of depth of path traveled would use a computed number of anticipated clicks [NNG 2007]. The study did not collect data about depth of path from symbolic representations, as might be displayed on pages as breadcrumb trails [Nielsen 2000]. The study would also not move files from their original locations within the structure of the site.

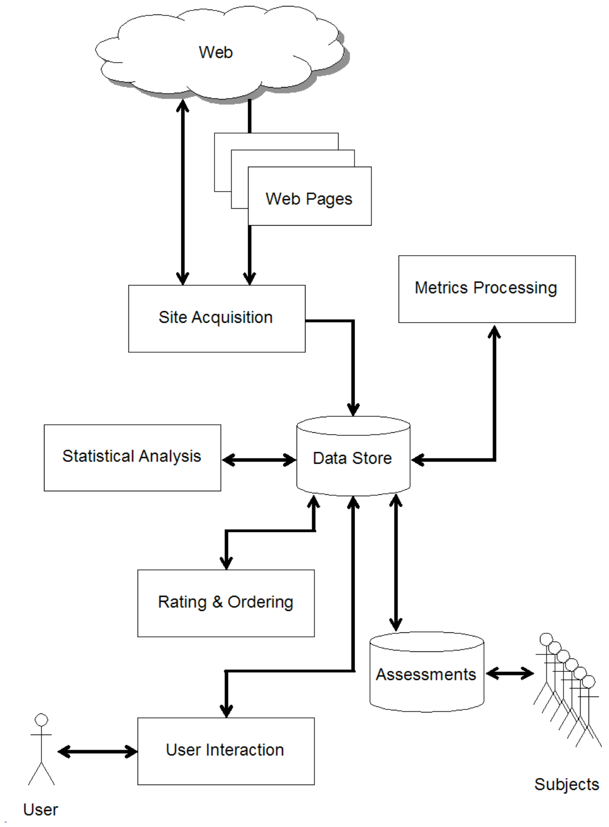


Figure 3. Data-centric design of system.

The Web Site

The study selected an informational site about the NASA Hubble Space Telescope that is made available at http://hubblesite.org through the Space Telescope Science Institute (STScI) [STScI 2010]. The International Academy of Digital Arts and Sciences (IADAS) recognized the site as Best Science Website during the Webby Awards in 2007 [IADAS 2010].

The crawler, HTTrack Website Copier, version 3.41-3 [Roche 2007], downloaded the Web site from hubblesite.org on August 20, 2007. It copied all files to a depth of three levels starting with the home page, and all files went to local storage. The approach only analyzed Web pages with file extensions of .html. Analysis was of static pages and did not count or analyze scripts or objects embedded in pages.

Web Page Metric Analyzer

The Web Page Metric Analyzer (WPMA) quantitatively measures certain properties of Web pages. Examples of properties include counts of HTML elements (e.g. number of headings, links), number of words in the body of the page, and sizes of image files. WPMA measures the following fifteen properties. An acronym follows a description of each test.

1. Number of all words displayed on a Web page (TW)
2. Number of words in body of Web page (BW)
3. Number of words not in body of Web page (NBW)
4. Number of words emphasized by bold face or italics (EW)
5. Number of text areas with a non-white background, with borders, with horizontal rule, or in a list (CLU)
6. Number of times blocks of text are not positioned flush left (NFL)
7. Number of unique colors of fonts used for all words (FC).
8. Number of internal and external hypertext links (LNK)
9. Number of images embedded in a page (not in scripts, applets, or objects) (IMG)
10. Percentage of Body words emphasized by bold face or italics (PEBW)
11. Percentage of Non-body Words to Total Words (PNBW)

12. Total of sizes of files of images used in a Web page (TIFS)
13. Total of Web Page File Size, Total of Image File Sizes, and size of file of Cascading Stylesheets for a Web page (TAMFS)
14. Size of file of Web page only without including sizes of files of images or CSS (WPFS)
15. Percent of Image File Size to Web Page File Size (PIWPFS)

Although the WPMA tests for differently colored fonts by inline style and deprecated color tag, it does not test for change of font color by CSS files. Consequently, analysis did not include counts of font color. The software tool tests properties associated with Web sites judged to exhibit good or poor quality [Ivory 2000]. In addition, the design of the WPMA measures properties that are not collected by the LIFT Machine.

The LIFT Machine

The LIFT Machine contains rules to measure general usability as applicable to people with and without visual and mobility limitations [Ivory and Chevalier 2002; NNG 2007]. The software tool generates assessments about the usability of Web pages by applying 27 tests in five categories [NNG 2007]. The categories and rules are as follows:

Effectiveness

1. Possible misspelled words detected
2. Errors possible from incorrect parsing of HTML page
3. Page should have only one BODY element
4. Do not use IFRAME in Web page
5. Do not use the MARQUEE element
6. Do not use the SPACER element

Flexibility

1. Frames should allow resizing
2. Make mail addresses explicitly stated

Navigability

1. Auto-loaded pages should have backward hypertext links duplicated
2. Auto-redirected hypertext links pointing forward to a visited page should be duplicated
3. No external hypertext links should be broken
4. In each Web page make it possible to return directly to the home page
5. Duplicate image map links in text
6. Give each hypertext link a label
7. Link to relative URL appears invalid
8. No local hypertext links should be broken
9. Display a logical path from the home page to the current Web page
10. A Web page should not have a hypertext link pointing back to itself

11. Use standard colors for hypertext links
12. Have appropriately targeted hypertext links in frames

Satisfaction

1. Check that the attributes of HEIGHT and WIDTH of a GIF image are the actual height and width of the image
2. Do not use the BLINK element

Efficiency

1. Image element should have size attributes
2. Do not put pages deeply within the site and away from the home page
3. Do not have unused map elements
4. Keep the size of the page small
5. Keep the average number of clicks to reach all pages in the site low

Descriptions of the Usability package, version 1.1, incorporated in the LIFT Machine, version 1.9, 2004-2005, by Usablenet, Inc. listed these tests and their descriptions [NNG 2007].

Scenarios of Contextual Use

The study had two groups of subjects analyze the same pages of a Web site under two scenarios of contextual use. Each scenario described people who require information from the site but for different purposes. The following are descriptions used in experiments with participants. Scenario No. 1 presents the following description to one group of subjects.

> "Assess the Web site of the NASA Hubble Space Telescope for use by middle school students. These students attend classes in physical science or environmental science. They all study the exploration of space. Their ages are between 12 and 15, and there are both boys and girls. All are familiar with how to use a Web browser."

The other description, Scenario No. 2, presents the following situation to the second group.

> "Assess the Web site of the NASA Hubble Space Telescope for use by astronomers. These scientists search for information, photos, and illustrations regarding astronomical phenomena and scientific instruments. They publish to scientific journals. Their ages are between 25 and 65, and there are both men and women. All are familiar with how to use a Web browser. Their academic credentials include a Masters or Ph.D. in physics or astronomy."

Precision and Accuracy

As described in the chapter "Approach," there are two kinds of precision defined and applied in this study. They are called "overall precision" and "precision at threshold." For both kinds, the numerator is the count of pages having both predicted and assessed ratings, i.e. a "pair," on the same side of a threshold. For precision at threshold, the denominator is the count of all predicted ratings that fall on the same side of the threshold as those counted for the numerator. Overall precision uses the total number of predicted ratings, which also equals the total number of pages tested.

To demonstrate how the two kinds of precision might be used, a manager applies a model that predicts ratings for severity of issues with Effectiveness. Developers are to inspect the use of the property (the predictor in the model) in all pages rating at or above the threshold. For example, a manager might find overall precision to be 25% and instructs developers to investigate those pages placed above the threshold. Examining only those pages with pairs at or above the threshold, developers find precision at threshold to be 75%. This means that 3 of 4 pages may have usability issues, and the system provides a list with pages in descending order by predicted severity of issue.

Accuracy is a measure of the mean of absolute values of residuals. A residual is the absolute value of the difference between assessed and predicted ratings for a measurement of property. If the mean of residuals is zero, the model is perfectly predictive [Badi et al. 2006; Neter 1996]. A perfectly predictive model would predict the assessed rating in the category correctly for every page tested. If comparing two models in the same category, the one with a smaller mean has higher accuracy.

The approach uses the mean of ratings predicted by a model as the threshold. . The decision to use the mean of predicted ratings was arbitrary. Consequently, two models might have different thresholds. In this study, there is no interpretation of a Web page as having "good" or "poor" usability based on how many ratings fall above or below a threshold. The study proposes an approach that allows users, such as Web site managers, to set their own requirements for thresholds.

This study tests if it is possible to find models in each category as well as for Overall Usability. In this early study, this is done with a single predictor. This study also tests whether it is possible to build multivariate models for Overall Usability. Practical application of the approach would test for other multivariate models in the categories of usability. If more than one model appears in the same category, the study compares precision and accuracy. Managers may choose which model to apply based on their needs for testing.

CHAPTER V

GUIDELINE

Subjects assessed Web pages from the selected site by identifying and rating severity of issues. They also rated pages for overall usability. They applied a guideline having five categories of usability, and the approach added a description for Overall Usability. Except for Overall Usability, the categories and their descriptions come from documentation available from the UsableNet Web site (http://www.usablenet.com), in the Describe Package under “Resources,” for the LIFT Machine, version 1.9, 2004-2005. (Access requires permission from UsableNet, Inc.). Subjects give a comprehensive rating, called the “assessed rating,” in each category as well as Overall Usability for the Web page they review. The following lists these and descriptions.

1. Effectiveness – Persons visiting the site should be able to accomplish their goals, such as finding information.
2. Flexibility – There should be more than one way to reach a goal. This also means that people should be able to use the site if the page appears in browser windows of different sizes.
3. Navigability – People should find it easy to learn where they are in a site as well as how to go elsewhere within the site. They should be able to remember where a page is if they return to the site.
4. Satisfaction – People should not tire or become upset when using a site. They should find the experience satisfying and satisfactory.

5. Visitor Efficiency – People should find that the ease of use and performance of the site are satisfactory. This can include obtaining pages, determining if those pages are useful, and finding hypertext links to other pages.
6. Overall Usability – the quality of usability of a Web page overall.

For categories, subjects rate pages with a scale from 0 (no issues) to 5 (severe issues). They also rate Overall Usability with a scale from 1 (very worst) to 10 (very best).

CHAPTER VI

EXPERIMENTS

Experiments took place in the spring semester of 2009 at the Evans Library Annex on the campus of Texas A&M University. To evaluate the approach, the study addressed the following questions:

1. Do groups of people give different ratings for categorical usability and Overall Usability of Web pages under different scenarios of use?
2. Do models have precision greater than 50% in categories of usability or for Overall Usability?
3. Are precision and accuracy different for models in the same category or for Overall Usability under the same scenario of use?
4. Are precision and accuracy different for models different in the same category or for Overall Usability under different scenarios of contextual use?

Twelve participants were randomly divided into two groups of six. The groups assessed the same downloaded version of HubbleSite [STScI 2007]. Each group performed assessments by applying different scenarios.

All subjects were graduate students with the Department of Computer Science and Engineering. All had taken advanced coursework in computer science. The first group assessed under a scenario for middle school students (ages 12 to 14, boys and girls) who sought information to use in a science report. This was Scenario No. 1, and the group had four men and two women. Of these subjects, one was between 18 and 25 years old. Two were between 25 and 30. Two were between 30 and 34, and one was between 40 and 45.

The second group assessing the usability of the site was given a scenario of astronomers and physicists (no age range, men and women) with graduate degrees in physics seeking information for scientific reports. This was Scenario No. 2. The second group had six men, of who four were between 25 and 30 years of age. One person was between 30 and 34, and the sixth person was between 40 and 45.

All subjects participated under controlled conditions. Each sat alone in a room with table, chair, laptop computer, mouse, and keyboard. The computer, an Apple MacBook Pro [Apple 2010] was not joined to a network but had a copy of the downloaded Web site stored locally. The computer was booted to Windows Vista to display an Internet Explorer v. 6 browser window [Microsoft 2010]. Each room had overhead lighting, central air conditioning, and no outside windows. Each room had an observational window by a single door facing a hallway. Subjects sat at the table and faced away from the door. All used paper forms to record their assessments and comments. All subjects used Microsoft Internet Explorer version 6 on Microsoft Windows Vista.

To prepare subjects for experiments, each received a written explanation about usability and was referred to Web sites about usability. Half of each group reported to have reviewed these materials. Only one reported to have visited HubbleSite but only to the level of home page. A proctor sat outside the room and out of direct view of the subject. (The proctor was the Principal Investigator.) Each experimental session lasted no more than two hours, and the proctor offered each subject at conclusion a gift card valued at $25.00. The design of the experiment had received approval by the Institutional Review Board of Texas A&M University for participation of humans in experiments.

Subjects could navigate the downloaded site freely but could only give assessments for certain pages. The approach selected 102 out of a total of 2,017 Web pages and assigned each a unique number, called a PageID. The number and name of page appeared in the title of the HTML heading and at the top of the browser window. There were no modifications to Web pages except for adding PageIDs and inactivating search capabilities built into pages. Pages receiving a PageID were those reachable by top and side menus as well as hypertext links within the body of pages. The home page also had a PageID.

The structure of the site was like a hierarchy with a top-level home page and with pages at lower levels linking back home as well as to other pages at the higher, the same, and deeper levels. When first encountering a page, subjects determined if it had a PageID. If so, they recorded the number and their ratings on a paper form.

During assessments, both groups used a list of ten questions to aid their exploration of the site. They were free to mark the number of questions answered as well as write comments. The following questions are taken from the list as examples:

- What are the dimensions of the Hubble Space Telescope?
- What is the latest news about the telescope and the project itself?

- Of what in space has the telescope taken images most recently?
- Where can you listen for news about the project?
- What is visible in tonight’s sky?
- Where is the Space Science Education Resource Directory?
- Where can you find calculators to find temperature, distance, and redshift?

CHAPTER VII

RESULTS

This is an overview of results with details in subsections to follow. Results confirmed that the approach builds models of usability for the categories tested as well as for Overall Usability, and that results indicate differences between ratings given under different contexts of use. The P-value for all comparisons was 0.05. The approach built and tested 15 bivariate models and two multivariate models for OU under Scenario 1. In contrast, under Scenario 2 the approach built one bivariate model (NAV) and two models by multiple regressions for OU. The approach also found that the means of ratings predicted by bivariate models for Navigability under the two scenarios of contextual use were not significantly different (independent samples t test.) Likewise, assessed ratings for Navigability between scenarios were not different. However, accuracies of bivariate models for Navigability between scenarios were significantly different. Comparing between scenarios, the means of ratings predicted by multivariate models for Overall Usability were significantly different. The means of assessed ratings for Overall Usability, as well as for Effectiveness, were also significantly different between scenarios. Models produced different coefficients of determination, accuracies, and precisions within categories under the same scenario as well as between scenarios.

Web pages have interfaces for which interactions are planned. If groups of people use interfaces for different purposes, the kinds and severities of issues may be different. Different kinds and numbers of models may reflect this. If models found under different scenarios produce similar ratings, people may report experiencing a common problem, such as with Navigability. If models rate issues with usability under one scenario but no model is found for that usability under a different scenario, it may be that one model does not identify issues experienced in common for those contextual uses. In this case, one model that tests interactions planned in the design of a Web site may not adequately address how people actually use the site.

At conclusion of experiments, both groups of subjects had assessed 63 Web pages in common. Selecting 32 of these pages randomly as the training set, the remaining 31 pages served as the prediction set. Analysis found correlations under each scenario between predictors (measurements of properties) and response variables (assessed ratings), and this permitted identifying and testing models.

Results for Assessments between Scenarios

Two groups under different scenarios of contextual use rated the same Web pages for five categories of usability and Overall Usability. By independent samples t test ($p < 0.05$, N = 63), assessed ratings for Effectiveness (EFF) and Overall Usability (OU) were significantly different (see Appendix.) The groups did not give significantly different ratings for Flexibility (FLX), Navigability (NAV), Satisfaction (SAT), and Efficiency (EFC.) Participants had indicated that middle school children might find the severity of issues with Effectiveness of pages examined to be greater than might be reported by astronomers. The mean of ratings for severity of issues with Effectiveness under Scenario 1 was higher than that reported under Scenario 2 (1.036 versus 0.691

respectively.) If pages have more severe issues with Effectiveness, they might also be less usable overall. Results with Overall Usability supported this proposition. The mean of ratings for Overall Usability was lower under Scenario 1 than for Scenario 2 (respectively 7.569 versus 8.087).

Examination of Models

The subsections that follow describe the models that were built. Each subsection either describes features or tests application of the models. This chapter addresses objectives presented in the "Introduction" and "Experiments." The final subsection has illustrations of the concept of precisions, as described in "Approach."

Bivariate, Linear Models

In all, the approach found six models with the WPMA and nine with the LIFT Machine under Scenario 1. Table 1 lists these models. Model 4 has the best fit to data with an F statistic of 20.284 and p = 0.000.

To understand variability of response variables accounted for by predictors, the approach calculated coefficients of determination (R^2). Adjusted R^2 describes how well a model may generalize if more predictors are added to the model. An indication that predictor(s) may be missing from a model is if R^2 and adjusted R^2 are significantly different [De Veaux, Velleman and Bock 2008]. Model 4 has the highest adjusted R^2, while all other models showed lower coefficients of determination.

Table 1.
Bivariate, linear models by WPMA,
scenario 1, training set (N = 31, p < 0.05)

Model	Resp. Var.	Pred.	Coeff.	Const.	F	Sig
1	OU	LNPNBW	0.227	6.827	4.688	0.038
2	FLX	EW	-0.013	1.055	5.640	0.024
3	FLX	LNLNK	-0.488	2.406	4.807	0.036
4	NAV	LNNBW	-0.780	4.110	20.284	0.000
5	NAV	LNLNK	-1.122	4.869	9.091	0.005
6	SAT	LNWPFS	1.130	-9.410	5.757	0.023

A test for independence among predicted results (Durbin Watson) identifies equations showing first-order autocorrelation among consecutive predicted results. Autocorrelation indicates some dependence, not independence, among predicted results. A result outside a range of 1.5 to 2.5 is a sign of autocorrelation [Neter 1996]. The study did not use a model, bivariate or multivariate, if autocorrelation was indicated.

Examining coefficients of determination, of the two models found for Flexibility (FLX) and Navigability (NAV), the highest R^2 and adjusted R^2 were respectively 0.158 and 0.130 for model 2 and 0.403 and 0.384 for model 4. In other categories, the R^2 and adjusted R^2 of model 1 (OU) were 0.135 and 0.106. For Satisfaction (SAT), the R^2 and adjusted R^2 of model 6 were 0.161 and 0.133. Except for model 4 (38.4%), predictors of models from WPMA accounted for less than 20% of the variability of assessed ratings.

Models found with the LIFT Machine are in Table 2. Like the WPMA, models appeared in three categories (NAV, OU, and SAT.) For NAV, model 7 had a R^2 and adjusted R^2 of 0.245 and 0.220 respectively. Of the three models (8, 9, and 10) for SAT, model 10 had the highest R^2 and adjusted R^2 with 0.287 and 0.263. Of the five models for Overall Usability, model 15 had highest R^2 and adjusted R^2 with 0.213 and 0.186. Interpreting these results, except for model 4, the models found with the LIFT Machine could describe more of the variability than those made using the WPMA.

Table 2.
Bivariate, linear models by LIFT Machine, scenario 1, training set (N = 31, p < 0.05)

Model	Resp. Var.	Pred.	Coeff.	Const.	F	Sig
7	NAV	lnEFCimg WithSizeFail	-0.497	1.119	9.722	0.004
8	SAT	NAVselfReferential PageFail	-0.683	2.053	5.194	0.030
9	SAT	EFCimg WithSizeFail	-0.104	1.585	7.808	0.009
10	SAT	lnEFCimg WithSizeFail	-0.462	1.529	12.047	0.002
11	OU	NAVselfReferential PageFail	0.857	6.764	5.598	0.025
12	OU	EFCimg WithSizeFail	0.101	7.442	4.575	0.041
13	LNOU	NAVself ReferentialPageFail	0.116	1.903	5.534	0.025
14	LNOU	EFCimg WithSizeFail	0.013	1.996	4.273	0.047
15	LNOU	lnEFCimg WithSizeFail	0.066	2.000	8.107	0.008

Table 3 shows the one model found under Scenario 2 (astronomers.) The R^2 and adjusted R^2 for model 16 were 0.641 and 0.630, which were higher than those found with any model within the category (Navigability) under Scenario 1. Under Scenario 2, no bivariate, linear models appeared with data measured with the WPMA.

Table 3.
Bivariate, linear models by LIFT Machine, scenario 2, training set (N = 31, $p < 0.05$)

Model	Resp. Var.	Pred.	Coeff.	Const.	F	Sig
16	NAV	EFCimgWithSizePass	0.932	0.782	11.762	0.002

With the LIFT Machine, predictors used in model 16 and model 7 are comparable. In each model, the predictor is called EFCimgWithSize. The LIFT Machine counts the number of image tags with attributes for height and width. These attributes control the displayed dimensions of images. A "Pass" is given if the image uses the attributes and a "Fail" if an image does not. This similarity between bivariate, linear models for Navigability is highlighted because predicted ratings are not significantly different between scenarios. These properties also appear as predictors in multivariate models.

The system is designed to apply multiple software tools when testing properties of Web pages. A research question was whether more than one model would appear under different contexts of use. Another research question was whether, within a category of usability, a model may have a predictor that accounts for more of the variability found in the response variable than is found with other models. This study provides evidence that answers “yes” to both of these research questions.

Comparing Predicted to Assessed Ratings for Each Model within Scenario

Application of models listed in Tables 1 through 3 to the prediction set generated predicted ratings. The study found no significant difference (paired t test, $p < 0.05$) between predicted and assessed ratings for any model except for those for Overall Usability. To test how interchangeable models are within the same category, the study found no significant difference (paired t test, $p < 0.05$) between mean predicted ratings for any combination of models in any category that was not Overall Usability.

None of the bivariate models for Overall Usability were interchangeable. Combinations of models 11 and 13, 11 and 14, 11 and 15, 12 and 14, and 12 and 15 had significant differences (paired t test, $p < 0.05$) between means of predicted ratings. (For models predicting the natural log of OU, statistical treatment used the inverse natural log.) This meant that it is not possible to use these models interchangeably, as is possible with bivariate models for Navigability, under Scenario 1.

Comparing Bivariate Models between Scenarios

Navigability is the only category that had bivariate models under both scenarios of contextual use. Comparing models for Navigability, assessed ratings between scenarios were not significantly different (independent samples t test, $t = 0.855$, sig. = 0.394 at $p < 0.05$). Likewise, by independent samples t test, model 4 ($t = -0.043$, sig. =

0.966,) model 5 (t = -1.215, sig. = 0.227,) and model 7 (t = -0.301, sig. = 0.764) for Navigability did not predict ratings under Scenario 1 that were significantly different ($p < 0.05$) from those generated by model 16 under Scenario 2. This concurrence about severity of issues with Navigability allows models to be compared. More information is available in the subsection, “Precision of All Models.”

Models 7 and 16 for NAV used the presence or absence of width and height attributes in image tags as a measured property (EFCimgWithSize). For model 7, as the number of pages containing images without size attributes (“Fail”) increased, the severity of issues with NAV decreased. For model 16, as the number of pages with images that passed this test (“Pass”) increased, severity of issues with NAV increased. These results are different perspectives on the same issue, and the severity of the issue from the perspectives of both scenarios was not different ($p < 0.05$.) However, it is worthwhile to note that comparing accuracies of models for Navigability between scenarios found significant differences. More information is in the subsection about accuracies of models that appears later within this chapter.

Models by Multiple Regressions

The study tests for Overall Usability by using multiple predictors of severity of issues. If predictors affecting the variability of a response variable are not included in the model, R^2 and adjusted R^2 may differ more. If included, the additional predictors may improve the accuracy of predictions [George and Mallery 2008].

Regression analysis found three multivariate models of Overall Usability. The study applied multiple regressions by the stepwise method of testing predictors successively against the response variable [Pallant 2007]. These models are in Table 4. In the stepwise method applied, the probability of F required for a predictor to enter the model is less than or equal to 0.05, and the F required to remove a predictor is greater than or equal to 0.10.

Table 4.
Models built for Overall Usability by multiple regressions (N = 63)

Model	Scenario	Equation
M1	1	OU = 1.249 * LNTAMFS – 0.636*LNIMG – 4.557
M2	1	OU = -1.949*EFCimgWithSizePass – 1.011*NAVbrokenLocalLinksFail + 7.975
M3	2	OU = 0.233 * LNPNBW + 7.235
M4	2	OU = -1.682*EFCimgWithSizePass – 0.618*NAVselfReferentialPageFail + 9.069

Predictors should show independence from one another in a population that is normally distributed. If predictors are described as showing multicollinearity, they are not independent but correlated with one another. Tolerance and VIF (Variance Inflation Factor) are two tests of multicollinearity. A value of 0.10 or less for Tolerance indicates that a predictor may be a combination of other predictors [George and Mallery 2008]. A VIF equal to or greater than 10 also indicates multicollinearity [Pallant 2007]. M1, M2, and M4 did not exhibit multicollinearity. Because M3 is bivariate, multicollinearity was not tested.

As shown in Table 5, values of R^2 (0.417) and adjusted R^2 (0.397) for M1 are similar, and the predictors account for almost 40% of the variability of the response variable. The predictors for M2 account for about 34.5% of the variability of ratings given for Overall Usability. In contrast, the variability of the response variable accounted for by LNPNBW is only 14.5%. Of the multivariate models, M4 has the lowest adjusted R^2 at 29.9%.

Table 5.
Coefficients of determination (R^2) by multivariate regressions
(N = 63, $p < 0.05$, df = 62)

Model	R	R^2	Adj. R^2	Std. Err. Est.	F	Sig.
M1	0.646	0.417	0.397	1.00	21.435	0.000
M2	0.605	0.366	0.345	1.04	17.348	0.000
M3	0.399	0.159	0.145	1.01	11.541	0.001
M4	0.567	0.321	0.299	0.92	14.192	0.000

Comparing Multivariate Models between Scenarios

For Overall Usability, the adjusted R^2 of M1 (0.397) and M2 (0.345) are higher than those of bivariate models under Scenario 1. (Model 15 had the highest adjusted R^2 of the bivariate models at 0.186.) In multivariate models, multiple predictors may account for more of the variability in assessed ratings than bivariate models are able.

Because multivariate models are found for each scenario, the approach tests for significant differences between combinations of predicted and assessed ratings for Overall Usability both within and between scenarios. By independent samples t test, results (in Appendix) show that assessed ratings for Overall Usability are different ($p < 0.05$) between scenarios. Likewise, predicted ratings are different ($p < 0.05$) between scenarios. As was observed with assessed ratings for Overall Usability, the mean of predicted ratings of Overall Usability was higher under Scenario 2 than Scenario 1. This indicated that the evaluation determined that astronomers might find the site to be more usable overall than middle school children might under their scenario.

In contrast, no significant difference ($p < 0.05$) appeared by paired t test between predicted and assessed ratings for Overall Usability (see Appendix) within each scenario. Ratings predicted by M1 or M2 are comparable to assessed ratings under Scenario 1, and comparisons between M3 and M4 to assessed ratings under Scenario 2 have the same result. The means of predicted and assessed ratings within each scenario differed by no more than 0.01.

Precisions of All Models

Table 6 lists results for both overall precision and precision at threshold for models found by simple linear regression. The study tests for models with the highest

Table 6.
Precisions of bivariate models with prediction set (N = 31)

Model	**Threshold**	**Overall Precision**		**Precision At Threshold**	
		Pairs At or Above	**Pairs Both Below**	**Pairs At or Above**	**Pairs Below**
OU					
1	7.554	*0.323	0.290	0.526	*0.750
11	7.759	0.032	0.452	0.167	0.560
12	7.680	0.129	*0.484	0.500	0.652
LNOU					
13	2.038	0.032	0.452	0.167	0.560
14	2.027	0.161	*0.484	*0.625	0.652
15	2.019	0.258	0.355	0.533	0.688
FLX					
2	0.670	0.258	0.258	0.500	0.533
3	0.649	*0.355	*0.355	*0.688	*0.733
NAV					
4	1.022	0.258	*0.452	0.667	*0.737
5	0.830	0.290	0.355	0.563	0.733
7	0.980	*0.419	0.258	*0.813	0.533
SAT					
6	1.440	0.161	*0.419	0.500	0.619
8	1.260	*0.516	0.129	*0.640	0.667
9	1.340	0.419	0.226	0.565	*0.875
10	1.400	0.290	0.323	0.563	0.667
NAV (Scenario 2)					
16	*0.993	*0.226	*0.516	*1.000	*0.696

precision. Bivariate models of Tables 1 through 3 are listed together. For precision at threshold, ratios were generally higher than for overall precision, of which few were as high as 50%.

As shown in Table 6, the kinds of precision used in this study varied across models. If more than one model was built in a category, no one model had highest values for both kinds of precision. In Table 6 and Table 7, an asterisk indicates highest value for the category and scenario. Table 7 has precisions for multivariate models. These observations apply to the site tested, and testing of other sites might yield different results. The double asterisk indicates a bivariate model in Table 7.

Table 7.
Precisions of models of Overall Usability by multiple regressions (N = 63)

Model	**Threshold**	**Overall Precision**		**Precision At Threshold**	
		Pairs At or Above	**Pairs Both Below**	**Pairs At or Above**	**Pairs Below**
Scenario 1					
M1	7.560	0.365	*0.381	0.742	*0.750
M2	6.558	*0.683	0.143	*0.896	0.600
Scenario 2					
**M3	0.809	0.429	*0.270	*0.750	0.603
M4	0.809	*0.460	0.222	0.707	*0.636

Accuracies of All Models

Accuracy is a measure of how well a model predicts actual values. The closer the mean of absolute values for unstandardized residuals is to zero, the higher the accuracy of the model [Badi, Bae, Moore, Meintanis, Zacchi, Hsieh, Shipman and Marshall 2006]. As shown in Table 8, of all models under Scenario 1, the most accurate was model 3 for Flexibility. This model had a mean of 0.48190. M1 was the most accurate among models for Overall Usability, but none of the models built for Overall Usability were more accurate than models found for other categories of usability. The accuracy of M4 under Scenario 2 was higher than that of M1. The accuracy of model 16 for Navigability under Scenario 2 was much higher than that of the most accurate model for Navigability (model 4, mean of 0.73889) under Scenario 1. In Table 8, asterisks indicate use of inverse natural log of absolute value of residuals.

Table 8.
Models with highest accuracies (N = 63)

Category of Usability	Model	Software Tool	Unstd. Residuals	
			Mean	Std.Dev.
Scenario 1				
FLX	3	WPMA	0.48190	0.38866
SAT	10	LIFT	0.71485	0.46273
NAV	4	WPMA	0.73889	0.53761
OU	M1	WPMA	*0.79143	*0.57619
Scenario 2				
NAV	16	LIFT	0.42763	0.34002
OU	M4	LIFT	*0.68513	*0.58191

For Navigability, model 16 (0.42763) was more accurate than model 4 (0.73889). This meant that the LIFT Machine provided a more accurate model under Scenario 2 than did the WPMA under Scenario 1. Table 8 lists the most accurate models by category in descending order of accuracy, e.g. the most accurate have the lowest means.

Testing for significant differences between models within categories, analysis revealed several characteristics. Using independent samples t-test, the mean accuracies of all models of Navigability under Scenario 1 were significantly different ($p < 0.05$) from that of model 16 under Scenario 2. Within categories under the same scenario by paired t test with P-value of 0.05, 8 of 21 unique combinations of models for Overall Usability did not have mean accuracies that were significantly different. Likewise, the accuracies of the one combination of two models found for Flexibility were different.

For Navigability, none of the accuracies from all combinations of three models were different. Finally, six of ten combinations of models for Satisfaction did not have significantly different mean accuracies. Under Scenario 2, the accuracies of M3 and M4 were not different ($p < 0.05$.) The consequence of these findings is that not all predictors in models for Overall Usability and Satisfaction will yield comparable accuracies within their categories. While there was no significant difference between accuracies for Navigability within Scenario 1, there were significant differences within the category between scenarios.

Principal Component Analysis

Principle component analysis (PCA) is an approach to describe variability within a set of scalar data [Pallant 2007; Tabachnick and Fidell 2007]. The five categories of usability assessed had scalar ratings (0–5.) Application of PCA did not include Overall

Usability. Although assessed ratings for OU are scalar, ratings are on a scale (1–10) that is different from that of the categories. Assessed ratings for OU are from very poor to very good while ratings in categories are from no detectable issues to very severe issues. The approach applied PCA to find a percentage of variability accounted for by the five categories. Under Scenario 1, the one component found accounted for 67.46% of variability in assessed ratings of severity of issues. Under Scenario 2, the approach again only found a single component, but it accounted for more, 72.43%, of the variability.

Precisions Plotted

Figures 4-6 illustrate bivariate models plotted against the prediction set. Selections of models apply to their category and show distributions of assessed and predicted ratings as well as thresholds. Figure 4 is of a model with highest likelihood using overall precision. Figure 5 is the plot of a model with highest likelihood using precision at threshold. Figure 6 shows a model that had highest likelihoods on both sides of a threshold. More figures are available in the Appendix.

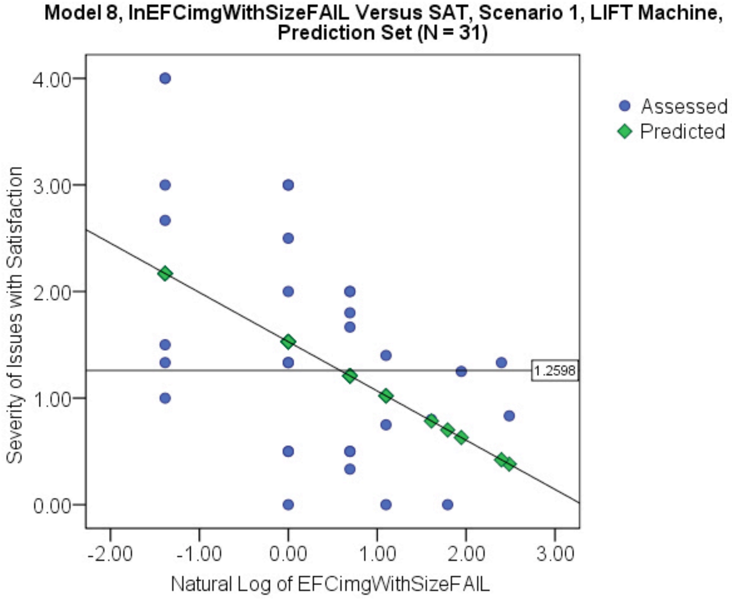


Figure 4. Highest overall precision above threshold for category SAT.

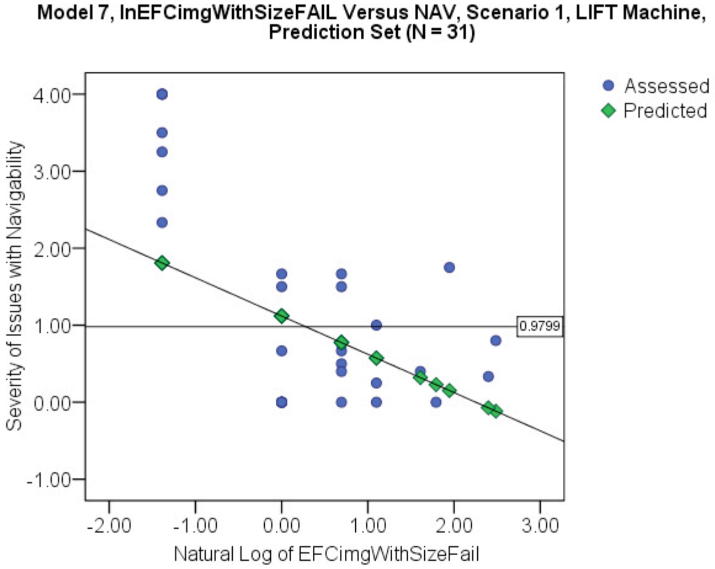


Figure 5. Highest precision at threshold above threshold for category NAV.

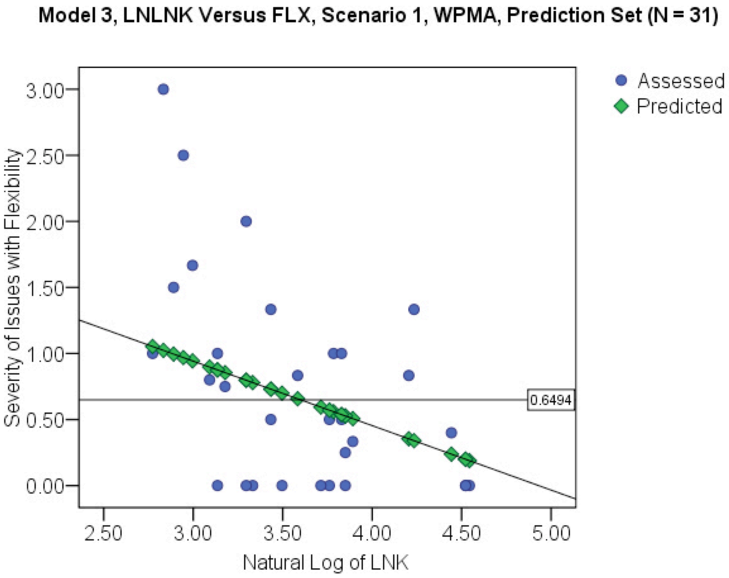


Figure 6. Highest precisions on both sides of threshold for any category.

CHAPTER VIII

DISCUSSION

Results showed that the approach used with the semi-automated system could find models under different scenarios of contextual use. The approach demonstrated advantages of using more than one software tool in analysis of user interfaces of informational HTML pages. Under Scenario 1 (middle school children), models predicted ratings for severity of usability issues in categories of Flexibility, Navigability, and Satisfaction. Under Scenario 2 (graduate-level astronomers,) the one model built predicted severity of issues with Navigability. The study applied multiple regression to ratings of Overall Usability and build models for each scenario. This study may benefit Web site managers who seek to identify Web pages that are likely to have (or not have) usability concerns above (or below) a threshold rating. They may choose to use different models to test how people might use their site under different contexts of use.

A question arose about how findings of usefulness of Web-based documents rated under document triage compared to results of the study. Document triage "is the practice of quickly determining the merit and disposition of relevant documents" (p. 130) [Bae et al. 2005]. It appears that quickly assessing usefulness and predicting usability of Web pages are approaches with different objectives. The former concerns content and the latter use of style.

The results of this study differ and are similar to those of Bae et al [Bae, Badi, Meintanis, Moore, Zacchi, Hsieh, Marshall and Shipman 2005]. Stylistic properties correlated with usability ratings in our study but did not appear to do so in theirs. The approaches used to test for “usable” and “useful” are different. It is worthwhile to note that the studies did not use the same set of documents or same groups of participants.

Briefly, Bae et al [Bae, Badi, Meintanis, Moore, Zacchi, Hsieh, Marshall and Shipman 2005] noted a positive relationship (Pearson coefficient of 0.397, $p < 0.05$) between total word count and page count with ratings given to documents. They reported no significant correlations at $p < 0.05$ between scores and counts of hypertext links, images, or file sizes. During interviews, two subjects reported that the length of documents influenced their scoring.

Results from this study with Overall Usability show similarities to those of Bae et al [Bae, Badi, Meintanis, Moore, Zacchi, Hsieh, Marshall and Shipman 2005], but the differences are notable. In our study, a positive association (Pearson coefficient of 0.368, $p < 0.05$) appeared between Overall Usability and non-body words (LNPNBW) under Scenario 1. However, no relationships appeared between assessed ratings and Total Words (TW) or Body Words (BW.) During triage, users sought documents with useful content. Participants in this study could seek pages with content that would answer questions presented at the start of experiments. However, they did not have an option to rate the usefulness of content of pages for answering those questions. In this study, participants did not associate amounts of textual content with usability issues for the site tested. Visual inspection of pages showed that many contained images and few, such as a glossary, consisted mainly of textual content.

In the scenario applied by Bae et al [Bae, Badi, Meintanis, Moore, Zacchi, Hsieh, Marshall and Shipman 2005], no significant relationships ($p < 0.05$) appeared between non-textual properties (number of hypertext links, images, and sizes of files) and ratings of usefulness. In contrast, this study found that counts of images (EFCimgWithSizePass) and broken links (NAVbrokenLocalLink under Scenario 1 and NAVselfReferentialPageFail under Scenario 2) correlated negatively with Overall Usability. If such properties do not affect the usefulness of document, they could affect how usable it is. However, the relationship between increasing numbers of images and declining ratings for Overall Usability and Navigability was unexpected until visual inspection of pages revealed that most pages contained images that linked to other pages. The multivariate models also showed a similar negative relationship between Overall Usability and number of images (EFCimgWithSizePass).

Document triage and usability assessment may reflect two different viewpoints about what enhances or impedes productive interaction with documents (which includes Web pages.) This study built models that were different in number, precision, and accuracy under different scenarios of contextual use. Based on results from this early study, future research would test for other multivariate models, scenarios, and Web sites.

CHAPTER IX

CONCLUSIONS

Software tools are useful for finding exceptions to rules about the usability of Web sites. Software professionals such as Web site managers may find these tools useful when deciding whether to change the amount of content or style of pages in order to improve experiences people have when visiting the sites. As the number of Web pages grows ever larger in sites, those who must find Web pages that are less usable are likely to find tools that list exceptions to rules helpful. However, concerns people have about usability of Web pages may differ. Might one group find a page very difficult to use while another experiences no difficulty? How well might automated analysis detect Web pages that are problematic to some groups but not to others on the basis of their informational needs? This research approached these questions by finding models that represent relationships between usability concerns and measurements taken by automated tools. Findings of this research support the concept that models may represent common as well as different concerns about usability by groups having different informational needs from a Web site.

The approach of the study designed, built, and tested a system that found models by methods of statistical analysis. The system collects measurements taken with software tools about Web pages and uses the data to find relationships with ratings of usability issues given in usability assessments of those pages. The current system requires human intervention to carry out assessments and statistical analysis. The design of the system accommodates automation by allowing use of different software tools to measure properties of Web pages. Heuristic assessments were performed under controlled experimental conditions, and the study applied different scenarios of use. This study built models that were different in number, coefficients of determination, precision, and accuracy under different scenarios of contextual use. Web site managers may find the approach beneficial when testing for Web pages that need inspection and possible rework for usability issues. Future studies may apply the approach to test other Web sites, automate more of the system, and build other types of models.

REFERENCES

AL-QAIMARI, G. AND MCROSTIE, D. 1999. KALDI: A CAUsE Tool for Supporting Testing and Analysis of User Interaction. In *People and Computers XV - Interaction without Frontiers, Joint Proceedings of HCI 2001 and IHM 2001*, 153–169.

APPLE 2010. Apple Inc, Cupertino, CA. http://www.apple.com (Retrieved October 8, 2010).

ARLOW, J. AND NEUSTADT, I. 2005. *UML 2 and the Unified Process:Practical Object-Oriented Analysis and Design*. Addison-Wesley, Upper Saddle River, NJ.

ATRC 2010. A-Prompt Web Accessibility Verifier Adaptive. http://www.aprompt.ca/ (Retrieved October 1, 2010).

ATTERER, R. 2008. Model-based automatic usability validation: a tool concept for improving web-based UIs. In *Proceedings of the 5th Nordic Conference on Human-computer Interaction: Building Bridges.* 13–22.

ATTERER, R., SCHMIDT, A. AND HUSSMAN, H. 2006. Extending web engineering models and tools for automatic usability validation. *Journal of Web Engineering 5*, 1, 43–64.

BADI, R., BAE, S., MOORE, J.M., MEINTANIS, K., ZACCHI, A., HSIEH, H., SHIPMAN, F. AND MARSHALL, C.C. 2006. Recognizing user interest and document value from reading and organizing activities in document triage. In *Proceedings of the 11th International Conference on Intelligent User Interfaces*. 218–225.

BAE, S., BADI, R., MEINTANIS, K., MOORE, J.M., ZACCHI, A., HSIEH, H., MARSHALL, C.C. AND SHIPMAN, F.M. 2005. Effects of Display Configurations on Document Triage. In *Human-Computer Interaction - INTERACT 2005*, M.F. Constabile and F. Patern, Eds. Springer-Verlag, Berlin, Germany, 130–143.

BAEZA-YATES, R. AND RIBEIRO, B.D.A.N. 1999. *Modern Information Retrieval*. ACM Press, New York, NY.

BEIREKDAR, A., KEITA, M., NOIRHOMME-FRAITURE, M., RANDOLET, F., VANDERDONCKT, J. AND MARIAGE, C. 2005. Flexible reporting for automated usability and accessibility evaluation of web sites. In *Human-Computer Interaction - INTERACT 2005, Lecture Notes in Computer Science*, M.F. Constabile and F. Paterno, Eds. Springer-Verlag, Berlin, Germany, 281–294.

BEIREKDAR, A., VANDERDONCKT, J. AND NOIRHOMME-FRAITURE, M. 2002. A framework and a language for usability automatic evaluation of web sites by static analysis of html source code. In *4th International Conference on Computer-Aided Design of User Interfaces (CADUI'2002)*. 15–17.

BEIREKDAR, A., VANDERDONCKT, J. AND NOIRHOMME-FRAITURE, M. 2003. KWARESMI - Knowledge-based Web Automated Evaluation with REconfigurable guidelineS optiMIzation. In *2nd International Conference on Universal Access in Human-Computer Interaction (UAHCI'2003)*. 1504–1508.

BISHOP, C. 2006. *Pattern Recognition and Machine Learning*. Springer, New York, NY.

BLACKMON, M.H., POLSON, P.G., KITAJIMA, M. AND LEWIS, C. 2002. Cognitive walkthrough for the web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing Our world, Changing Ourselves*. 463–470.

BMC 2010. *BMC Software, Inc.*, Houston, TX. http://www.bmc.com (Retrieved September 24, 2010).

BRAJNIK, G. 2000a. Automatic web usability evaluation: What needs to be done? In *Proceedings of the Sixth Conference on Human Factors & the Web*, P. Kortum and E. Eudzinger Eds. University of Texas, Austin, TX.

BRAJNIK, G. 2000b. Automatic web usability evaluation: Where is the limit? In *Proceedings of the Sixth Conference on Human Factors & the Web*. P. Kortum and E. Eudzinger Eds. University of Texas, Austin, TX.

BRAJNIK, G. 2004. Using automatic tools in accessibility and usability assurance processes. In *UI4All LNCS*, C. Stary and C. Stephanidis, Eds. Springer-Verlag, Berlin, Germany, 219–234.

BYRNE, M.D., WOOD, S.D., FOLEY, J.D., KIERAS, D.E. AND SUKAVIRIYA, P.N. 1994. Automating interface evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating Interdependence*. 232–237.

CARD, S.K., MORAN, T.P. AND NEWELL, A. 1980. The keystroke-level model for user performance time with interactive systems. *Commun. ACM 23*, 7, 396–410.

CARD, S.K., MORAN, T.P. AND NEWELL, A. 1983. *The Psychology of Human-Computer Interaction*. Lawrence Elbaum Associates, Inc., Hillsdale, NJ.

CERI, S., DANIEL, F., MATERA, M. AND FACCA, F.M. 2007. Model-driven development of context-aware Web applications. *ACM Trans. Internet Technol. 7*, 1, 1–33.

CERI, S., FRATERNALI, P. AND BONGIO, A. 2000. Web modeling language (WebML): a modeling language for designing web sites. *Computer Networks 33*, 1–6, 137–157.

CHI, E.H., PIROLLI, P. AND PITKOW, J. 2000. The scent of a site: a system for analyzing and predicting information scent, usage, and usability of a Web site. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, The Hague, The Netherlands. 161–168.

CHI, E.H., ROSIEN, A. AND HEER, J. 2003. LumberJack: intelligent discovery and analysis of web user traffic composition. In *WEBKDD 2002 - Mining Web Data for Discovering Usage Patterns and Profiles*, B.M. Masand, M. Spiliopoulou, J. Srivastava, and O.R. Zaiane Eds. Springer, Edmonton, Canada, 1–16.

CHI, E.H., ROSIEN, A., SUPATTANASIRI, G., WILLIAMS, A., ROYER, C., CHOW, C., ROBLES, E., DALAL, B., CHEN, J. AND COUSINS, S. 2003. The bloodhound project: automating discovery of web usability issues using the InfoScent™ simulator. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 505–512.

COOPER, A., REIMANN, R. AND CRONIN, D. 2007. *About face 3 : the Essentials of Interaction Design*. Wiley, Indianapolis, IN.

CROFT, W.B., METZLER, D. AND STROHMAN, T. 2009. *Search Engines : Information Retrieval in Practice*. Addison-Wesley, Boston, MA.

DE VEAUX, R.D., VELLEMAN, P.F. AND BOCK, D.E. 2008. *Intro Stats*. Addison-Wesley, Reading, MA.

DEITEL, H.M. 2001. *XML : How to Program*. Prentice Hall, Upper Saddle River, NJ.

DIX, A. 1998. *Human-computer Interaction*. Prentice Hall, New York, NY.

FINLAY, J. AND DIX, A. 2002. *An Introduction to Artificial Intelligence*. Routledge, London, England.

FORD, S. AND BOYARSKI, D. 1997. Design @ Carnegie Mellon: a web story. In *Proceedings of the 2nd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques.* 121–124.

FOWLER, M. 2004. *UML Distilled : a Brief Guide to the Standard Object Modeling Language*. Addison-Wesley, Boston, MA.

GEORGE, D. AND MALLERY, P. 2008. *SPSS for Windows Step by Step : a Simple Guide and Reference, 15.0 Update*. Pearson, Boston, MA.

GULLI, A. AND SIGNORINI, A. 2005. The indexable web is more than 11.5 billion pages. In *Proceedings of the Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*. 902–903.

HELFRICH, B. AND LANDAY, J.A. 1999. QUIP: Quantitative user interface profiling. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.37.2367 (Retrieved October 7, 2010).

HONG, J.I. AND LANDAY, J.A. 2001. WebQuilt: a framework for capturing and visualizing the web experience. In *Proceedings of the 10th International Conference on the World Wide Web*. 717–724.

HULME, G.V. 2000. BMC Integrates Site-Monitoring Tool. http://www.informationweek.com/807/bmc.htm (Retrieved September 24, 2010).

IADAS 2010. The Webby Awards International Academy of Digital Arts and Sciences. http://www.iadas.com (Retrieved October 7, 2010).

IBM. International Business Machines Corp., Armonk, NY. http://www.ibm.com (Retrieved October 7, 2010).

IBM 2010. WebSphere Application Server v4.0x. In *6.6.21: Performing tasks with the resource analyzer*. http://publib.boulder.ibm.com/infocenter/wasinfo/v4r0/index.jsp?topic=/com.ibm.websphere.v4.doc/wasa_content/060621.html (Retrieved October 7, 2010).

IEEE-SA 2008. *IEEE Standards Definition Database*. http://dictionary.ieee.org (Retrieved October 7, 2010).

ISO 1998. 9241-11:1998(E). In *Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: guidance on usability.* International Organization for Standardization, Geneva, Switzerland, 22.

ITAW 2010. Section 508, U.S. General Services Administration, Washington, D.C. http://www.section508.gov (Retrieved October 7, 2010).

IVORY, M.Y. 2000. Web TANGO: towards automated comparison of information-centric web site designs. In *Proceedings of the CHI '00 Extended Abstracts on Human Factors in Computing Systems*. 329–330.

IVORY, M.Y. 2003. *Automated Web Site Evaluation : Researchers' and Practitioners' Perspectives*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

IVORY, M.Y. AND CHEVALIER, A. 2002. *UW-CSE-02-10-01: A Study of Automated Web Site Evaluation Tools.* University of Washington.

IVORY, M.Y. AND HEARST, M.A. 2001. The state of the art in automating usability evaluation of user interfaces. *ACM Comput. Surv. 33*, 4, 470–516.

IVORY, M.Y. AND HEARST, M.A. 2002. Statistical profiles of highly-rated web sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing Our World, Changing Ourselves*. 367–374.

IVORY, M.Y. AND MEGRAW, R. 2005. Evolution of web site design patterns. *ACM Trans. Inf. Syst. 23*, 4, 463–497.

IVORY, M.Y., SINHA, R.R. AND HEARST, M.A. 2000. Preliminary findings on quantitative measures for distinguishing highly rated information-centric web pages. In *Proceedings of the Sixth Conference on Human Factors & the Web*. http://webtango.berkeley.edu/papers/hfw00/hfw00.ps (Retrieved January 23, 2010).

IVORY, M.Y., SINHA, R.R. AND HEARST, M.A. 2001. Empirically validated web page design metrics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 53–60.

JOHN, B.E. AND KIERAS, D.E. 1996. The GOMS family of user interface analysis techniques: comparison and contrast. *ACM Trans. Comput.-Hum. Interact. 3*, 4, 320–351.

KAPITSAKI, G.M., KATEROS, D.A., PREZERAKOS, G.N. AND VENIERIS, I.S. 2009. Model-driven development of composite context-aware web applications. *Inf. Softw. Technol. 51*, 8, 1244–1260.

KIERAS, D. 1997. A guide to GOMS model usability evaluation using NGOMSL. In *Handbook of Human-Computer Interaction*, M. Helander, T.K. Landauer and P. Prabhu, Eds. Elsevier, Amsterdam, The Netherlands, 733–766.

KIERAS, D.E., WOOD, S.D., ABOTEL, K. AND HORNOF, A. 1995. GLEAN: a computer-based tool for rapid GOMS model usability evaluation of user interface designs. In *Proceedings of the 8th Annual ACM Symposium on User Interface and Software Technology*. 91–100.

LIMBOURG, Q. AND VANDERDONCKT, J. 2004. UsiXML: A user interface description language supporting multiple levels of independence. In *Engineering Advanced Web Applications*, M. Matera and S. Coma, Eds. Rinton Press, Paramus, NJ, 325–338.

LYNCH, P.J. AND HORTON, S. 2008. *Web Style Guide : Basic Design Principles for Creating Web Sites*. Yale University Press, New Haven, CT.

MA, J., ZHANG, Z. AND GARCIA, R. 2007. Automatically determining web site comprehensibility. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1027758## (Retrieved October 8, 2010).

MANNING, C.D., RAGHAVAN, P. AND SCHÈUTZE, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, NY.

MERRIAM-WEBSTER INC. 2003. *Merriam-Webster's Collegiate Dictionary*. Merriam-Webster, Springfield, MA.

MICROSOFT 2010. Microsoft, Redmond, WA. http://www.microsoft.com (Retrieved October 8, 2010).

NETER, J. 1996. *Applied Linear Regression Models*. Irwin, Chicago, IL.

NETER, J. AND WASSERMAN, W. 1974. *Applied Linear Statistical Models; Regression, Analysis of Variance, and Experimental Designs*. R.D. Irwin, Homewood, IL.

NETRAKER 2010. NetRaker experience recorder. NetRaker Corp., Mountain View, CA. http://www.netraker.com/nrinfo/products/nrer.asp (Retrieved September 24, 2010).

NIELSEN, J. 1992. Finding usability problems through heuristic evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 373–380.

NIELSEN, J. 1993. *Usability Engineering*. Academic Press, Boston, MA.

NIELSEN, J. 1994. Usability inspection methods. In *Proceedings of the Conference Companion on Human Factors in Computing Systems*, Boston, MA, 413–414.

NIELSEN, J. 2000. *Designing Web Usability*. New Riders, Indianapolis, IN.

NIELSEN, J. 2005. Ten usability heuristics. http://www.useit.com/papers/heuristic/heuristic_list.html (Retrieved September 24, 2010).

NNG 2007. LIFT machine, Nielsen-Norman Group, Usablenet Inc, New York, NY. http://www.usablenet.com (Retrieved August 20, 2007).

NORMAN, D.A. 2002. *The Design of Everyday Things*. Basic Books, New York, NY.

OMG 2010. *Object Management Group, Inc.*, Needham, MA. http://www.omg.org (Retrieved January 23, 2010).

PALLANT, J. 2007. *SPSS Survival Manual : a Step by Step Guide to Data Analysis Using SPSS for Windows*. McGraw-Hill, Maidenhead, England.

PRESSMAN, R.S. 2005. *Software Engineering : a Practitioner's Approach*. McGraw-Hill, Boston, MA.

ROCHE, X. 2007. HTTrack XR&CO, Paducah, KY. http://www.httrack.com (Retrieved August 20, 2007).

RUSSELL, S.J. AND NORVIG, P. 2003. *Artificial Intelligence : a Modern Approach*. Prentice Hall, Upper Saddle River, NJ.

SCHWABE, D. AND ROSSI, G. 1998. Developing hypermedia applications using OOHDM. In *Proceedings of the Workshop on Hypermedia Development Process, Methods and Models (Hypertext'98*). 1–20.

SHANNON, D.M. AND DAVENPORT, M.A. 2001. *Using SPSS® to Solve Statistical Problems : a Self-instruction Guide*. Prentice Hall, Upper Saddle River, NJ.

SHARP, H., ROGERS, Y. AND PREECE, J. 2007. *Interaction Design : Beyond Human-computer Interaction*. Wiley, Hoboken, NJ.

SHNEIDERMAN, B. 1997. *Designing the User Interface: Strategies for Effective Human-computer Interaction*. Addison-Wesley, Boston, MA.

SHNEIDERMAN, B. AND PLAISANT, C. 2010. *Designing the User Interface : Strategies for Effective Human-computer Interaction*. Addison-Wesley, Boston, MA.

SPSS 2006. SPSS 15.0 for Windows. SPSS, Inc., Chicago, IL. http://www.spss.com (Retrieved October 8, 2010).

STSCI 2007. HubbleSite Space Telescope Science Institute, Baltimore, MD. http://www.stsci.edu (Retrieved August 20, 2007).

STSCI 2010. Astronomy Resources at STScI Space Telescope Science Institute (STScI), Baltimore, MD. http://www.stsci.edu/resources/ (Retrieved October 8, 2010).

SUCHMAN, L.A. 1987. *Plans and Situated Actions : the Problem of Human-machine Communication*. Cambridge University Press, New York, NY.

TABACHNICK, B.G. AND FIDELL, L.S. 2007. *Using Multivariate Statistics*. Pearson, Boston, MA.

TIDWELL, J. 2006. *Designing Interfaces*. O'Reilly, Sebastpol, CA.

TIEDTKE, T., MARTIN, C. AND GERTH, N. 2002. AWUSA - a tool for automated website usability analysis. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.83.7285 (Retrieved October 8, 2010).

TROYER, O.M.F.D. AND LEUNE, C.J. 1998. WSDM: A user centered design method for web sites. In *Proceedings of the 7th International World Wide Web Conference*. 85-94.

VANDERDONCKT, J. AND BEIREKDAR, A. 2005. Automated web evaluation by guideline review. *Journal of Web Engineering 4*, 2, 102-117.

W3C 1999. W3C Web Content Accessibility Guidelines 1.0 W3C Recommendation 5-May-1999. http://www.w3.org/TR/WCAG10/ (Retrieved October 8, 2010).

W3C 2003. Extensible Markup Language (XML). http://www.w3.org/XML/ (Retrieved October 8, 2010).

W3C 2005. Essential Components of Web Accessibility. http://www.w3.org/WAI/intro/components.php (Retrieved October 8, 2010).

W3C 2006. Complete List of Web Accessibility Evaluation Tools. http://www.w3.org/WAI/RC/tools/complete (Retrieved October 8, 2010).

W3C 2008. Web Content Accessibility Guidelines (WCAG) Overview, S.L. HENRY Ed. World Wide Web Consortium, Cambridge, MA.

W3C 2009. About W3C World Wide Web Consortium. http://www.w3.org/Consortium/ (Retrieved October 8, 2010).

W3C 2010. World Wide Web Consortium. http://www.w3c.org (Retrieved October 8, 2010).

YAN, P., ZHANG, Z. AND GARCIA, R. 2007. Automatic website comprehensibility evaluation. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. 191-197.

APPENDIX A

Terminology

Term	Description
(the) approach	Application of the system in this study.
heuristic	“Involving or serving as an aid to learning, discovery, or problem-solving by experiments and especially trial-and-error methods; also: of or relating to exploratory problem-solving techniques that utilize self-educating techniques (as the evaluation of feedback) to improve performance.” [Merriam-Webster Inc. 2003]
participant	Person who performs evaluations of user interfaces.
practitioner	Person engaged as a designer, webmaster, usability engineer, and other professionals who design, build, or maintain Web sites [Ivory 2003].
rule of thumb	“1 : a method of procedure based on experience and common sense. 2 : a general principle regarded as roughly correct but not intended to be scientifically accurate.” [Merriam-Webster Inc. 2003]
specification	“A set of conditions and requirements of precise and limited application that provide a detailed description of a procedure, process, material, product, or service for use primarily in procurement and manufacturing.” [IEEE-SA 2008]
UML™	Unified Modeling Language. The Unified Modeling Language through the Object Management Group (OMG) allows specification, visualization, and documentation of software systems [Arlow and Neustadt 2005; OMG 2010].
Web site manager	Person who is a manager of practitioners and who may also at times be a practitioner.
XML	Extensible Markup Language. An Extensible Markup Language derived from SGML (Standard Generalized Markup Language) designed to provide flexibility in publishing as well as exchange of data, such as over the Web [Deitel 2001; W3C 2003].

APPENDIX B

Skewness of Assessed Ratings for Both Scenarios, N = 63

Usability	Min	Max	Mean	Median	Skewness	
					Statistic	Std. Error
Scenario No. 1						
Effectiveness	0.0	2.5	1.036	1.000	0.240	0.302
Flexibility	0.0	3.0	0.696	0.500	1.032	0.302
Navigability	0.0	4.0	1.022	0.500	1.296	0.302
Satisfaction	0.0	4.0	1.385	1.000	0.789	0.302
Efficiency	0.0	4.0	1.200	1.000	0.793	0.302
Overall Usability	4.0	10.0	7.569	7.500	-0.388	0.302
Scenario No. 2						
Effectiveness	0.0	4.0	0.691	0.600	2.400	0.302
Flexibility	0.0	4.0	0.809	0.600	1.820	0.302
Navigability	0.0	2.5	0.871	0.750	0.614	0.302
Satisfaction	0.0	4.5	1.235	1.000	1.651	0.302
Efficiency	0.0	4.0	0.983	1.000	2.151	0.302
Overall Usability	4.5	10.0	8.087	8.500	-1.060	1.291

Normality of Assessed Ratings, N = 63

Usability	Mean	Median	Std. Dev.	Var.	Kolmogorov-Smirnov Test	
					Statistic, df = 63	Sig. (2-tailed)
Scenario No. 1						
Effectiveness	1.036	1.000	0.706	0.499	0.128	0.012
Flexibility	0.696	0.500	0.693	0.480	0.158	0.001
Navigability	1.022	0.500	1.221	1.491	0.212	0.000
Satisfaction	1.385	1.000	1.012	1.023	0.172	0.000
Efficiency	1.200	1.000	0.938	0.879	0.100	0.188
Overall Usability	7.569	7.500	1.288	1.659	0.099	0.199
Scenario No. 2						
Effectiveness	0.691	0.600	0.645	0.416	0.157	0.001
Flexibility	0.809	0.600	0.791	0.626	0.182	0.000
Navigability	0.871	0.750	0.673	0.453	0.154	0.001
Satisfaction	1.235	1.000	0.866	0.750	0.189	0.000
Efficiency	0.983	1.000	0.759	0.576	0.205	0.000
Overall Usability	8.087	8.500	1.096	1.201	0.171	0.000

Skewness of Properties Measured by WPMA (N = 63)

Property	Min	Max	Mean	Median	Skewness	Kurtosis
TW	6	3157	297.490	208.000	5.206	33.154
BW	0	2901	215.030	113.000	5.306	33.762
NBW	6	284	82.460	60.000	1.353	1.658
EW	0	98	30.190	31.000	1.079	2.945
CLU	1	17	5.300	5.000	1.470	3.745
NFL	0	24	2.350	0.000	3.312	12.717
LNK	16	101	41.490	34.000	1.376	0.776
IMG	1	66	24.030	13.000	0.640	-1.259
PEBW	0	194.118	36.115	20.202	1.656	2.973
PNBW	0	1188.235	105.451	50.732	4.713	28.673
TIFS	670	154990	35819.950	23508.000	1.504	1.886
TAMFS	17292	234776	70009.860	56858.000	1.622	4.174
WPFS	7515	53760	14717.080	13554.000	3.219	16.353
PIWPFS	0.036	0.923	0.594	0.635	-0.963	0.292

Standard error for Skewness of all measured properties was 0.302. Standard error for Kurtosis of all measured properties was 0.595.

Normality of Measurements of Properties Measured by WPMA (N = 63)

Property	**Mean**	**Median**	**Std. Deviation**	**Variance**	**Kolmogorov-Smirnov**	
					Statistic, df = 63	**Sig. (2-tailed)**
TW	297.49	208.00	427.052	182373.3	0.247	0.000
BW	215.00	113.00	399.637	159709.5	0.295	0.000
NBW	82.46	60.00	62.527	3909.672	0.225	0.000
EW	30.19	31.00	18.475	341.318	0.091	0.200*
CLU	5.30	5.00	2.938	8.633	0.216	0.000
NFL	2.35	0.00	4.393	19.295	0.296	0.000
LNK	41.49	34.00	23.209	538.641	0.216	0.000
IMG	24.03	13.00	22.099	488.386	0.209	0.000
PEBW	36.12	20.20	39.510	1561.02	0.205	0.000
PNBW	105.45	50.73	167.040	27901.52	0.264	0.000
TIFS	35819.95	23508.00	35239.265	1E+009	0.235	0.000
TCPS	19472.83	25709.00	14889.957	2E+008	0.244	0.000
TAMFS	70009.86	56858.00	38600.099	1E+009	0.154	0.001
WPFS	14717.08	13554.00	6837.920	5E+007	0.162	0.000
PIWPFS	0.59	0.64	0.218	0.047	0.150	0.001

An asterisk indicates a significance of at least 0.200.

Skewness in Effectiveness Measured by LIFT Machine (N = 63)

Property	Min	Max	Mean	Median	Skewness	Kurtosis
EFFspell_checkerFAIL	1	103	7.600	5.000	6.626	46.357

The standard error for skewness was 0.302 and for kurtosis was 0.595.

Normality of Measurements in Effectiveness, LIFT Machine (N = 63)

Property	Mean	Median	Std. Dev.	Var.	Kolmogorov-Smirnov	
					Statistic	Sig (2-tailed)
EFFspell_checkerFAIL	7.600	5.000	13.012	169.308	0.314	0.000

Skewness in Navigability Measured by LIFT Machine (N = 63)

Property	Min	Max	Mean	Median	Skewness	Kurtosis
NAVbrokenExternalLinks PASS	0	1	0.90	1.00	-2.825	6.179
NAVbrokenExternalLinks FAIL	0	14	0.30	0.00	7.661	59.925
NAVemptyAnchorLinks PASS	0	1	0.90	1.00	-2.825	6.179
NAVemptyAnchorLinks FAIL	0	7	0.24	0.00	5.680	35.940
NAVbrokenLocalLinks PASS	0	1	0.90	1.00	-2.825	6.179
NAVbrokenLocalLinks FAIL	0	1	0.10	0.00	2.825	6.179
NAVmissingLogicalPath PASS	0	1	0.95	1.00	-4.353	17.502
NAVmissingLogicalPath FAIL	0	1	0.05	0.00	4.353	17.502
NAVselfReferentialPage PASS	0	1	0.05	0.00	4.353	17.502
NAVselfReferentialPage FAIL	0	2	1.16	1.00	0.429	0.778
NAVstdLinksFontAndColor PASS	0	1	0.86	1.00	-2.091	2.451
NAVstdLinksFontAndColor FAIL	0	2	0.29	0.00	2.091	2.451

The standard error of skewness was 0.302 and Kurtosis was 0.595.

Normality of Measurements in Navigability by LIFT Machine (N = 63)

Property	Mean	Median	Standard Deviation	Var.	Kolmogorov-Smirnov	
					Statistic	Sig (2-tailed)
NAVbrokenExternalLinks PASS	0.90	1.00	0.296	0.088	0.531	0.000
NAVbrokenExternalLinks FAIL	0.30	0.00	1.775	3.150	0.472	0.000
NAVemptyAnchorLinks PASS	0.90	1.00	0.296	0.088	0.531	0.000
NAVemptyAnchorLinks FAIL	0.24	0.00	0.995	0.991	0.499	0.000
NAVbrokenLocalLinks PASS	0.90	1.00	0.296	0.088	0.531	0.000
NAVbrokenLocalLinks FAIL	0.10	0.00	0.296	0.088	0.531	0.000
NAVmissingLogicalPath PASS	0.95	1.00	0.215	0.046	0.540	0.000
NAVmissingLogicalPath FAIL	0.05	0.00	0.215	0.046	0.540	0.000
NAVselfReferentialPage PASS	0.05	0.00	0.215	0.046	0.540	0.000
NAVselfReferentialPage FAIL	1.16	1.00	0.482	0.232	0.423	0.000
NAVstdLinksFontAndColor PASS	0.86	1.00	0.353	0.124	.514	0.000
NAVstdLinksFontAndColor FAIL	0.29	0.00	0.705	0.498	0.514	0.000

Skewness in Satisfaction Measured by LIFT Machine (N = 63)

Property	Min	Max	Mean	Median	Skewness	Kurtosis
SATimgStretching PASS	0	1	0.98	1.00	-7.937	63.000
SATimgStretching FAIL	0	1	0.02	0.00	7.937	63.000

The standard error of skewness was 0.302 and Kurtosis was 0.595.

Normality of Measurements in Satisfaction by LIFT Machine (N = 63)

Property	Mean	Median	Std. Dev.	Var.	Kolmogorov-Smirnov	
					Statistic	Sig (2-tailed)
SATimgStretching PASS	0.98	1.00	0.126	0.016	0.534	0.000
SATimgStretching FAIL	0.00	0.00	0.125	0.016	0.534	0.000

Skewness in Efficiency Measured by LIFT Machine (N = 63)

Property	Min	Max	Mean	Median	Skewness	Kurtosis
EFCimgWithSize PASS	0	1	0.16	0.00	1.914	1.716
EFCimgWithSize FAIL	0	19	2.73	1.00	2.538	7.513
EFCunvisibleImage MapsPASS	0	1	0.97	1.00	-5.473	28.867
EFCunvisibleImage MapsFAIL	0	1	0.03	0.00	5.473	28.867
EFCpageSizeSmall PASS	0	1	0.73	1.00	-1.063	-0.901
EFCpageSizeSmall FAIL	0	1	0.27	0.00	1.063	-0.901

The standard error of skewness was 0.302 and Kurtosis was 0.595.

Normality of Measurements in Efficiency by LIFT Machine (N = 63)

Property	Mean	Median	Standard Deviation	Var.	Kolmogorov-Smirnov	
					Statistic	Sig (2-tailed)
EFCimgWithSize PASS	0.16	0.00	0.368	0.1360	0.508	0.000
EFCimgWithSize FAIL	2.73	1.00	3.530	12.458	0.296	0.000
EFCunvisibleImage MapsPASS	0.97	1.00	0.177	0.031	0.540	0.000
EFCunvisibleImage MapsFAIL	0.03	0.00	0.177	0.031	0.540	0.000
EFCpageSizeSmall PASS	0.73	1.00	0.447	0.200	0.457	0.000
EFCpageSizeSmall FAIL	0.27	0.00	0.447	0.200	0.457	0.000

Coefficients of Determination (R^2) of Models With Data of Training Set (N = 32)

Mod.	Scenario	Resp. Var.	Predictor	R	R^2	Adj. R^2
1	1	OU	LNPNBW	0.368	0.135	0.106
2	1	FLX	EW	0.398	0.158	0.130
3	1	FLX	LNLNK			
4	1	NAV	LNNBW	0.635	0.403	0.384
5	1	NAV	LNLNK			
6	1	SAT	LNWPFS	0.401	0.161	0.133
7	1	NAV	lnEFCimgWith SizeFail	0.495	0.245	0.220
8	1	SAT	NAVselfReferential PageFail	0.384	0.148	0.119
9	1	SAT	EFCimgWith SizeFail	0.454	0.207	0.180
10	1	SAT	lnEFCimgWith SizeFail	0.535	0.287	0.263
11	1	OU	NAVselfReferential PageFail	0.397	0.157	0.129
12	1	OU	EFCimgWith SizeFail	0.364	0.132	0.103
13	1	LNOU	NAVselfReferential PageFail	0.395	0.156	0.128
14	1	LNOU	EFCimgWith SizeFail	0.353	0.125	0.095
15	1	LNOU	lnEFCimgWith SizeFail	0.461	0.213	0.186
16	2	NAV	EFCimgWith SizePass	0.801	0.641	0.630

Correlations of All Models Found, Training Set (N = 32)

Model	Scen.	Resp. Var.	Predictor	Corr.	Sig ($p < 0.05$)
1	1	OU	LNPNBW	0.368	0.038
2	1	FLX	EW	-0.398	0.024
3	1	FLX	LNLNK	-0.372	0.036
4	1	NAV	LNNBW	-0.635	0.000
5	1	NAV	LNLNK	-0.482	0.005
6	1	SAT	LNWPFS	0.401	0.023
7	1	NAV	lnEFCimgWith SizeFail	-0.495	0.004
8	1	SAT	NAVselfReferential PageFail	-0.384	0.030
9	1	SAT	EFCimgWith SizeFail	-0.454	0.009
10	1	SAT	lnEFCimgWith SizeFail	-0.535	0.002
11	1	OU	NAVselfReferential PageFail	0.397	0.025
12	1	OU	EFCimgWith SizeFail	0.364	0.041
13	1	LNOU	NAVselfReferential PageFail	0.395	0.025
14	1	LNOU	EFCimgWith SizeFail	0.353	0.047
15	1	LNOU	lnEFCimgWith SizeFail	0.461	0.008
16	2	NAV	EFCimgWithSizePass	0.650	0.000

Independent Samples t-test of Assessed Ratings Between Scenarios (N = 63)

	Levene's Test for Equality of Variances		**t**	**sig.**	**df**	**Mean Diff., std err.**
	F	**sig.**				
EFF	4.551	0.035	2.864	0.005	122.995	0.3450, 0.1205
FLX	0.021	0.884	-0.852	0.396	124.000	-0.1129, 0.1325
SAT	4.861	0.029	0.897	0.371	124.000	0.1505, 0.1678
EFC	8.140	0.005	1.430	0.155	124.000	0.2174, 0.1520
NAV	16.209	0.000	0.855	0.394	124.000	0.1503, 0.1757
OU	2.001	0.160	-2.428	0.017	120.906	-0.5174, 0.2131

Levene's test for equality of variances is applied in F test. If significance level of t is less than 0.05, equal variances are not assumed.

Paired t-test of Assessed Versus Predicted Ratings Within Category Within Scenario 1, Prediction Set (N = 31)

Resp. Var.	Paired With Predicted Resp. Var.	Mean	Std. Dev.	Std. Error of Mean	t	sig. (2-tailed)
FLX	FLX_EW	0.07298	0.74577	0.13394	0.545	0.590
	FLX_LNLNK	0.09359	0.68373	0.12280	0.762	0.452
NAV	NAV_LNNBW	0.19332	1.00247	0.18005	1.074	0.292
	NAV_LNLNK	0.38529	1.09182	0.19610	1.965	0.059
	lnEFCimgWithSizeFail	0.23565	1.05425	0.18935	1.245	0.223
SAT	SAT_LNWPFS	0.07111	1.12551	0.20215	0.352	0.727
	SAT_NAVselfReferential PageFail	0.25091	1.13337	0.20356	1.233	0.227
	SAT_EFCimgWithSize Fail	0.17066	1.03599	0.18607	0.917	0.366
	SAT_lnEFCimgWithSize Fail	0.11102	0.94078	0.16897	0.657	0.516
OU	NAVselfReferentialPage Fail	-0.38181	1.56180	0.28051	-1.361	0.184
	EFCimgWithSize Fail	-0.18187	1.44989	0.26041	-0.698	0.490
LNOU	NAVselfReferentialPage Fail	-0.05938	0.22544	0.04049	-1.467	0.153
	EFCimgWithSizeFail	-0.03277	0.21066	0.03784	-0.866	0.393
	lnEFCimgWithSizeFail	-0.04014	0.18935	0.03401	-1.180	0.247

Paired t-test of Assessed Versus Predicted Ratings Within Category Within Scenario 2, Prediction Set (N = 31)

Resp. Var.	Paired With Predicted Resp. Var.	Mean	Std. Dev.	Std. Error of Mean	t	sig. (2-tailed)
NAV	NAV_EFCimgWithSizePass	-0.00002	0.62220	0.11175	0.000	1.000

Paired t-test Between Predicted Ratings Within Category of Usability Within Scenario 1 (N = 63)

Paired Models	Response Variables	Mean	Std. Dev.	Std. Error of Mean	t	sig. (2-tailed)
2, 3	FLX_EW, FLX_LNLNK	0.01278	0.22755	0.02867	0.446	0.657
4, 5	NAV_LNNBW, NAV_LNLNK	0.09457	0.65679	0.08275	1.143	0.258
4, 7	NAV_LNNBW, NAV_ lnEFCimg WithSizeFail	0.02024	0.44003	0.05544	0.365	0.716
5, 7	NAV_LNLNK, NAV_lnEFCimg WithSizeFail	-0.07432	0.66761	0.08411	-0.884	0.380
6, 8	SAT_LNWPFS, SAT_NAVselfReferential PageFail	0.08935	0.55011	0.06931	1.289	0.202
6, 9	SAT_LNWPFS, SAT_EFCimg WithSizeFail	0.04987	0.61788	0.07785	0.641	0.524
6, 10	SAT_LNWPFS, SAT_lnEFCimg WithSizeFail	0.02052	0.71086	0.08956	0.229	0.820
8, 9	SAT_NAVselfReferential PageFail, SAT_EFCimg WithSizeFail	-0.03948	0.30813	0.03882	-1.017	0.313
8, 10	SAT_NAVselfReferential PageFail, SAT_lnEFCimg WithSizeFail	-0.06883	0.41687	0.05252	-1.311	0.195
9, 10	SAT_EFCimgWithSizeFail, SAT_lnEFCimgWithSizeFail	-0.02936	0.28376	0.03575	-0.821	0.415
1, 11	OU_LNPNBW, OU_NAVselfReferential PageFail	-0.10101	0.52991	0.06676	-1.513	0.135
1, 12	OU_LNPNBW, OU_EFCimgWithSizeFail	-0.06172	0.47279	0.05957	-1.036	0.304
1, 13	OU_LNPNBW, LNOU_NAVselfReferential PageFail	-0.02662	0.54221	0.06831	-0.390	0.698
1, 14	OU_LNPNBW, LNOU_EFCimgWithSizeFail	0.02236	0.48222	0.06075	0.368	0.714
1, 15	OU_LNPNBW, LNOU_lnEFCimgWithSizeFail	0.03438	0.52169	0.06573	0.523	0.603

Paired t-test Between Predicted Ratings Within Category of Usability Within Scenario 1 (N = 63)

Paired Models	Response Variables	Mean	Std. Dev.	Std. Error of Mean	t	sig. (2-tailed)
11, 12	OU_NAVselfReferentialPage Fail, OU_EFCimgWithSizeFail	0.03929	0.34230	0.04313	0.911	0.366
11, 13	OU_NAVselfReferentialPage Fail, LNOU_NAVselfReferential PageFail	0.07439	0.02831	0.00357	20.856	0.000
11, 14	OU_NAVselfReferentialPage, LNOU_EFCimgWifthSizeFail	0.12337	0.35049	0.04416	2.794	0.007
11, 15	OU_NAVselfReferentialPage Fail, LNOU_lnEFCimgWithSize Fail	0.13539	0.45649	0.05751	2.354	0.022
12, 13	OU_EFCimgWithSizeFail, LNOU_NAVselfReferential PageFail	0.03510	0.35156	0.04429	0.792	0.431
12, 14	OU_EFCimgWithSizeFail, LNOU_EFCimgWithSizeFail	0.08408	0.02015	0.00254	33.119	0.000
12, 15	OU_EFCimgWithSizeFail, LNOU_lnEFCimg WithSizeFail	0.09610	0.30848	0.03886	2.473	0.016
13, 14	LNOU_NAVselfReferential PageFail, LNOU_EFCimgWithSizeFail	0.04898	0.35898	0.04523	1.083	0.283
13, 15	LNOU_NAVselfReferential PageFail, LNOU_lnEFCimg WithSizeFail	0.06100	0.45984	0.05793	1.053	0.296
14, 15	LNOU_EFCimgWithSizeFail, LNOU_lnEFCimg WithSizeFail	0.01203	0.31341	0.03949	0.305	0.762

Models 13, 14, and 15 predict ratings of OU as their natural logs. To compare these models to models yielding OU, the inverse natural log of LNOU is used.

Independent Samples t-test Using Predicted Ratings for Navigability Between Scenarios (N = 63)

Paired Models	Response Variables	F, sig.	t	sig.	Mean Diff, Std. Err.
4, 16	NAV_LNNBW, NAV_EFCimgWith SizePass	12.785, 0.000	-0.043	0.966	-0.0043, 0.1013
5, 16	NAV_LNLNK, NAV_EFCimgWith SizePass	12.723, 0.001	-1.215	0.227	-0.0989, 0.0814
7, 16	NAV_lnEFCimgWith SizeFail, NAV_EFCimgWith SizePass	13.420, 0.000	-0.301	0.764	-0.0246, 0.0816

Correlations of Paired Models By Category Within Scenario 1 (N = 63)

Paired Models	Response Variables Tested	N	Corr.	Sig.
2, 3	FLX_EW, FLX_LNLNK	63	0.547	0.000
4, 5	NAV_LNLNK, NAV_LNNBW	63	0.499	0.000
4, 7	NAV_LNNBW, NAV_ lnEFCimg WithSizeFail	63	0.797	0.000
5, 7	NAV_LNLNK, NAV_lnEFCimg WithSizeFail	63	0.259	0.041
6, 8	SAT_LNWPFS, SAT_NAVselfReferential PageFail	63	-0.087	0.499
6, 9	SAT_LNWPFS, SAT_EFCimgWithSizeFail	63	-0.252	0.046
6, 10	SAT_LNWPFS, SAT_lnEFCimgWithSizeFail	63	-0.175	0.170
8, 9	SAT_NAVselfReferential PageFail, SAT_EFCimgWithSizeFail	63	0.613	0.000
8, 10	SAT_NAVselfReferential PageFail, SAT_lnEFCimgWithSizeFail	63	0.581	0.000

Correlations of Paired Models By Category Within Scenario 1 (N = 63)

Paired Models	Response Variables Tested	N	Corr.	Sig.
9, 10	SAT_EFCimgWithSizeFail, SAT_lnEFCimgWithSizeFail	63	0.840	0.000
1, 11	OU_LNPNBW, OU_NAVselfReferentialPageFail	63	0.202	0.113
1, 12	OU_LNPNBW, OU_EFCimgWithSizeFail	63	0.279	0.027
1, 13	OU_LNPNBW, LNOU_NAVselfReferential PageFail	63	0.206	0.105
1, 14	OU_LNPNBW, LNOU_EFCimgWithSizeFail	63	0.273	0.030
1, 15	OU_LNPNBW, LNOU_lnEFCimgWithSizeFail	63	0.463	0.000
11, 12	OU_NAVselfReferentialPageFail, OU_EFCimgWithSizeFail	63	0.613	0.000
11, 13	OU_NAVselfReferentialPageFail, LNOU_NAVselfReferentialPageFail	63	0.999	0.000
11, 14	OU_NAVselfReferentialPage, LNOU_EFCimgWifthSizeFail	63	0.605	0.000
11, 15	OU_NAVselfReferentialPageFail, LNOU_lnEFCimgWithSizeFail	63	0.593	0.000
12, 13	OU_EFCimgWithSizeFail, LNOU_ NAVselfReferentialPageFail	63	0.622	0.000
12, 14	OU_EFCimgWithSizeFail, LNOU_EFCimgWithSizeFail	63	0.999	0.000
12, 15	OU_EFCimgWithSizeFail, LNOU_lnEFCimgWithSizeFail	63	0.863	0.000
13, 14	LNOU_NAVselfReferentialPageFail, LNOU_EFCimgWithSizeFail	63	0.614	0.000
13, 15	LNOU_NAVselfReferentialPageFail, LNOU_lnEFCimgWithSizeFail	63	0.595	0.000
14, 15	LNOU_EFCimgWithSizeFail, LNOU_lnEFCimgWithSizeFail	63	0.847	0.000

Models 13, 14, and 15 predict ratings of OU as their natural logs. To compare these models to models yielding OU, the inverse natural log of LNOU is used.

Correlations of Paired Models for Navigability Between Scenarios (N = 63)

Paired Models	Response Variables	N	Corr.	Sig.
4, 16	NAV_LNNBW, NAV_EFCimgWith SizePass	63	0.797	0.000
5, 16	NAV_LNLNK, NAV_EFCimgWith SizePass	63	0.562	0.000
7, 16	NAV_lnEFCimgWith SizeFail, NAV_EFCimgWith SizePass	63	0.719	0.000

Unstandardized and Standardized Residuals, Scenario 1, WPMA (N = 63)

Mod.	Resp. Var.	Resid.	Min	Max	Mean	Std Dev	Var.	Skew.	Kurt.
1	OU	Unstd	-3.23	2.43	-0.0867	1.17835	1.389	-0.377, 0.302	0.467, 0.595
		Std.	-2.67044	2.13405	0.00000	1.00000	1.00000		
2	FLX	Unstd	-0.86	2.21	0.033	0.65035	0.423	0.887, 0.302	0.907, 0.595
		Std.	-1.37353	3.33928	0.00000	1.00000	1.00000		
3	FLX	Unstd	-0.88	1.98	0.0461	0.61997	0.384	0.800, 0.302	0.610, 0.595
		Std.	-1.48707	3.11393	0.00000	1.00000	1.00000		
4	NAV	Unstd	-1.31	2.49	0.0960	0.91328	0.834	0.649, 0.302	-0.053, 0.595
		Std.	-1.54477	2.62039	0.00000	1.00000	1.00000		
5	NAV	Unstd	-1.35	2.65	0.1905	1.02448	1.050	0.770, 0.302	-0.266, 0.595
		Std.	-1.50467	2.39975	0.00000	1.00000	1.00000		
6	SAT	Unstd	-2.10	2.87	0.0340	0.98274	0.966	0.640, 0.302	0.649, 0.595
		Std.	-2.16974	2.88345	0.00000	1.00000	1.00000		

A comma separates values of Skewness and Kurtosis from their significance.
Statistics in this table are not based on absolute values of residuals.

Unstandardized and Standardized Residuals, Scenario 1, LIFT (N = 63)

Mod.	Resp. Var	Resid.	Min	Max	Mean	Std. Dev.	Var.	Skew.	Kurt.
7	NAV	Obs.	-1.12	2.19	0.1162	0.99012	0.980	0.522, 0.302	-0.664, 0.595
		Std.	-1.24753	2.09653	0.0000	1.0000	1.000		
8	SAT	Obs.	-1.37	3.31	0.1234	0.99732	0.995	0.897, 0.302	0.691, 0.595
		Std.	-1.49738	3.19818	0.0000	1.0000	1.000		
9	SAT	Obs.	-1.48	2.42	0.0839	0.92690	0.859	0.600, 0.302	-0.306, 0.595
		Std.	-1.68830	2.51494	0.0000	1.0000	1.000		
10	SAT	Obs.	-1.53	1.83	0.0545	0.85457	0.730	0.442, 0.302	-0.744, 0.595
		Std.	-1.85302	2.07823	0.0000	1.0000	1.000		
11	SAT	Obs.	-3.98	2.38	-0.1877	1.31870	1.739	-0.404, 0.302	0.619, 0.595
		Std.	-2.87429	1.94638	0.0000	1.0000	1.000		
12	OU	Obs.	-3.44	2.46	-0.1484	1.24622	1.553	-0.221, 0.302	-0.048, 0.595
		Std.	-2.64288	2.09065	0.0000	1.0000	1.000		
13	OU	Obs.	-3.96	2.47	-0.1133	1.32426	1.754	-0.409, 0.302	0.643, 0.595
		Std.	-2.90256	1.95015	0.0000	1.0000	1.000		
14	LNOU	Obs.	-3.36	2.54	-0.0643	1.24855	1.559	-0.215, 0.302	-0.056, 0.595
		Std.	-2.63925	2.08918	0.0000	1.0000	1.000		
15	LNOU	Obs.	-2.74	2.61	-0.0523	1.16726	1.363	-0.114, 0.302	-0.261, 0.595
		Std.	-2.30514	2.28160	0.0000	1.0000	1.000		

For models 14 and 15, residuals are observed OU minus inverse log of predicted LNOU.

Statistics in this table are not based on absolute values of residuals.

Unstandardized and Standardized Residuals, Scenario 2, LIFT (N = 63)

Mod.	Resp. Var.	Resid.	Min	Max	Mean	Std. Dev.	Var.	Skew.	Kurt.
16	NAV	Obs.	-0.78	1.22	-0.0585	0.54601	0.298	0.692, 0.302	0.140, 0.595
		Std.	-1.3250	2.3379	0.0000	1.00000	1.0000		

Statistics in this table are not based on absolute values of residuals.

Accuracy of Models (N = 63)

Model	Response Var.	Mean	Std. Dev.
M1	OU	0.79143	0.57619
M2	OU	0.85713	0.55200
1	OU	0.91238	0.74225
12	OU	1.02255	0.71501
15	LNOU	*1.13558	*1.11030
14	LNOU	*1.14881	*1.11815
13	LNOU	*1.15229	*1.13433
3	FLX	0.48190	0.38866
2	FLX	0.52714	0.37390
4	NAV	0.73889	0.53761
7	NAV	0.79661	0.59075
5	NAV	0.79857	0.66158
10	SAT	0.71485	0.46273
6	SAT	0.75111	0.62850
9	SAT	0.75866	0.52968
8	SAT	0.77298	0.63445
11	SAT	1.04670	0.81374
16	NAV	0.42762	0.34002
M4	OU	0.68513	0.58191
M3	OU	0.73196	0.68328

*Inverse natural log of mean rating. Accuracy is the mean of absolute value of residuals.

Paired t test Comparison of Models Within Category Using Absolute Values of Residuals (N = 63)

Pairs	Pairs	t	Sig. (2-tailed)
OU	M1-M2	-1.043	0.301
	M1-1	1.458	0.150
	M1-12	3.197	0.002
	M1-15	-5.604	0.000
	M1-14	-5.735	0.000
	M1-13	-5.875	0.000
	M2-1	0.818	0.417
	M2-12	2.271	0.027
	M2-15	-4.687	0.000
	M2-14	-4.847	0.000
	M2-13	-5.028	0.000
	1-12	-1.959	0.055
	1-15	-2.828	0.006
	1-14	-3.070	0.003
	1-13	-3.208	0.002
	12-15	-1.593	0.116
	12-14	-1.838	0.071
	12-13	-1.948	0.056
	15-14	2.332	0.023
	15-13	2.084	0.041
	14-13	0.848	0.400
FLX	2-3	1.688	0.096
NAV	4-7	-1.199	0.235
	5-7	0.030	0.976
SAT	10-6	0.503	0.617
	10-9	1.373	0.175
	10-8	1.229	0.224
	10-11	-3.374	0.001
	6-9	-0.129	0.898
	6-8	-0.353	0.725
	6-11	-3.354	0.001
	9-8	0.431	0.668
	9-11	-3.217	0.002
	8-11	-3.153	0.002
NAV	16	NA	NA
OU	M3-M4	0.886	0.379

Inverse natural log of absolute values used for residuals by models 13, 14, and 15.

Independent Samples t test Comparisons Between Absolute Values of Residuals for NAV Models Between Scenarios 1 and 2 (N = 63)

Models	t, sig	F, sig
4-16	3.884, 0.000	5.505, 0.021
5-16	3.958, 0.000	17.559, 0.000
7-16	4.297, 0.000	21.320, 0.000

Models Found for Overall Usability by Multiple Regressions (N = 63)

Model	Scenario	Equation
M1	1	OU = 1.249 * LNTAMFS – 0.636*LNIMG – 4.557
M2	1	OU = -1.949*EFCimgWithSizePass – 1.011*NAVbrokenLocalLinksFail + 7.975
M3	2	OU = 0.233 * LNPNBW + 7.235
M4	2	OU = -1.682*EFCimgWithSizePass – 0.618*NAVselfReferentialPageFail + 9.069

Correlation Coefficients of Multivariate Regression (N = 63)

Model	Desc.	Predictors	R	R^2	Adj. R^2	Std. Err. of Est.	df	F	Sig.
M1	Scen.1, WPMA	LNTAMFS, LNIMG	0.646	0.417	0.397	1.00	62	21.435	0.000
M2	Scen.1, LIFT	EFCimgWith SizePass, NAVbrokenLocal LinksFail	0.605	0.366	0.345	1.04	62	17.348	0.000
M3	Scen.2, WPMA	LNPNBW	0.399	0.159	0.145	1.013	62	11.541	0.001
M4	Scen.2, LIFT	EFCimgWith SizePass, NAVselfReferential PageFail	0.567	0.321	0.299	0.92	62	14.192	0.000

Models of Overall Usability Found by Multiple Regressions (N = 63)

Mod.	Desc.	Pred.	Coeff., sig.	Const., sig.	t test, sig	Multicoll.	
						Tol.	VIF
M1	Scen. 1, WPMA	LNTAMFS	1.249, 0.262	-4.557, 2.812	4.760, 0.000	0.884	1.131
		LNIMG	-0.636, 0.109		-5.849, 0.000	0.884	1.131
M2	Scen. 1, LIFT	EFCimgWith SizePass	-1.949	6.964, 0.430	-5.423, 0.000	1.000	1.000
		NAVbroken LocalLinksFail	-1.011		-2.261, 0.027	1.000	1.000
M3	Scen. 2, WPMA	LNPNBW	0.233	7.235, 0.281	3.397	1.000	1.000
M4	Scen. 2, LIFT	EFCimgWith SizePass	-1.682, 0.326	9.069, 0.326	-5.165, 0.000	0.945	1.058
		NAVself Referential PageFail	-0.618, 0.249		-2.483, 0.016	0.945	1.058

Precisions of Bivariate Models with Prediction Set (N = 31)

Mod.	Threshold	Overall Precision		Precision At Threshold	
		Pairs At or Above	Pairs Both Below	Pairs At or Above	Pairs Below
OU					
1	7.554	*0.323	0.290	*0.526	*0.750
11	7.759	0.032	0.452	0.167	0.560
12	7.680	0.129	*0.484	0.500	0.652
LNOU					
13	2.038	0.032	0.452	0.167	0.560
14	2.027	0.161	*0.484	*0.625	0.652
15	2.019	*0.258	0.355	0.533	*0.688
FLX					
2	0.670	0.258	0.258	0.500	0.533
3	0.649	*0.355	*0.355	*0.688	*0.733
NAV					
4	1.022	0.258	*0.484	*0.667	*0.789
5	0.830	0.290	0.355	0.563	0.733
7	0.980	*0.355	0.355	0.563	0.733
SAT					
6	1.440	0.161	*0.419	0.500	0.619
8	1.260	*0.516	0.129	*0.640	0.667
9	1.340	0.419	0.226	0.565	*0.875
10	1.400	0.290	0.355	0.563	0.733
NAV (Scenario 2)					
16	*0.993	*0.226	*0.516	*1.000	*0.667

* Asterisks denote highest values for category.

Precisions of Models by Multiple Regressions (N = 63)

Mod.	**Threshold**	**Overall Precision**		**Precision At Threshold**	
		Pairs At or Above	**Pairs Both Below**	**Pairs At or Above**	**Pairs Below**
Scenario 1					
M1	7.560	0.365	0.381	0.742	0.750
M2	6.558	0.683	0.143	0.896	0.600
Scenario 2					
*M3	0.809	0.429	0.270	0.750	0.603
M4	0.809	0.460	0.222	0.707	0.636

* Bivariate model

Correlations and Independent Samples t test Between Scenarios 1 and 2 Using Multivariate Models for Overall Usability (N = 63, p < 0.05, df = 62)

Comb.	Means, std. dev.	Correlation, 2-tailed sig.	t test, 2-tailed sig.	F, sig.	df	Mean Diff, Std. Err.
M1	7.5700, 0.8311	0.510, 0.000	-4.361, 0.000	28.960, 0.000	93.802	-0.5159, 0.1183
M3	8.0859, 0.4367					
M1	7.5700, 0.8311	0.604, 0.000	-3.946, 0.000	6.711, 0.011	114.795	-0.5159, 0.1307
M4	8.0859, 0.6212					
M2	7.5693, 0.7797	0.668, 0.000	-4.588, 0.000	18.622, 0.000	97.407	-0.5166, 0.1126
M3	8.0859, 0.4367					
M2	7.5693, 0.7797	0.797, 0.000	-4.113, 0.000	2.686, 0.104	118.105	-0.5166, 0.1256
M4	8.0859, 0.6212					

Levene’s test for equality of variances is applied in F test. If significance level of t is < 0.05, equal variances are not assumed.

Comparisons Between Assessed Overall Usability and Multivariate Models for Overall Usability by Paired Samples t Test (N = 63, p < 0.05, df = 62)

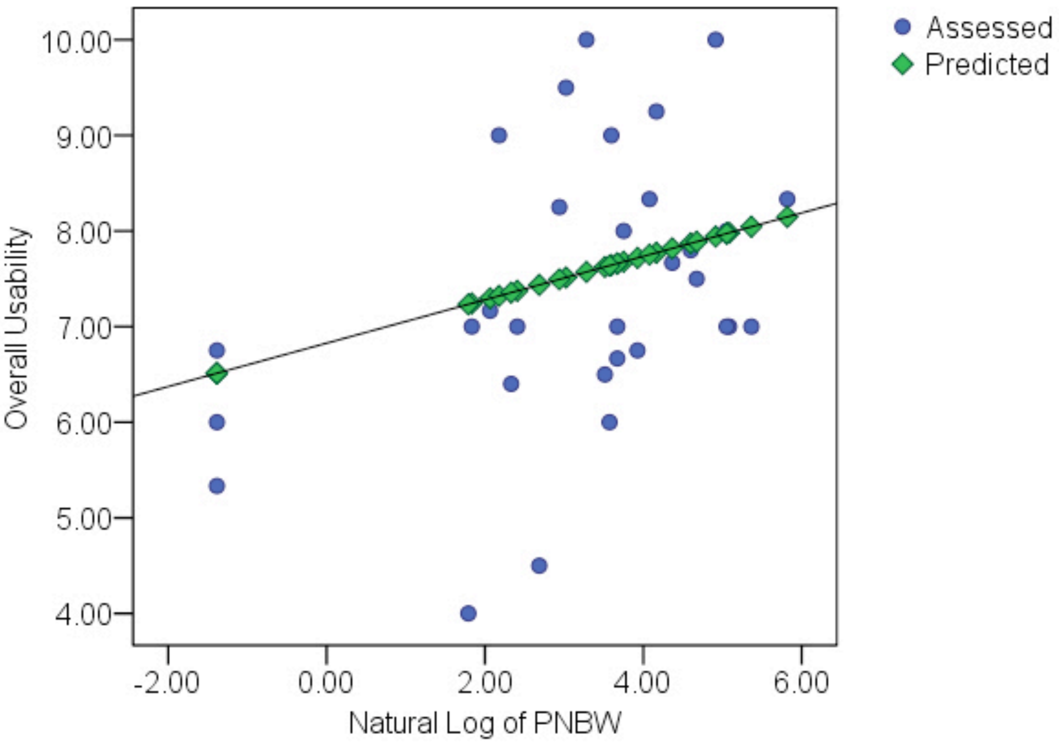
Combinations	**Mean**	**Correlation, 2-tailed sig.**	**t test, 2-tailed sig.**	**Different?**
Assessed Scenario 1	7.5693500	0.646, 0.000	-0.006, 0.996	No
M1	7.5700389			
Assessed Scenario 1	7.5693500	0.605, 0.000	0.000, 1.000	No
M2	7.5693492			
Assessed Scenario 2	8.0867725	0.399, 0.001	0.007, 0.995	No
M3	8.0859377			
Assessed Scenario 2	8.0867725	0.567, 0.000	0.007, 0.994	No
M4	8.0859206			

Comparisons Between Multivariate Models for Overall Usability Within Scenario by Paired Samples t Test (N = 63, p < 0.05, df = 62)

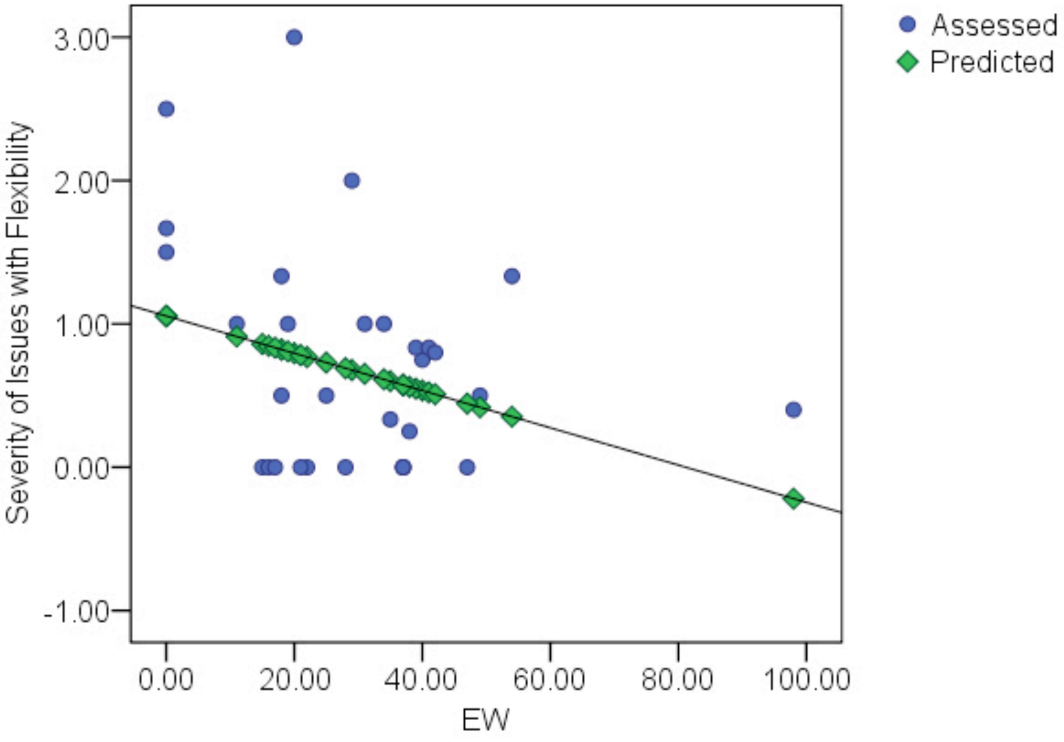
Combinations	Mean	Correlation, 2-tailed sig.	t test, 2-tailed sig.	Different?
M1	7.5700389	0.776, 0.000	0.010, 0.992	No
M2	7.5693492			
M3	8.0859377	0.608, 0.000	0.000, 1.000	No
M4	8.0859206			

APPENDIX C

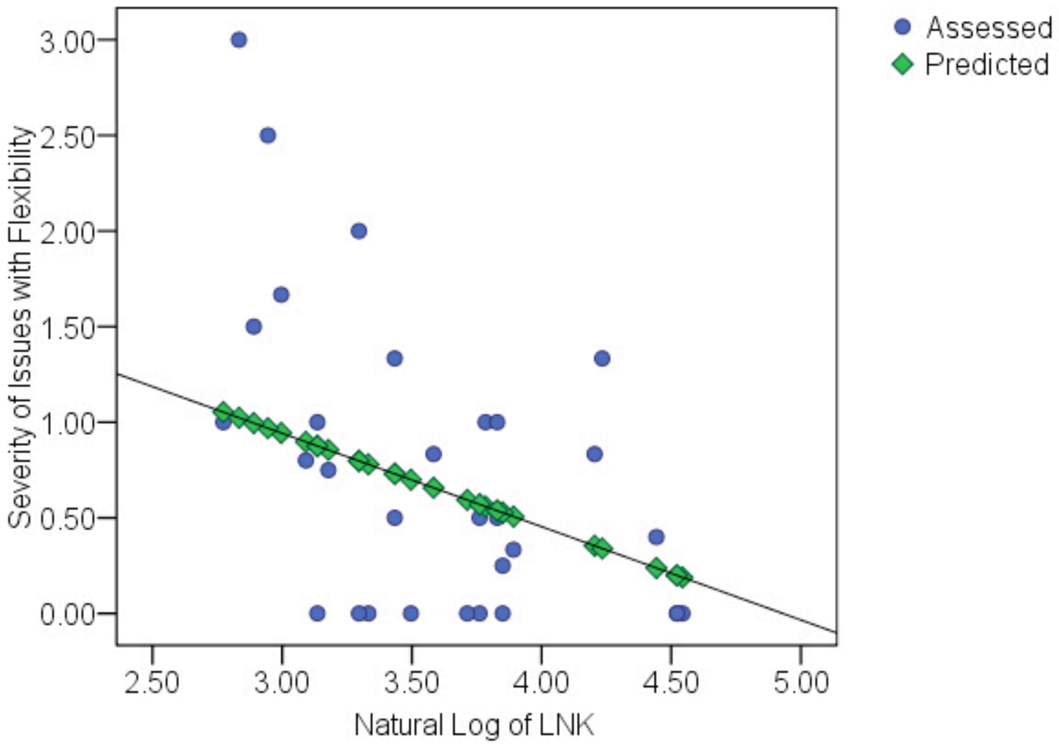
Model 1, LNPNBW Versus OU, Scenario 1, WPMA, Prediction Set (N = 31)



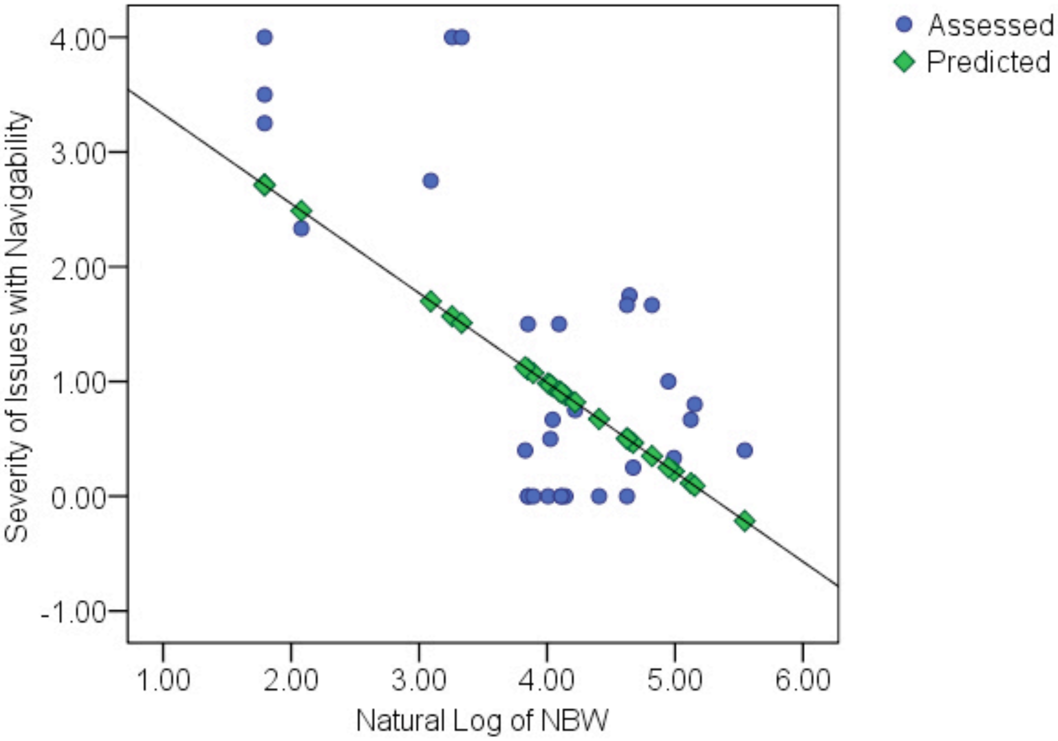
Model 2, FLX Versus EW, Scenario 1, WPMA, Prediction Set (N = 31)



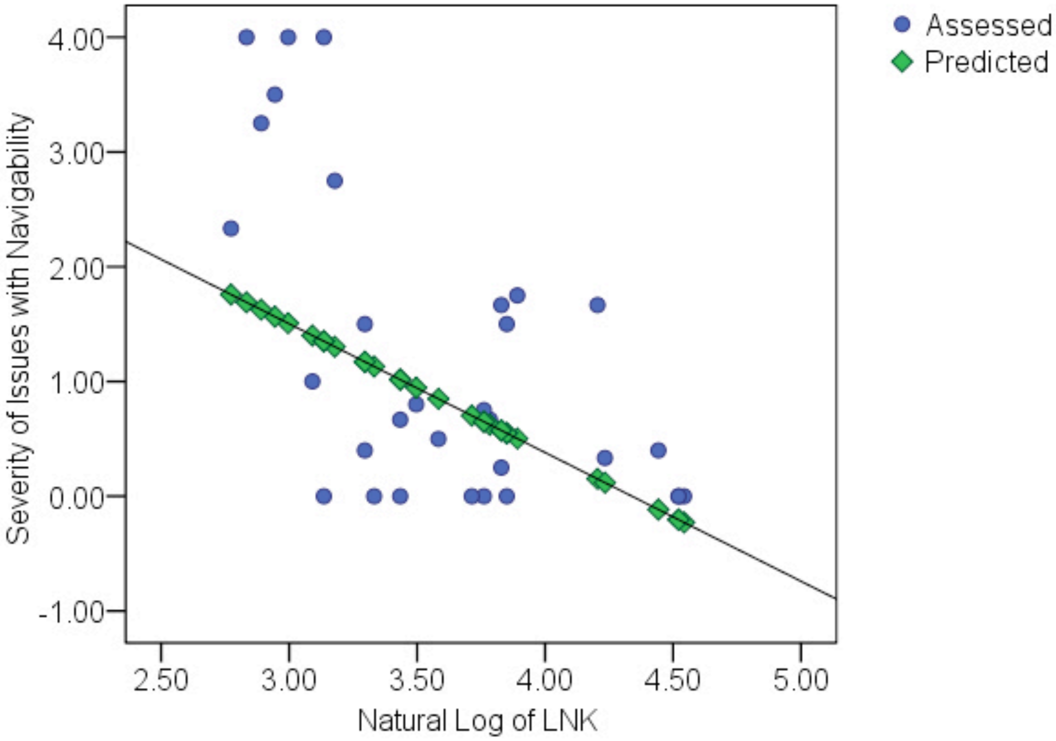
Model 3, LNLNK Versus FLX, Scenario 1, WPMA, Prediction Set (N = 31)



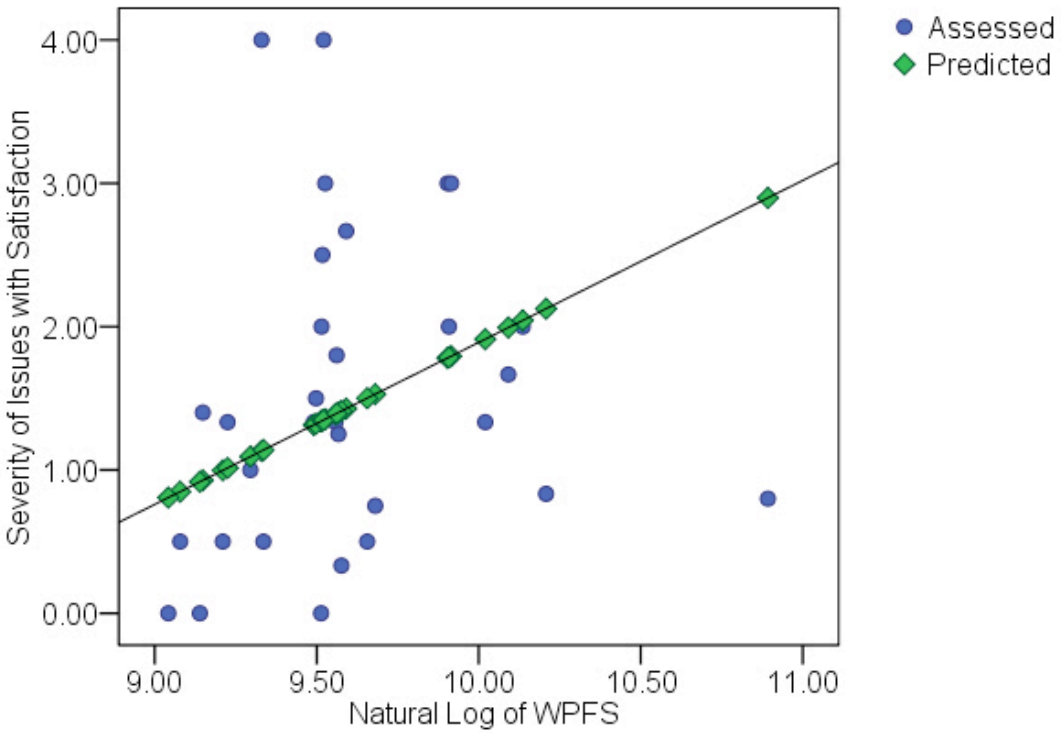
Model 4, LNNBW Versus NAV, Scenario 1, WPMA, Prediction Set (N = 31)

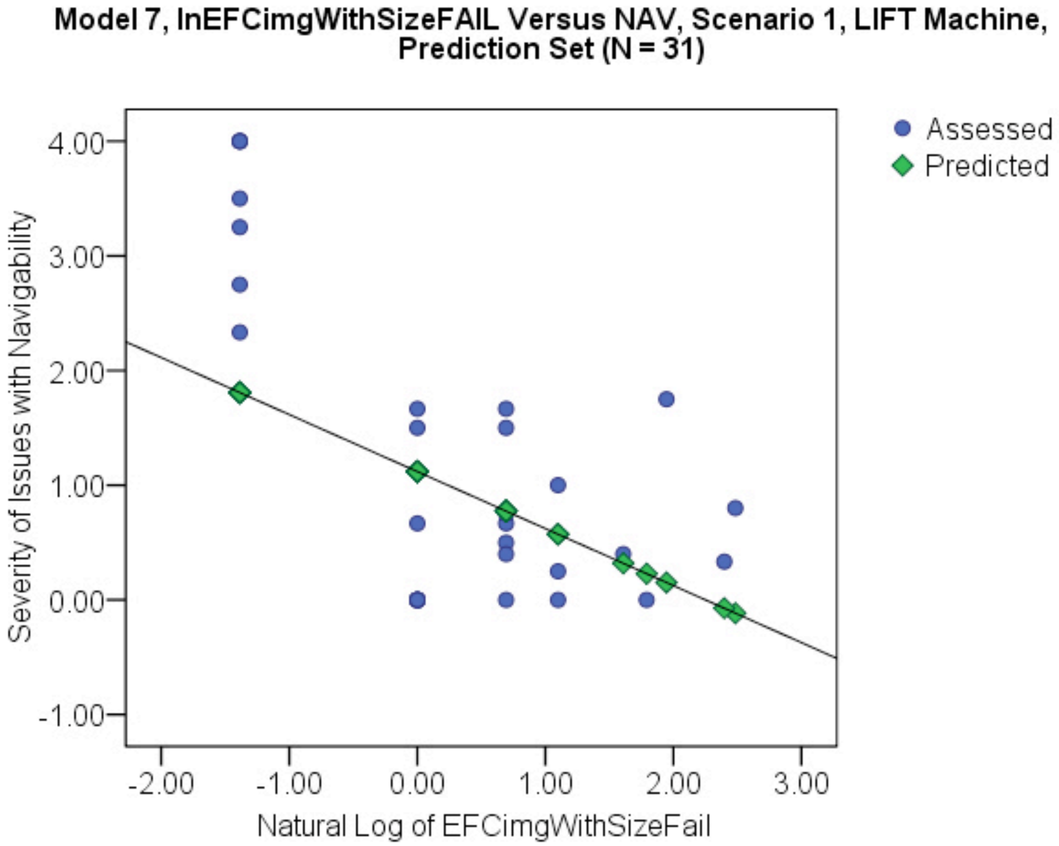


Model 5, LNLNK Versus NAV, Scenario 1, WPMA, Prediction Set (N = 31)

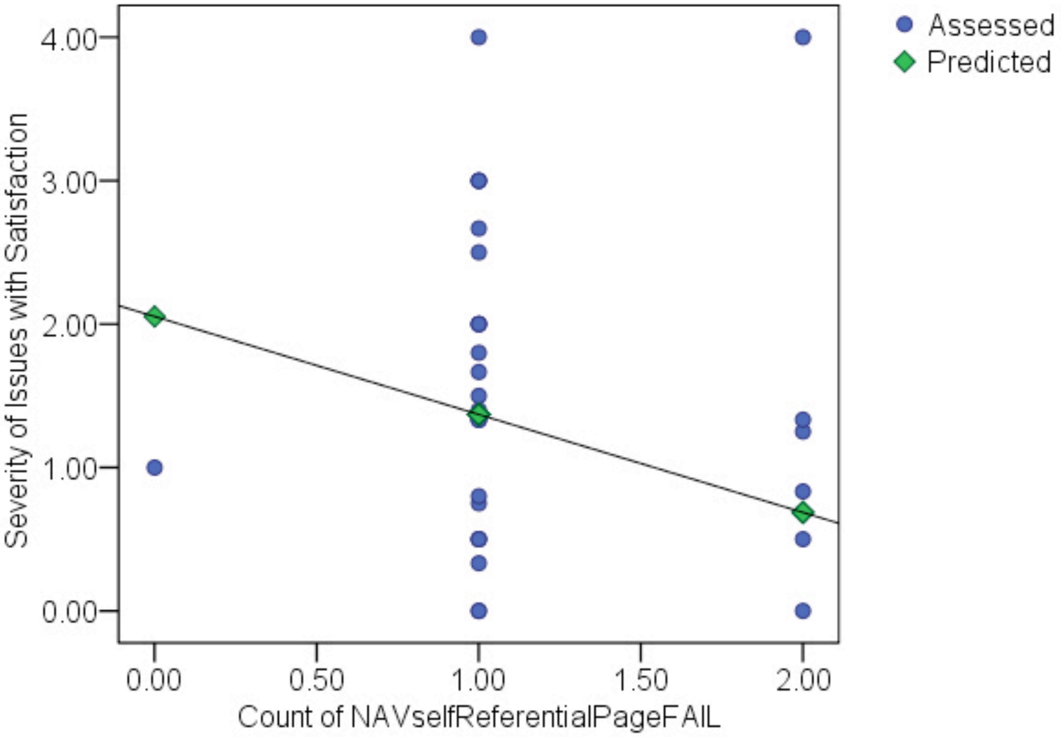


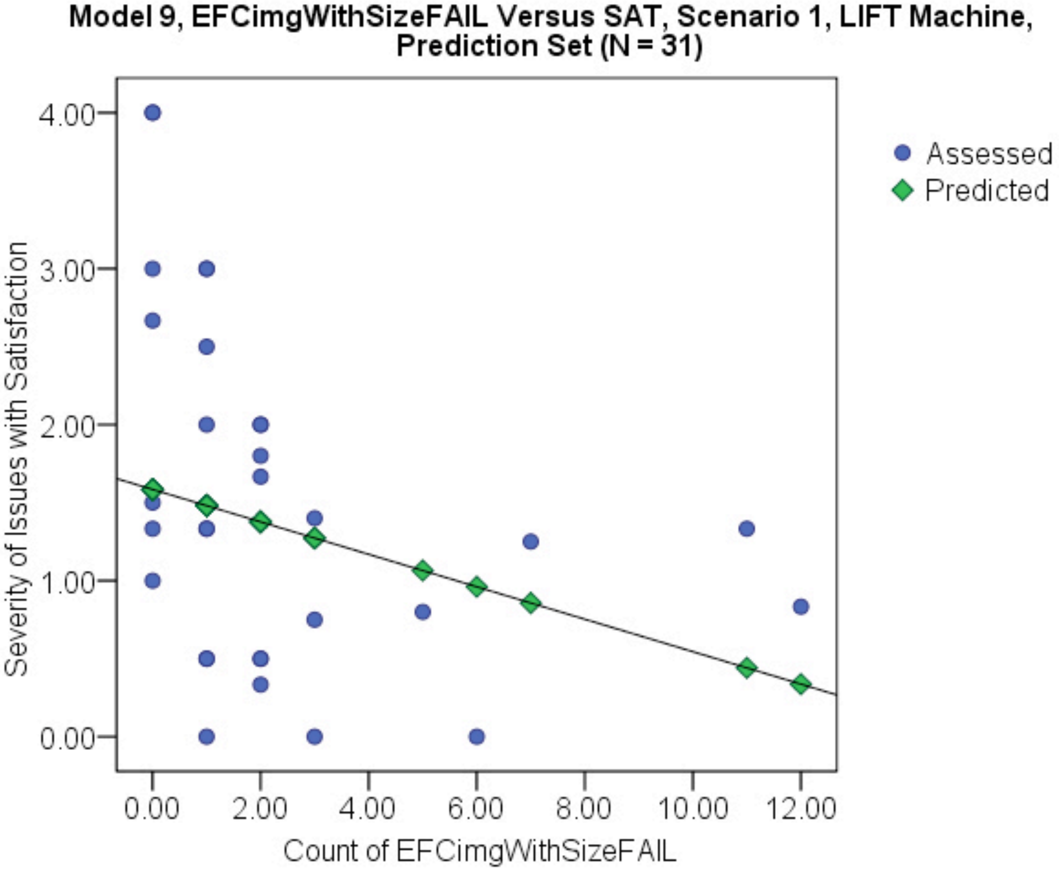
Model 6, LNWPFS Versus SAT, Scenario 1, WPMA, Prediction Set (N = 31)



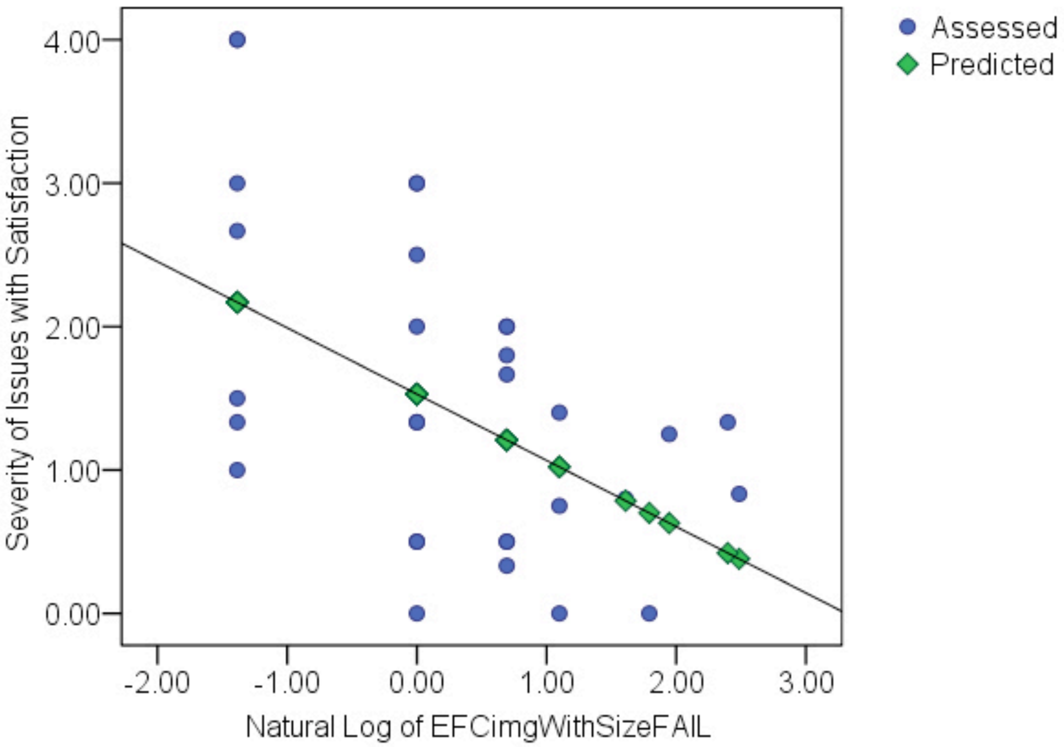

Model 7, lnEFCimgWithSizeFAIL Versus NAV, Scenario 1, LIFT Machine,
Prediction Set (N = 31)
● Assessed
◆ Predicted
Severity of Issues with Navigability
4.00
3.00
2.00
1.00
0.00
-1.00
-2.00
-1.00
0.00
1.00
2.00
3.00
Natural Log of EFCimgWithSizeFail

Model 8, NAVselfReferentialPageFAIL Versus SAT, Scenario 1, LIFT Machine, Prediction Set (N = 31)

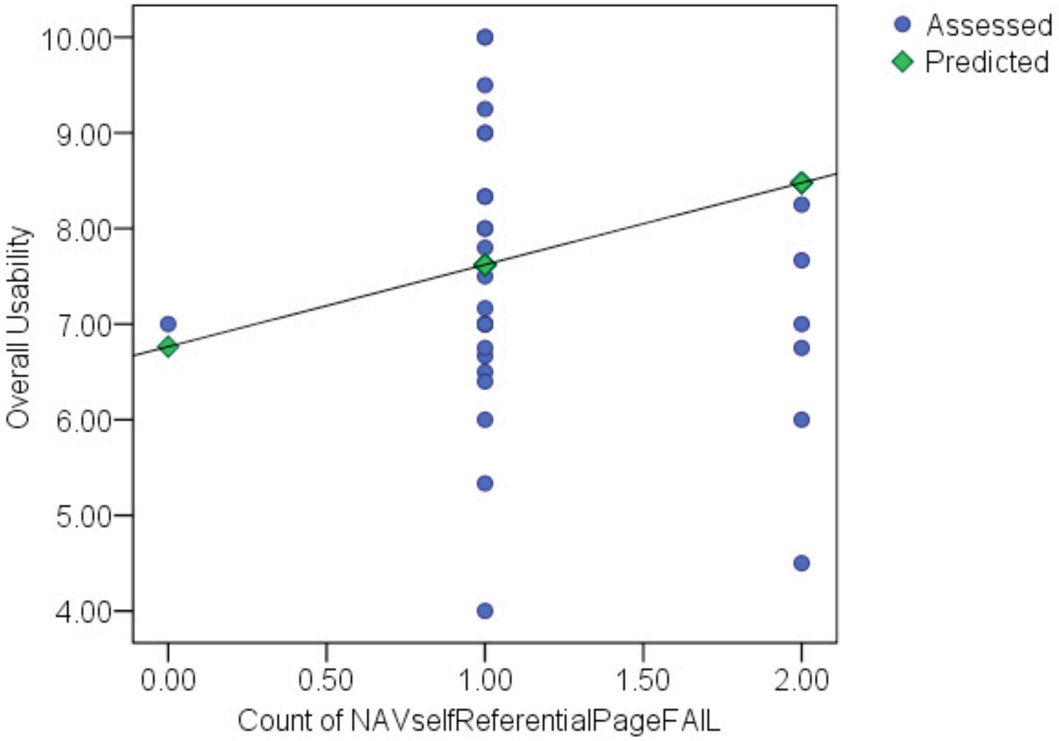



Model 9, EFCimgWithSizeFAIL Versus SAT, Scenario 1, LIFT Machine, Prediction Set (N = 31)
Severity of Issues with Satisfaction
4.00
3.00
2.00
1.00
0.00
0.00
2.00
4.00
6.00
8.00
10.00
12.00
Count of EFCimgWithSizeFAIL
● Assessed
◆ Predicted

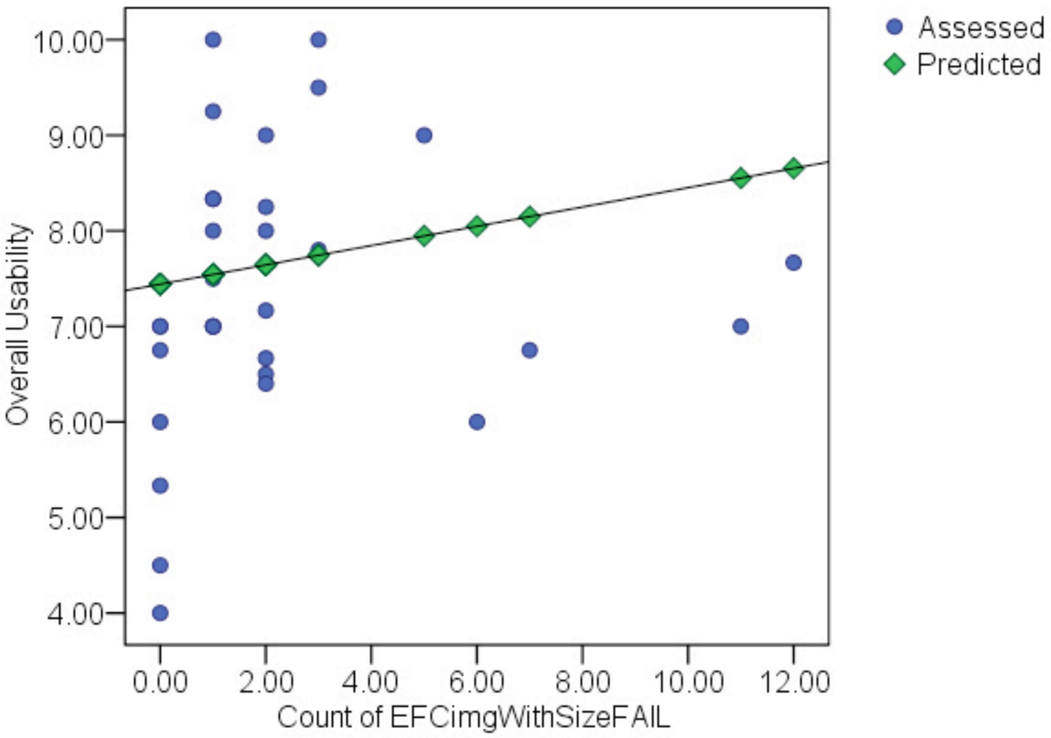
Model 10, lnEFCimgWithSizeFAIL Versus SAT, Scenario 1, LIFT Machine, Prediction Set (N = 31)



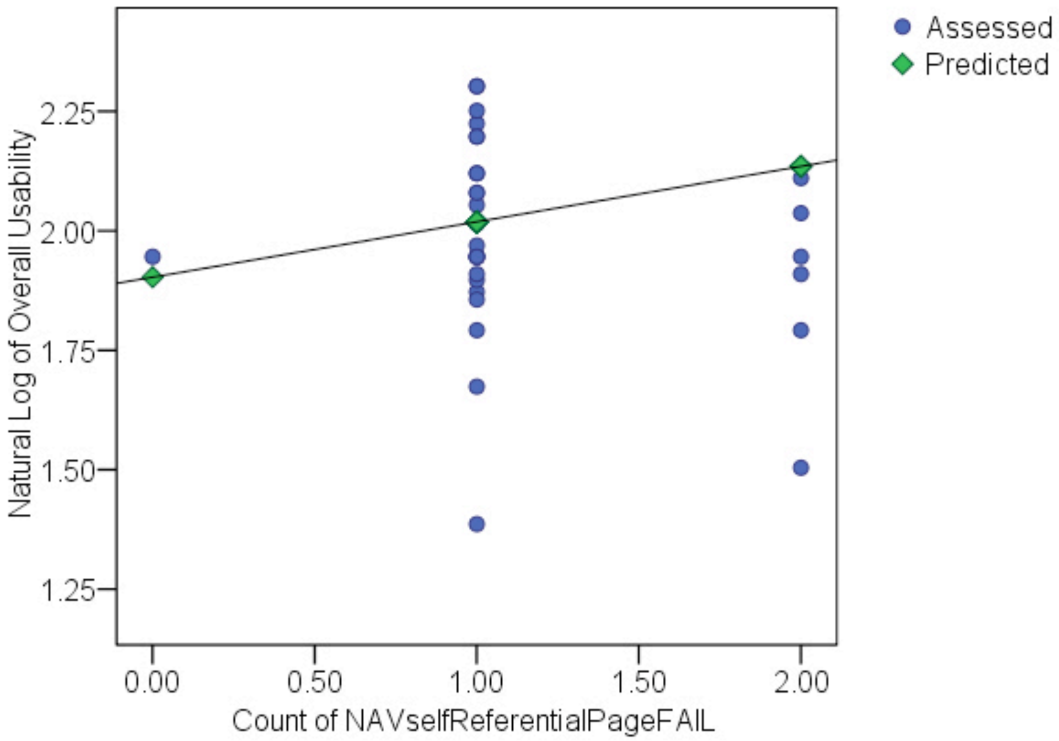
Model 11, NAVselfReferentialPageFAIL Versus OU, Scenario 1, LIFT Machine, Prediction Set (N = 31)



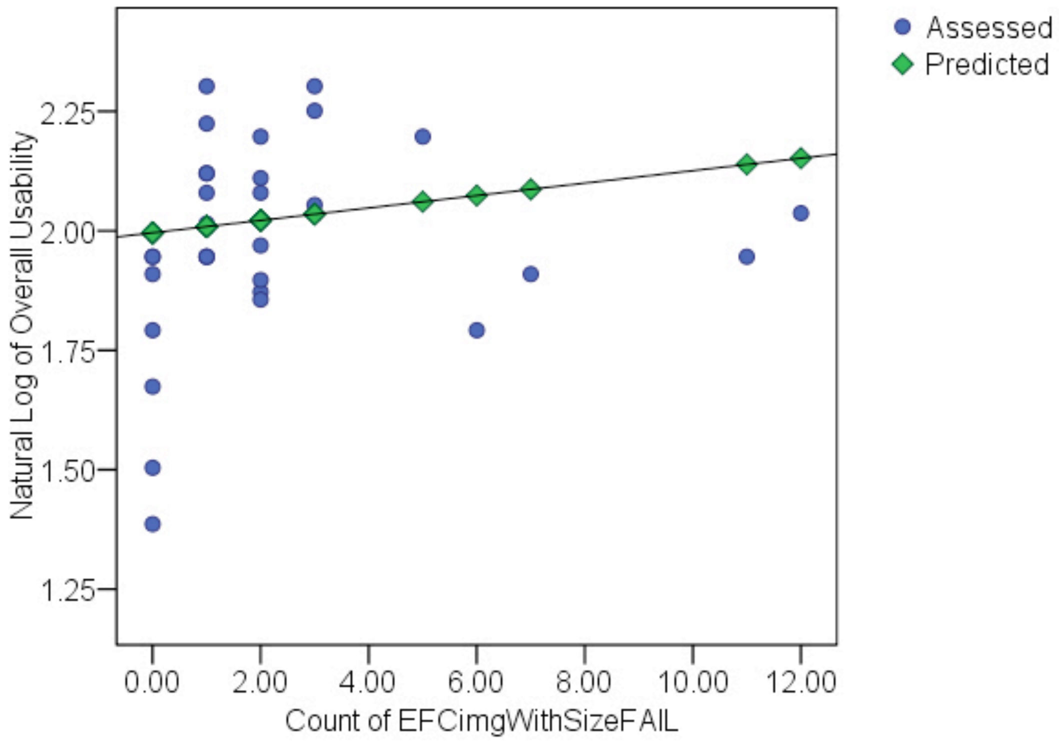
Model 12, EFCimgWithSizeFAIL Versus OU, Scenario 1, LIFT Machine, Prediction Set (N = 31)



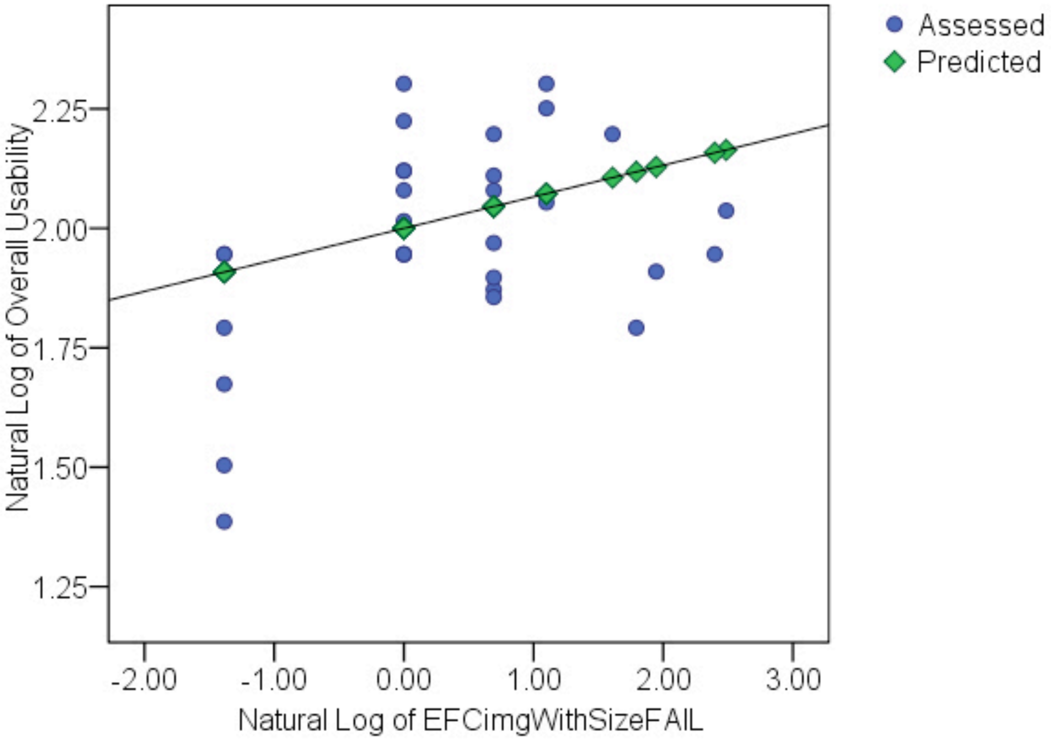
Model 13, NAVselfReferentialPageFAIL Versus LNOU, Scenario 1, LIFT Machine, Prediction Set (N = 31)

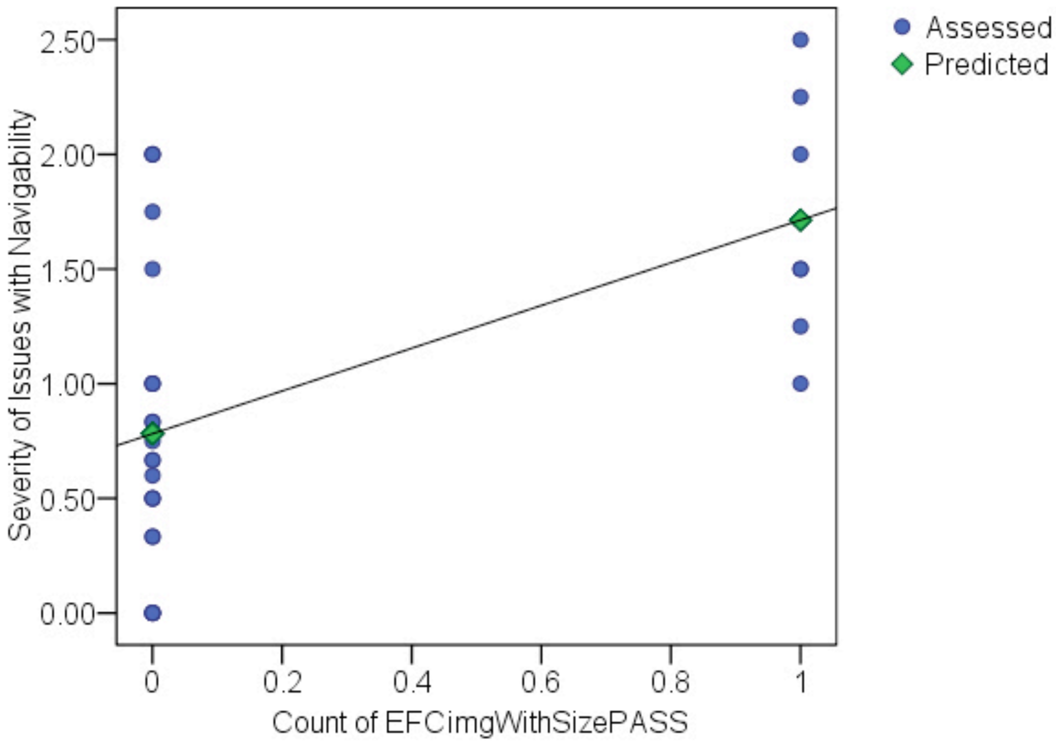


Model 14, EFCimgWithSizeFAIL Versus LNOU, Scenario 1, LIFT Machine, Prediction Set (N = 31)



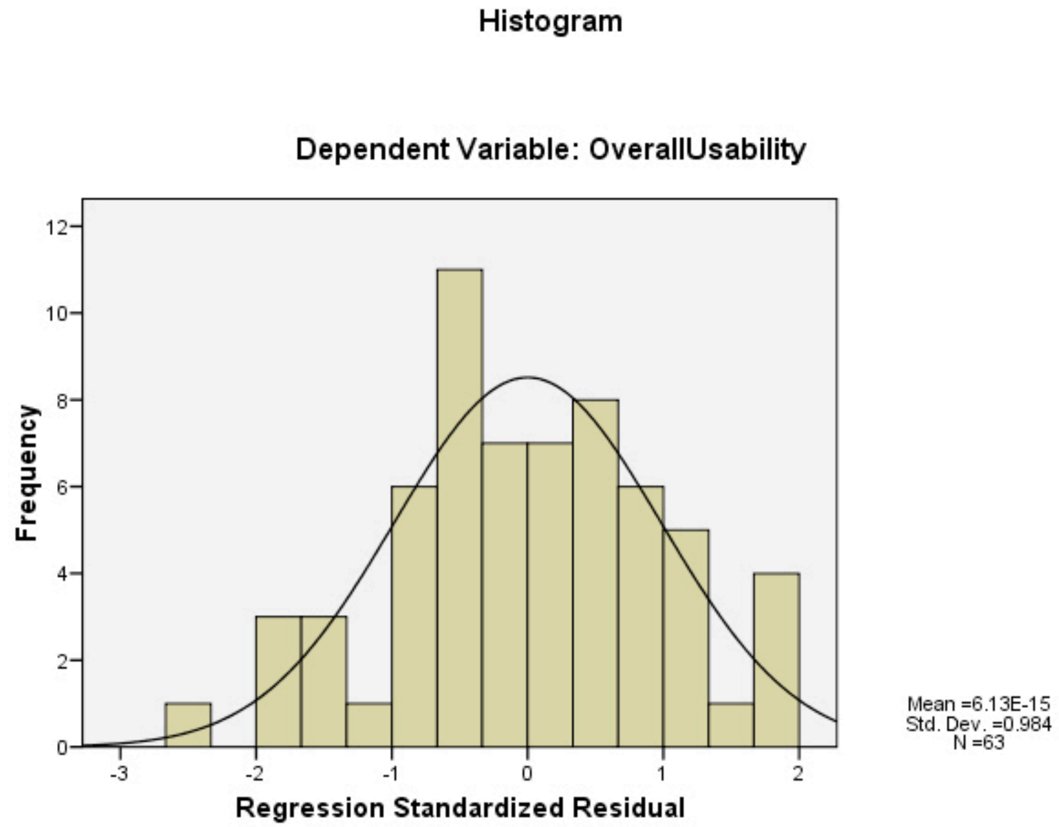
Model 15, lnEFCimgWithSizeFAIL Versus LNOU, Scenario 1, LIFT Machine, Prediction Set (N = 31)



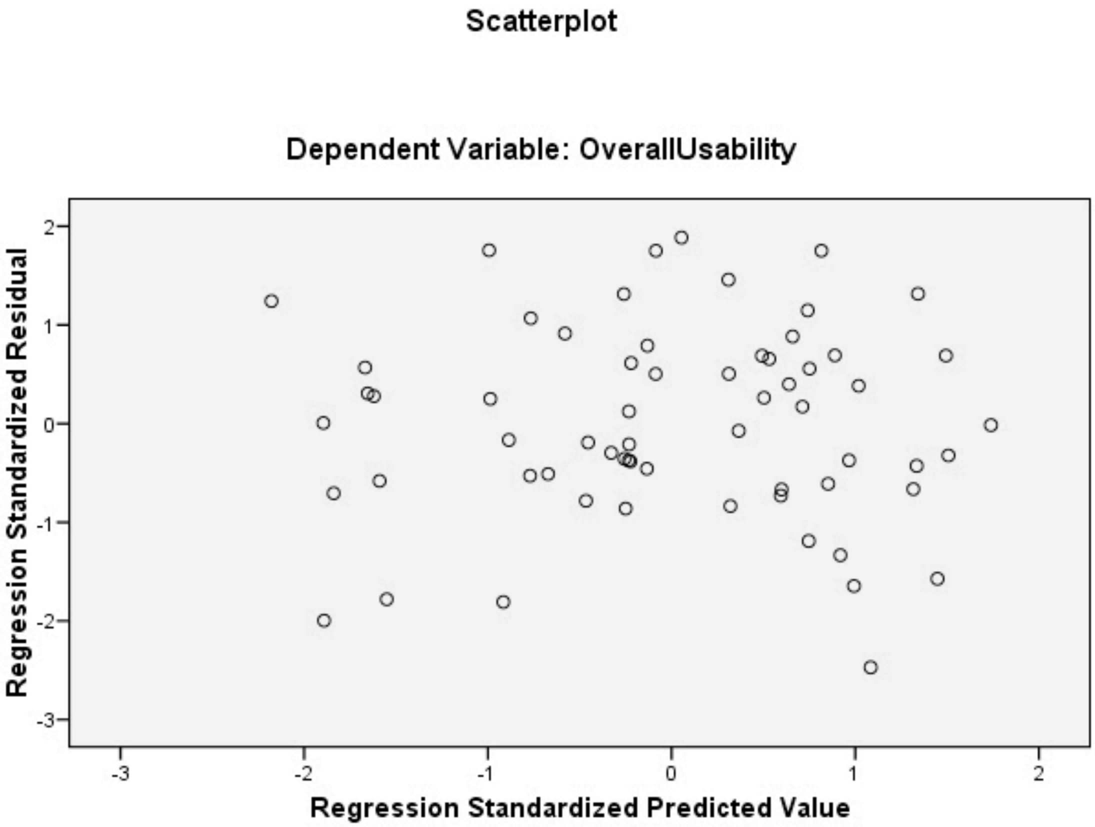


Model 16, EFCimgWithSizePASS Versus NAV, Scenario 2, LIFT Machine, Prediction Set (N = 31)

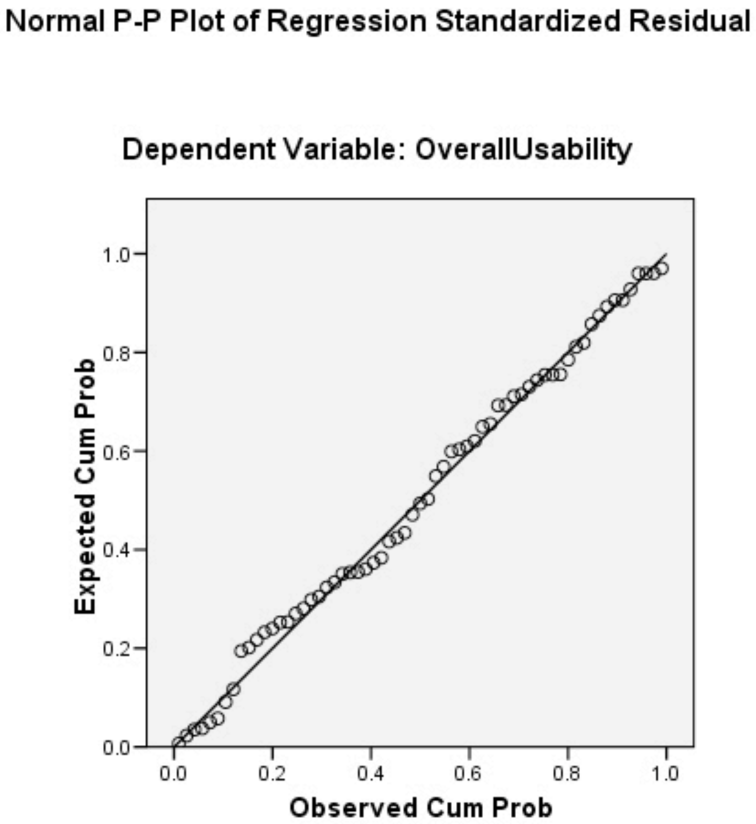
APPENDIX D



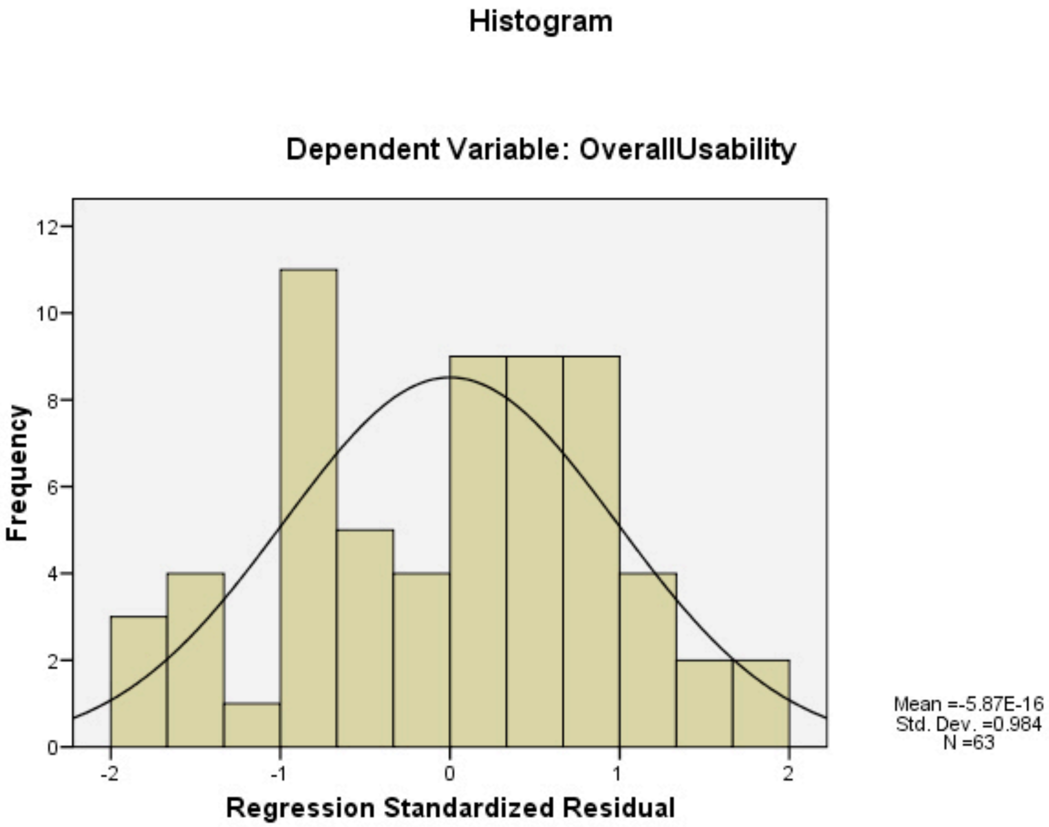
For Multivariate Regression, Scenario 1, WPMA, by SPSS v. 15 [SPSS 2006].



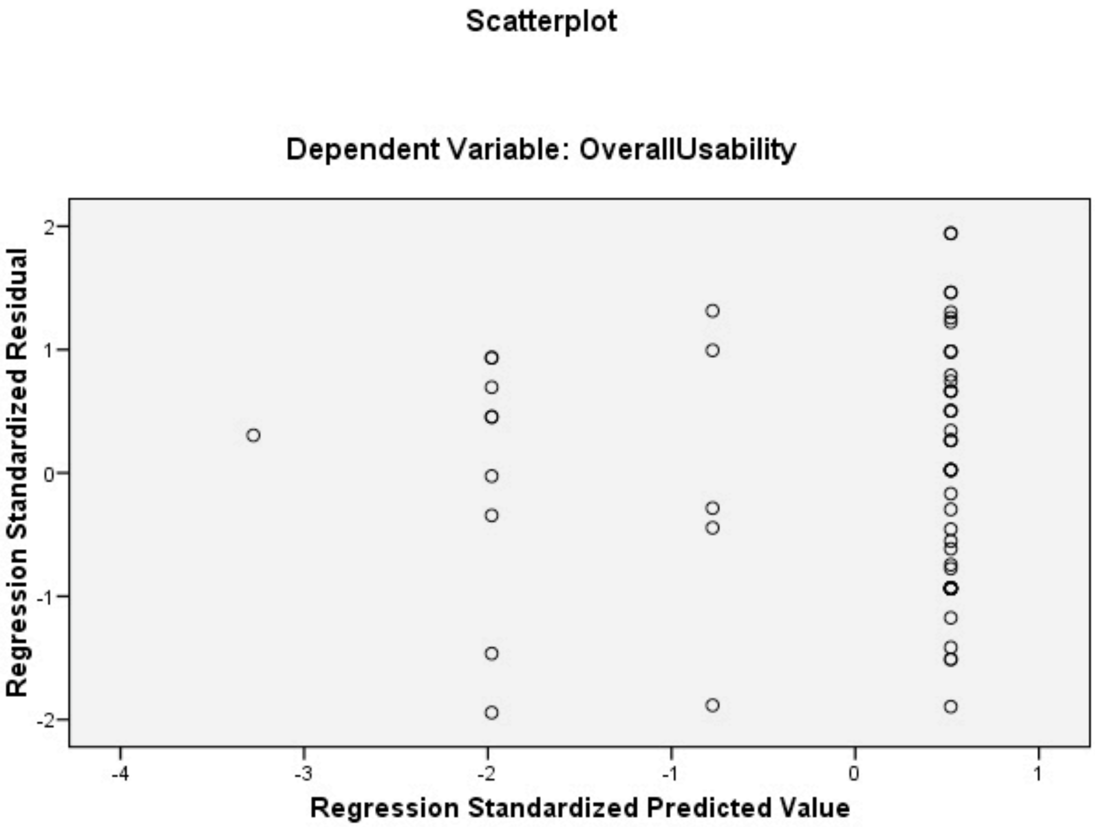
For Multivariate Regression, Scenario 1, WPMA, by SPSS v. 15 [SPSS 2006]



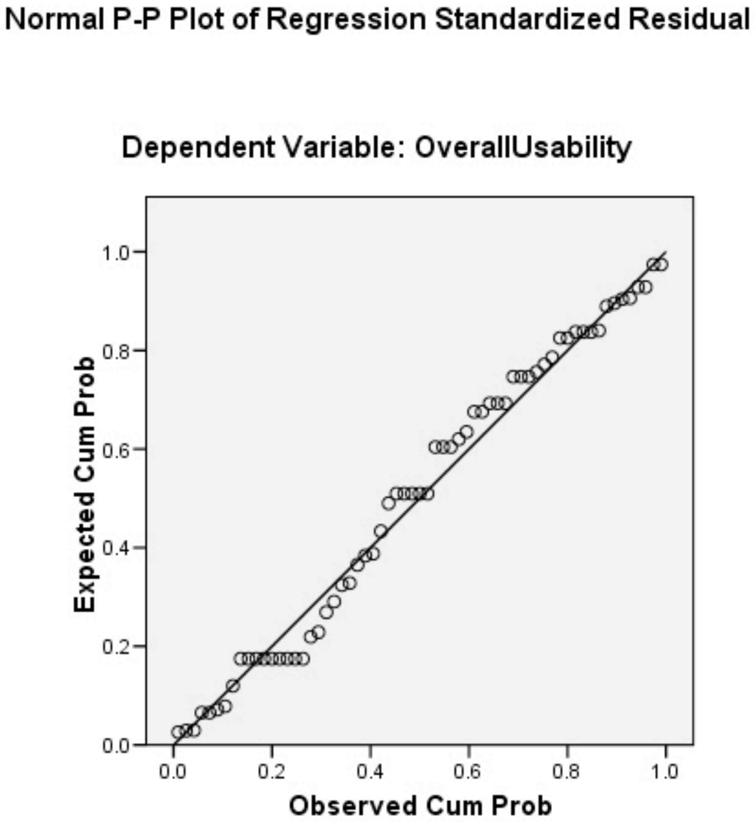
For Multivariate Regression, Scenario 1, WPMA, by linear regression, SPSS v. 15 [SPSS 2006]



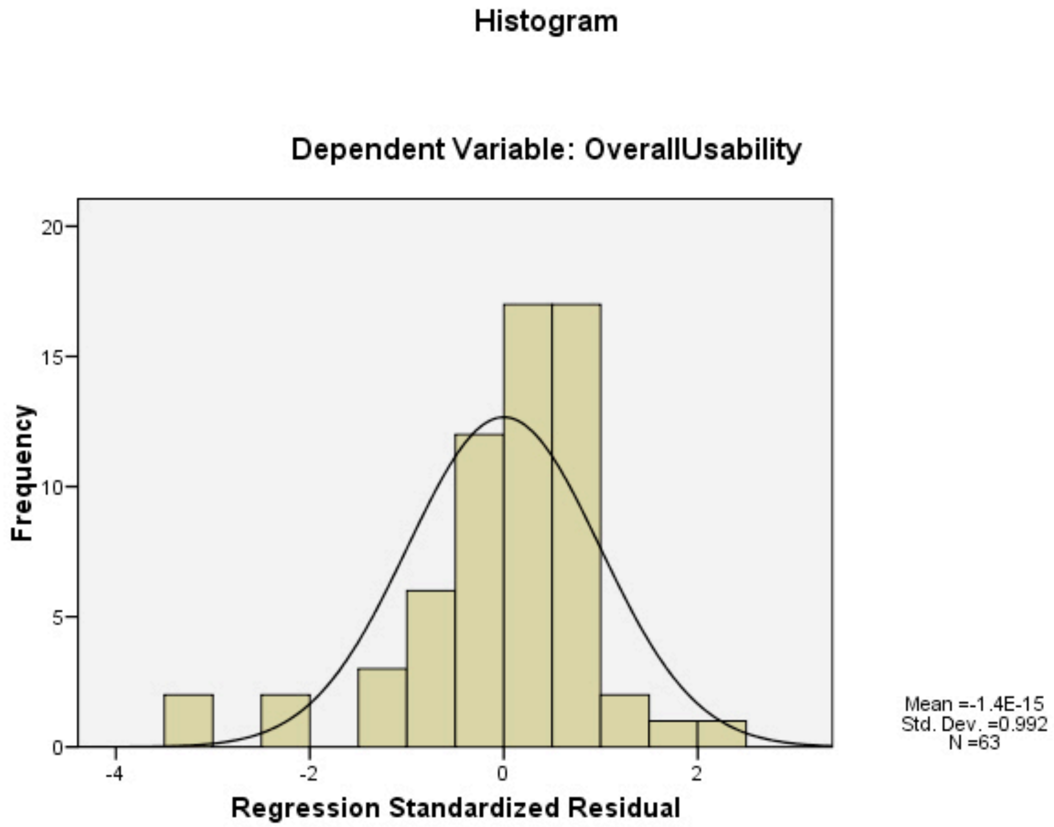
For Multivariate Regression, Scenario 1, LIFT Machine, by Linear Regression, SPSS v. 15



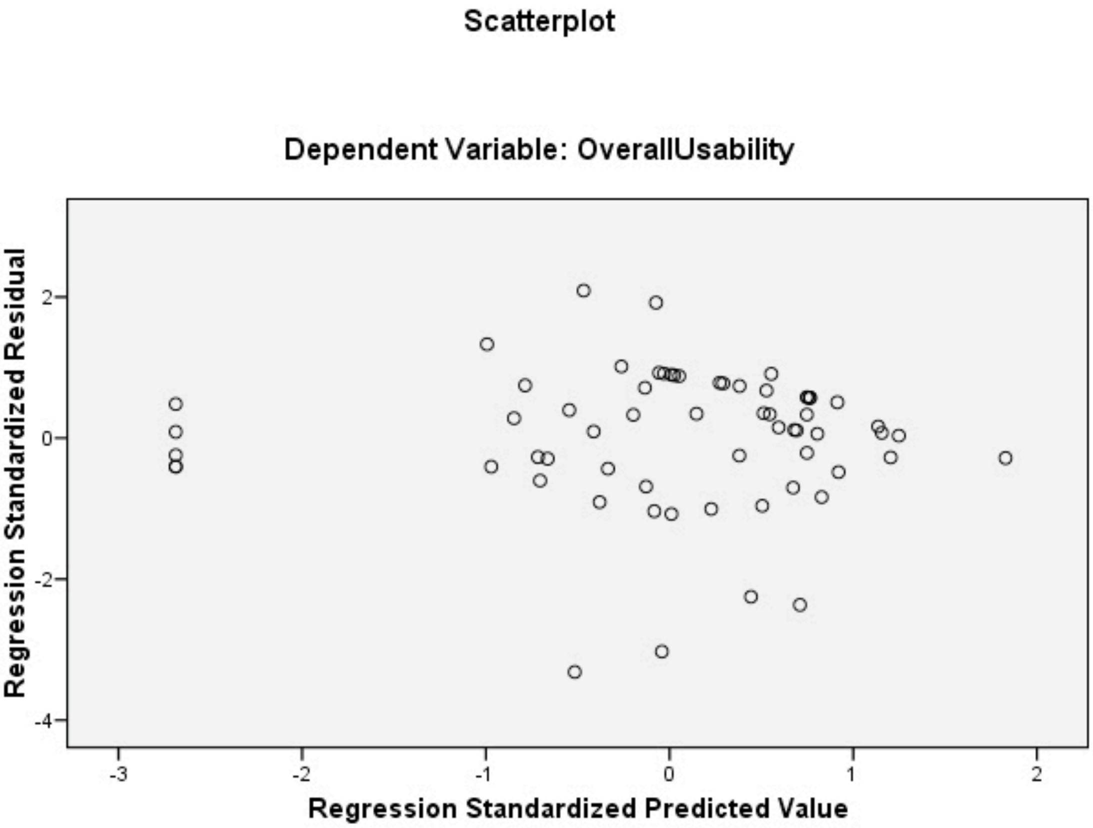
For Multivariate Regression, Scenario 1, LIFT Machine, by SPSS v. 15



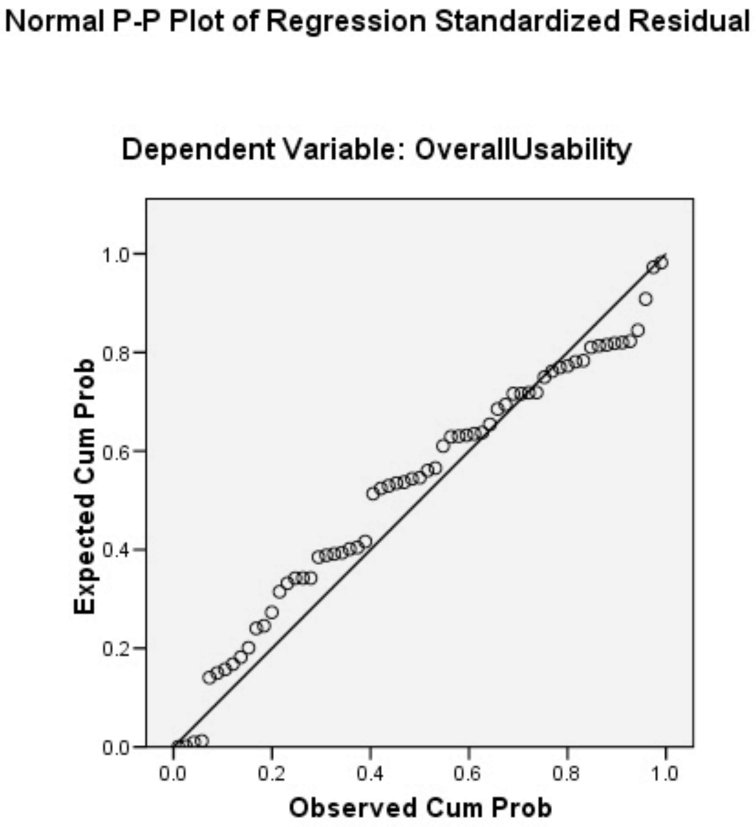
For Multivariate Regression, Scenario 1, LIFT Machine, by Linear Regression, SPSS v. 15



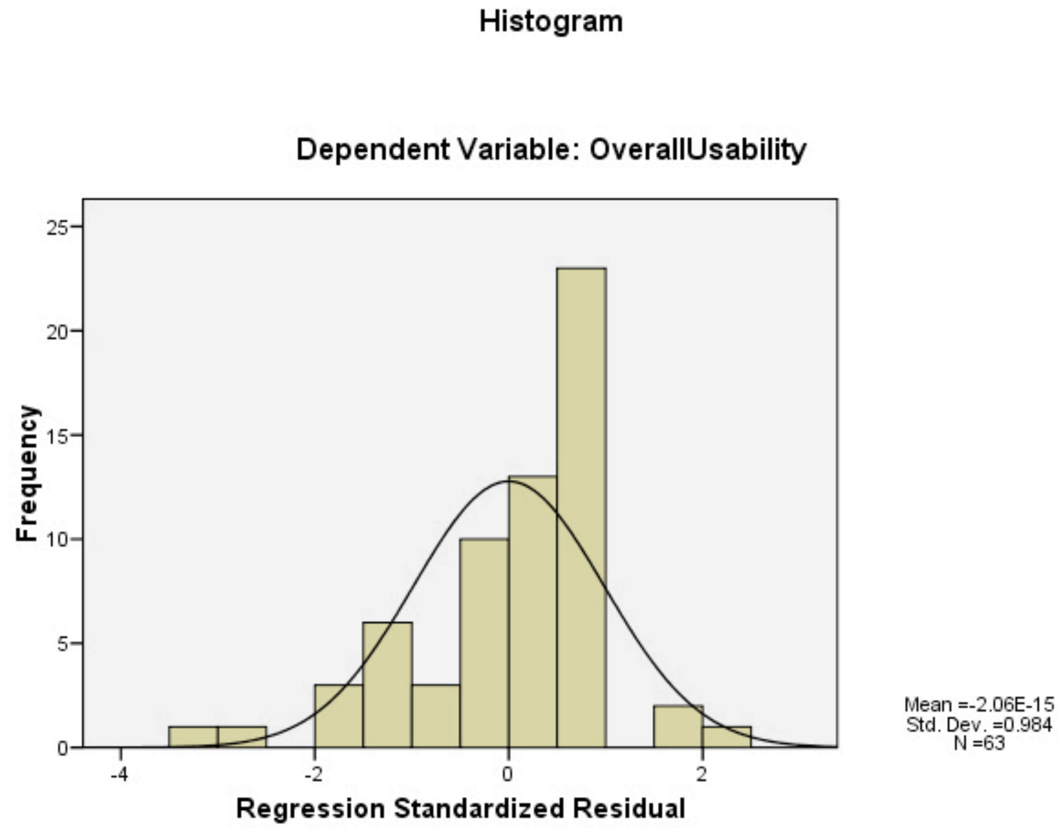
For Multivariate Regression, Scenario 2, WPMA, by Linear Regression, SPSS v. 15



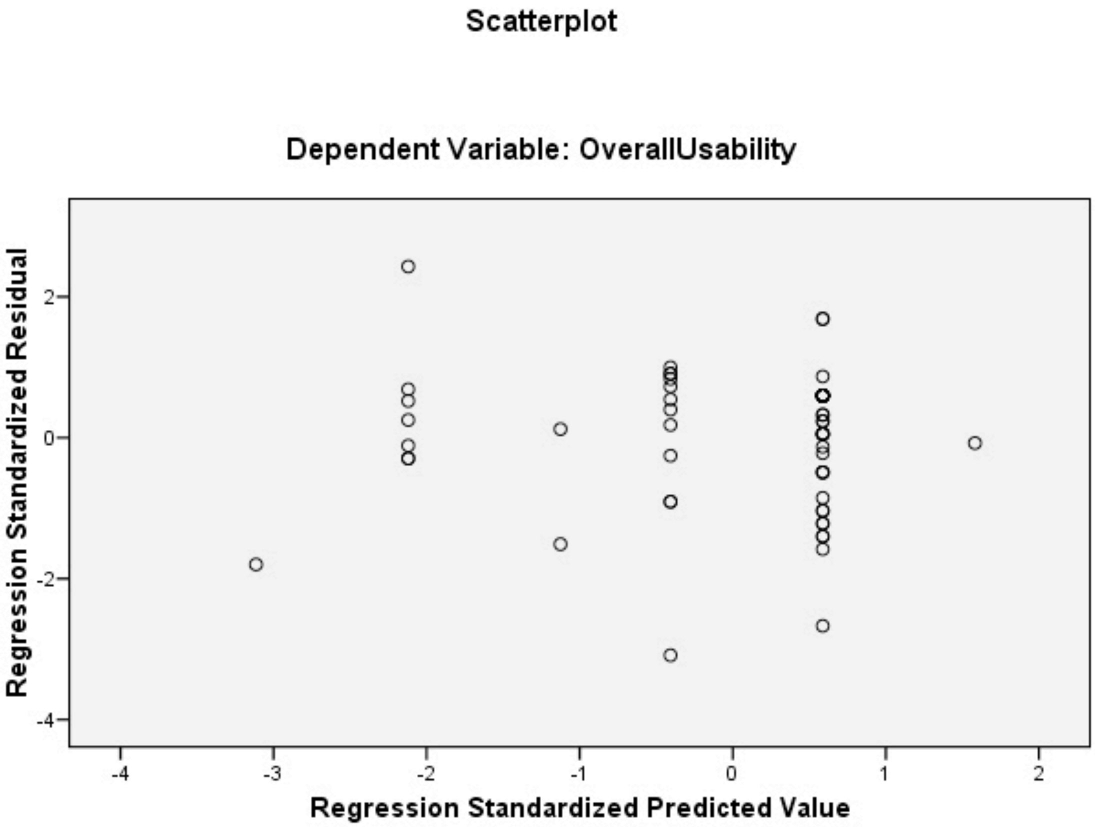
For Multivariate Regression, Scenario 2, WPMA, by Linear Regression using SPSS v. 15.



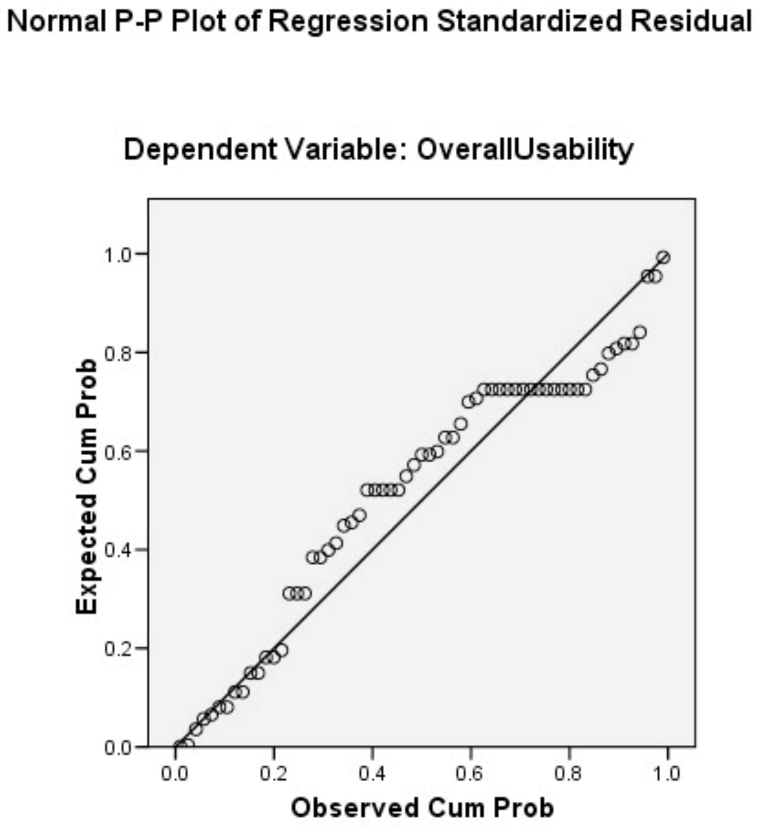
For Multivariate Regression, Scenario 2, WPMA, by Linear Regression, SPSS v. 15.



For Multivariate Regression, Scenario 2, LIFT Machine, by Linear Regression, SPSS v. 15.



For Multivariate Regression, Scenario, LIFT Machine, by Linear Regression with SPSS v. 15.



For Multivariate Regression, Scenario 2, LIFT Machine, by Linear Regression with SPSS v. 15.

APPENDIX E

Pages Not Processed by WPMA

The WPMA miscounted properties for two pages (PageIDs 4 and 13):

PageID 4 –
C:\Experiment\Hubble2\Hubble\Dissertation\hubblesite.org\explore_astronomy\hubbles_universe\index.html

PageID 13 –
C:\Experiment\Hubble2\Hubble\Dissertation\hubblesite.org\explore_astronomy\skywatch\index.html

The software tool did not measure the following properties correctly:

1. Number of words not in body of Web page (NBW)
2. Number of text areas with a non-white background, with borders, with a horizontal rule, or as a list (CLU)
3. Number of times blocks of text were not positioned flush left (NFL)
4. Number of internal and external hypertext links (LNK)
5. Number of images embedded in a page (not in scripts, applets, or objects) (IMG)
6. Percentage of Non-body Words to Total Words (PNBW)
7. Total of sizes of all files of images of a Web page (TIFS)
8. Percent of Image File Size to Web Page File Size (PIWPFS)

The reason for the error with measurements was code for embedded video found in the two Web pages above.

Pages Not Processed by LIFT Machine

The LIFT Machine did not analyze four Web pages (PageIDs 29, 31, 48, and 54) assessed by subjects in experiments.

PageID 29 –

hubblesite.org\newscenter\archive\browse\index.html.

LIFT processed –

hubblesite.org/newscenter/archive/releases/image_category/*
hubblesite.org/newscenter/archive/releases/index/0
hubblesite.org/newscenter/archive/releases/miscellaneous/*
hubblesite.org/newscenter/archive/releases/nebula/*
hubblesite.org/newscenter/archive/releases/news_nugget
hubblesite.org/newscenter/archive/releases/solar-system/*
hubblesite.org/newscenter/archive/releases/star/*
hubblesite.org/newscenter/archive/releases/star-cluster/*
hubblesite.org/newscenter/archive/releases/survey/*
hubblesite.org/newscenter/archive/releases/video_category/*
hubblesite.org/newscenter/archive/releases/

PageID 31 –

hubblesite.org\newscenter\hubble_on_the_go\inbox_astronomy\index.html

LIFT processed –

hubblesite.org/newscenter/hubble_on_the_go/inbox_astronomy/help/
hubblesite.org/newscenter/hubble_on_the_go/inbox_astronomy/mailsample.html

PageID 48 –

hubblesite.org\reference_desk\glossary\index.html

LIFT processed –

hubblesite.org/reference_desk/glossary/index.php?range=c-d
hubblesite.org/reference_desk/glossary/index.php?range=h-k
hubblesite.org/reference_desk/glossary/index.php?range=o-p
hubblesite.org/reference_desk/glossary/index.php?range=t-z
hubblesite.org/reference_desk/glossary/index.php?topic=topic_astronomy
hubblesite.org/reference_desk/glossary/index.php?topic=topic_galaxies
hubblesite.org/reference_desk/glossary/index.php?topic=topic_light
hubblesite.org/reference_desk/glossary/index.php?topic=topic_physics
hubblesite.org/reference_desk/glossary/index.php?topic=topic_stars

PageID 54 –

hubblesite.org\gallery\movie_theater\index.html

*LIFT processed –*None

Procedure for Converting Data for Statistical Processing

Tests for regression use original values of predictors as well as their values transformed to natural logs [De Veaux, Velleman and Bock 2008]. Visual inspection of predictors plotted against response variables may suggest curved as well as linear patterns. Some assessed ratings and measurement from the software tools have the value of zero. In this case, tests that use the natural log of zero substitute 0.25 for zero. For assessed ratings, $1\text{x}10^{-9}$ replaces zero if using the natural log of values.

APPENDIX F

Procedure for Making a Hubble Data File

1. Open IE and navigate to the Web page to process.
2. At command line in Paul/Documents/workspace/WebPageMetricAnalyzer/bin enter the following:

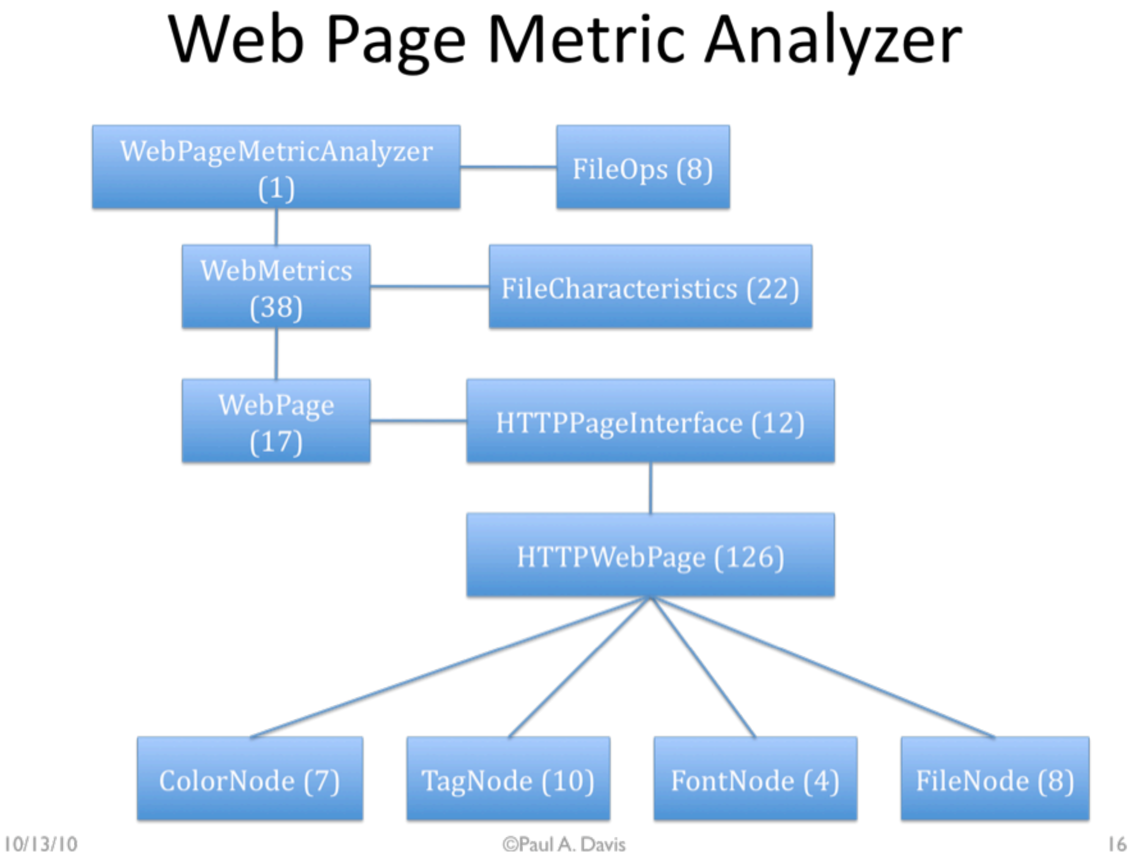
   ```
   java –classpath .
   webPageMetricAnalyzer.WebPageMetricAnalyzer <path and
   name of file> > hubble_<page_no>.dat
   ```

3. Select <path and name of file> from the address line of IE. Select Edit and Copy from the menu. Click on the command window. Select Command Prompt icon, Edit, and Paste. Use the address of the page exactly as it appeared in the Address line of the browser.
4. Using the number of the Hubble page, fill out `hubble_<page_no>.dat`
5. Hit enter and let the program run.
6. Open a text editor, navigate to the data file, and open it. If image file names and sizes appear, then it's good to use. If they don't, you'll have to work with the address line as shown in the browser. Be aware that the program is not written to parse the address line of some browsers, such as Safari earlier than version 5.

APPENDIX G

Web Page Metric Analyzer V. 1.1 Files, Source Lines of Code (SLOC), and Comments

File Name	SLOC and comments
HTTPWebPage.java	4923
FileCharacteristics.java	529
WebMetrics.java	339
WebPageMetricAnalyzer.java	236
FileOps.java	133
WebPage.java	102
TagNode.java	46
FileNode.java	41
HTTPPageInterface.java	38
ColorNode.java	31
FontNode.java	21
Total	6439



Web Page Metric Analyzer

WebPageMetricAnalyzer (1)

FileOps (8)

WebMetrics (38)

FileCharacteristics (22)

WebPage (17)

HTTPPageInterface (12)

HTTPWebPage (126)

ColorNode (7)

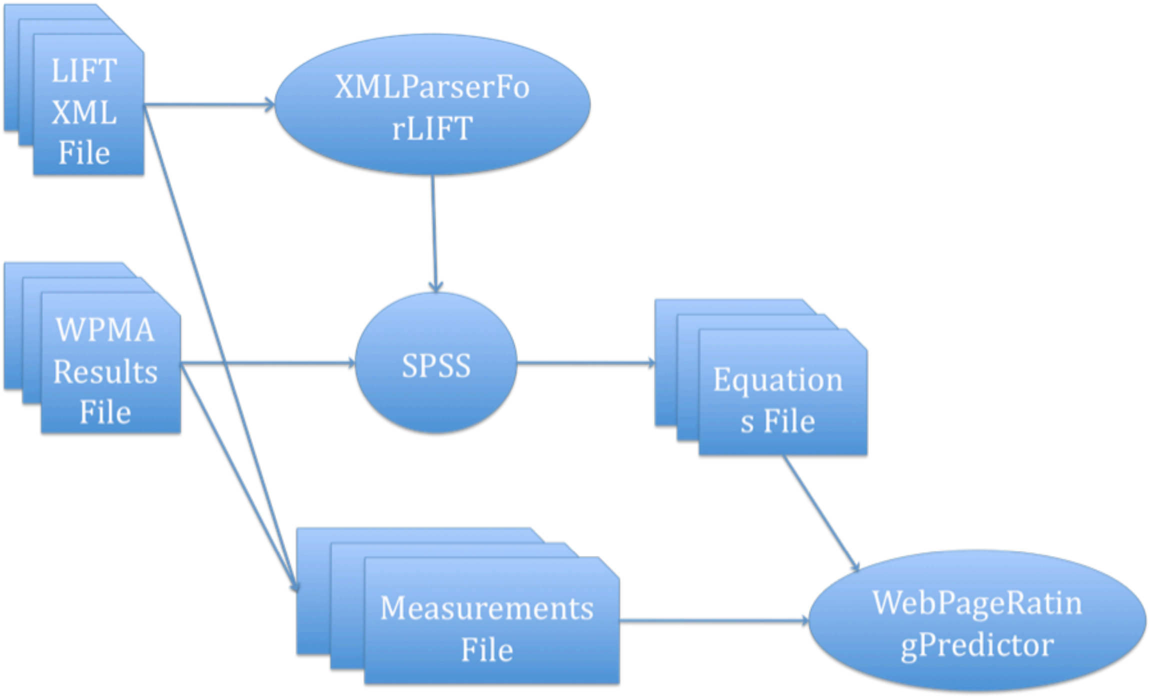
TagNode (10)

FontNode (4)

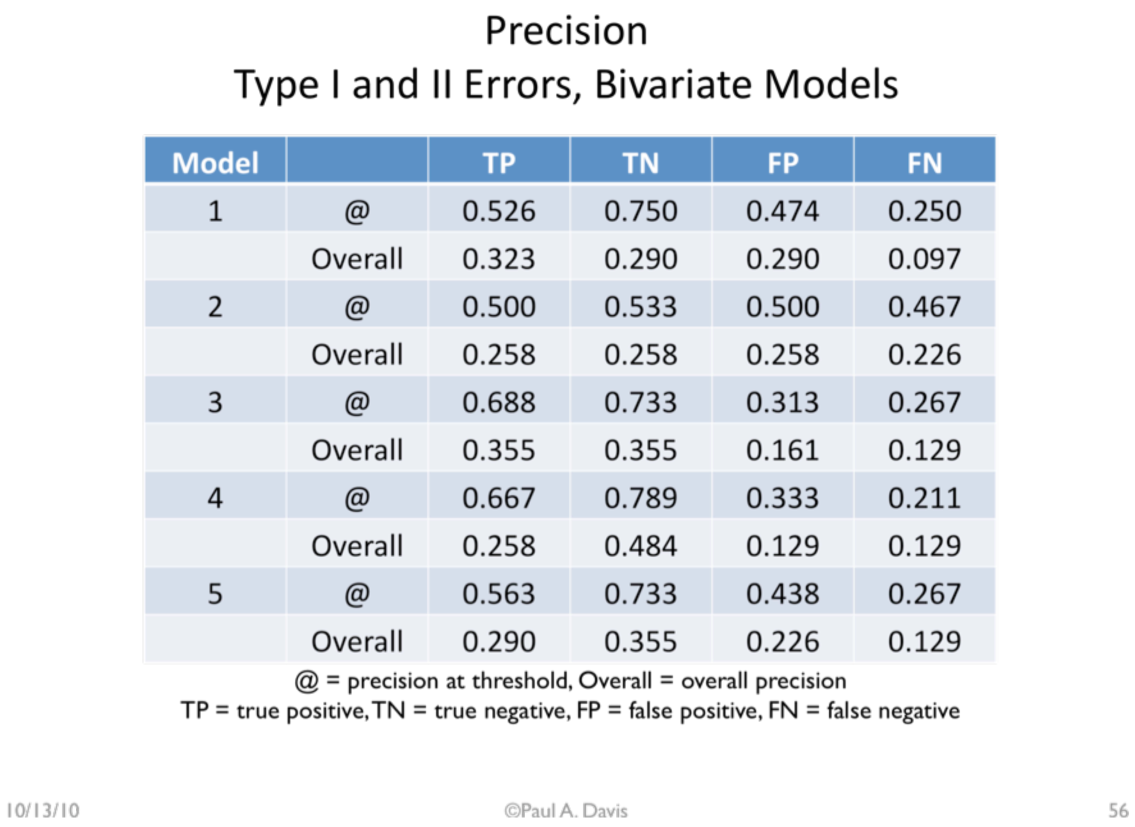
FileNode (8)

10/13/10 ©Paul A. Davis 16

Processing of Data



10/13/10 ©Paul A. Davis 17

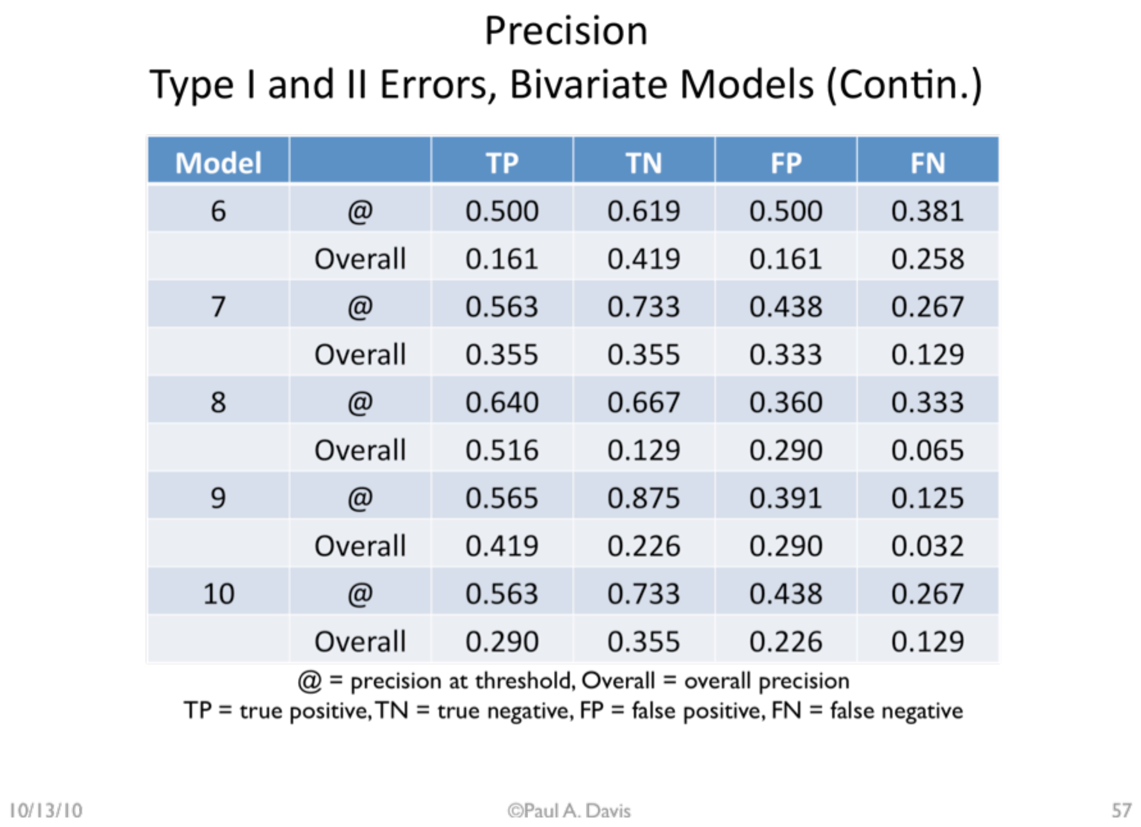


Precision
Type I and II Errors, Bivariate Models

Model		TP	TN	FP	FN
1	@	0.526	0.750	0.474	0.250
	Overall	0.323	0.290	0.290	0.097
2	@	0.500	0.533	0.500	0.467
	Overall	0.258	0.258	0.258	0.226
3	@	0.688	0.733	0.313	0.267
	Overall	0.355	0.355	0.161	0.129
4	@	0.667	0.789	0.333	0.211
	Overall	0.258	0.484	0.129	0.129
5	@	0.563	0.733	0.438	0.267
	Overall	0.290	0.355	0.226	0.129

@ = precision at threshold, Overall = overall precision
TP = true positive, TN = true negative, FP = false positive, FN = false negative

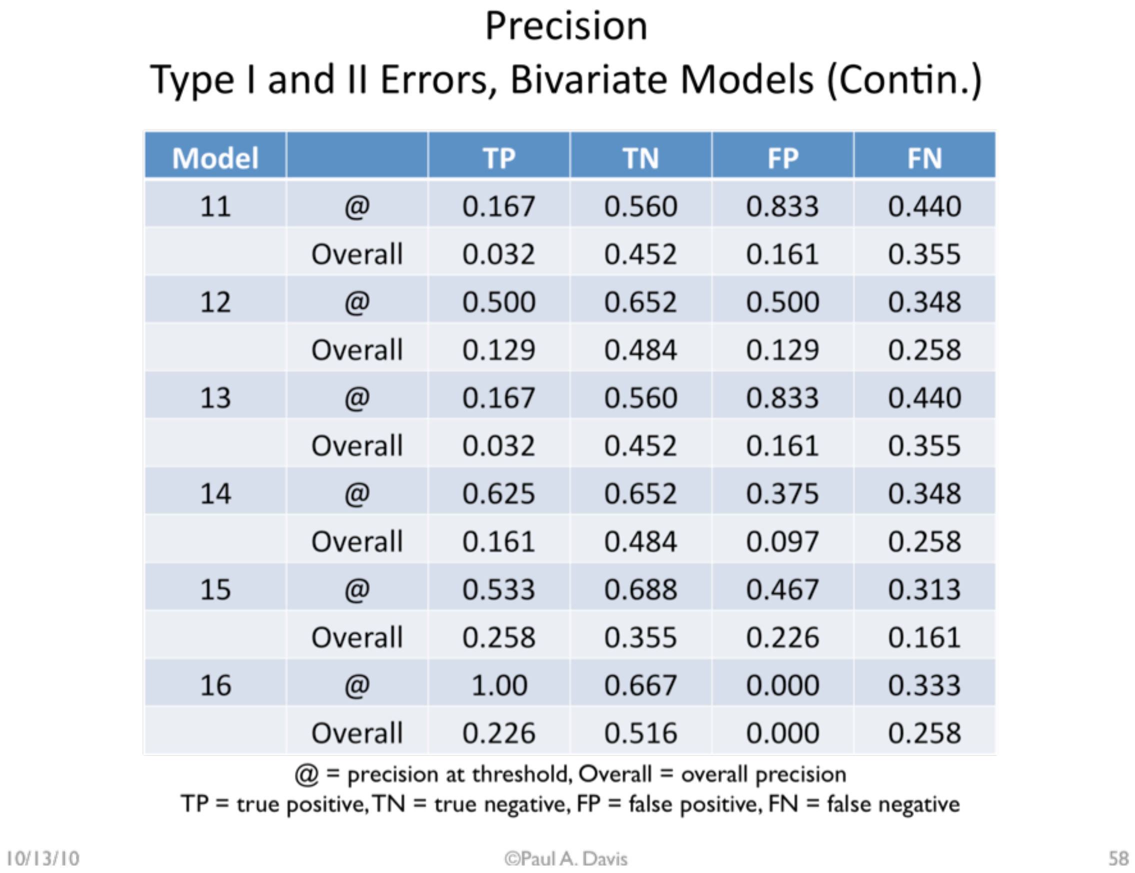
10/13/10 ©Paul A. Davis 56



Precision
Type I and II Errors, Bivariate Models (Contin.)

Model		TP	TN	FP	FN
6	@	0.500	0.619	0.500	0.381
	Overall	0.161	0.419	0.161	0.258
7	@	0.563	0.733	0.438	0.267
	Overall	0.355	0.355	0.333	0.129
8	@	0.640	0.667	0.360	0.333
	Overall	0.516	0.129	0.290	0.065
9	@	0.565	0.875	0.391	0.125
	Overall	0.419	0.226	0.290	0.032
10	@	0.563	0.733	0.438	0.267
	Overall	0.290	0.355	0.226	0.129

@ = precision at threshold, Overall = overall precision
TP = true positive, TN = true negative, FP = false positive, FN = false negative



Precision
Type I and II Errors, Bivariate Models (Contin.)

Model		TP	TN	FP	FN
11	@	0.167	0.560	0.833	0.440
	Overall	0.032	0.452	0.161	0.355
12	@	0.500	0.652	0.500	0.348
	Overall	0.129	0.484	0.129	0.258
13	@	0.167	0.560	0.833	0.440
	Overall	0.032	0.452	0.161	0.355
14	@	0.625	0.652	0.375	0.348
	Overall	0.161	0.484	0.097	0.258
15	@	0.533	0.688	0.467	0.313
	Overall	0.258	0.355	0.226	0.161
16	@	1.00	0.667	0.000	0.333
	Overall	0.226	0.516	0.000	0.258

@ = precision at threshold, Overall = overall precision
TP = true positive, TN = true negative, FP = false positive, FN = false negative

APPENDIX H

Thank you for volunteering for this research.

There are three documents attached to this e-mail. One is an information sheet that lists categories of usability. The other is a Consent Form. The third is the document you are reading.

Before starting the experiment, you'll need to do the following things:

1. Read these attached documents before coming to the experiment.
2. Practice identifying usability problems at sites listed at;
http://www.webbyawards.com (examples of good design,)
http://www.webpagesthatsuck.com/ (examples of poor design,) and
http://www.useit.com (general information about usability.)

If you agree to participate, please sign the document and bring it with you. I'll have extra copies available. If you decline to participate, please let me know as soon as possible. Please be sure to read about identifying usability problems and practice with the sites listed above.

The procedure of the experiment is as follows:

In a pre-survey interview, the Priniciple Investigator will collect your name, age within an age range, experience in years using Web browsers, academic major and classification, gender, and years of prior experience with designing Web sites.

The experiment is performed under controlled conditions. The location will be the Evans Library Annex. If a place in Evans Annex is not available, the H.R. Bright Building at 3112 TAMU or Evans Library itself are other locations. If the experiment takes place in a closed room, the PI will sit outside the room. If you allow, the experiment will be recorded. If carried out at another location, the subject might sit in a study area. In either case, with the PI available nearby. Each participant will sit at a desk before a computer, view a sequence of web pages, and answer questions on paper forms, called data forms. The forms have check or option boxes as well as text areas. The questions are about the kind and severity of usability issues they might find in the web pages they review. Each may verbalize or write observations at any time as well.

The survey ends when the person has finished analyzing all web pages or when time runs out. Time limit is two hours. There will be a post-survey interview, which may be declined. The offer of the gift card is at completion of the survey.

A Brief Introduction To Usability and Web Pages

Paul A. Davis, January 16, 2008

Usability has several definitions. Basically, it is about making a product that people find to be easy, enjoyable, and effective to use [1]. Here are some characteristics of a usable Web site:

1. Effective
   Persons visiting a Web site should accomplish their goals, such as finding information or performing an online transaction.

2. Flexible
   The Web site should allow a person to reach a goal both quickly and easily. It should give people more than one way to reach their goals. New users might want more guidance, while expert users might want less.

3. Navigable
   People should find it easy to know where they are in a Web site. For example, they should find it easy to click on hyperlinks and go where they want to.

4. Satisfactory
   A Web site should not make people tired or angry. Instead, they should find it satisfying to use.

5. Searchable
   A Web page should provide information about its content to search engines so that they can index it. To do this, the coding of the page should have page descriptions, keywords, and lists of important words for both page and Web site.

6. Efficient
   People should find the ease of use and performance of the site to be satisfactory. For example, they should be able to find pages, determine if those pages are useful for them, and find other pages in a manner that they find satisfactory.

There are many other ways to define usability. The above list of six has just a few of many possible characteristics.

If you would like to learn more about usability, see the following:

1. Rogers, Y., Preece, J., Sharp, H. Interaction Design, 2nd ed. John Wiley & Sons, Ltd. Chichester, England. 2007.

2. Nielsen, J. Usability Engineering. Academic Press, Inc. Boston, Massachusetts. 1993.

APPENDIX I

The ID number of the Web page is _______.

Web Page Usability Assessment Form

Describe the usability problem.

Select a category and rate the problem in that category.

□ Effectiveness – Persons visiting the site should be able to accomplish their goals, such as finding information.

___1 ___2 ___3 ___4 ___5

□ Flexibility – There should be more than one way to reach a goal. This also means that people should be able to use the site if the page appears in browser windows of different sizes.

___1 ___2 ___3 ___4 ___5

□ Navigability – People should find it easy to learn where they are in a site as well as how to go elsewhere within the site. They should be able to remember where a page is if they return to the site.

___1 ___2 ___3 ___4 ___5

□ Satisfaction – People should not tire or become upset when using a site. They should find the experience satisfying and satisfactory.

___1 ___2 ___3 ___4 ___5

□ Visitor efficiency – People should find that the ease of use and performance of the site are satisfactory. This can include obtaining pages, determining if those pages are useful, and finding hypertext links to other pages.

___1 ___2 ___3 ___4 ___5

Rate of the overall usability of the Web page.

___1 ___2 ___3 ___4 ___5 ___6 ___7 ___8 ___9 ___10

Optional comments.

Note: Please analyze one problem at a time per form. Write one check mark by one rating. A rating of "1" means a very weak problem. A rating of "5" means a very strong problem. You may select more than one category but be sure to select only one rating in each category chosen. For overall usability, the scale is "1" as lowest and "10" as highest.

Note: the PI instructed volunteers to write "0" in categories if they detected no issues.

SCENARIO

You have been hired as a contractor to assess the usability of a Web site. The purpose of your contract is to find usability issues within a particular context. Do not assess the performance of the site but only the usability of its Web pages. Each page has an identifying number in the title of its header.

NAME AND PURPOSE OF PRODUCT

This is the Web site for the NASA Hubble Space Telescope. The purpose of the Web site is to provide information to the public.

CONTEXT OF USE

Assess the Web site of the NASA Hubble Space Telescope for use by middle school students. These students attend classes in physical science or environmental science. They all study the exploration of space. Their ages are between 12 and 15, and there are both boys and girls. All are familiar with how to use a Web browser.

INSTRUCTIONS

Base your assessment on five categories of usability: efficiency, navigability, flexibility, satisfaction, and effectiveness. In your assessment, always record the number of the Web page at the top of the form. As you discover issues (in a particular category of usability) with a Web page, rate the severity of the issue in that category. A rating of 1 means a very minor issue, whereas a rating of 5 is a severe issue. If there is no issue in a category, write a zero.

Also remember to always rate a page for the overall quality of usability. A rating of 1 is given to pages with very lowest quality, and a rating of 10 is for pages with very highest quality.

SCENARIO

You have been hired as a contractor to assess the usability of a Web site. The purpose of your contract is to find usability issues within a particular context. Do not assess the performance of the site but only the usability of its Web pages. Each page has an identifying number in the title of its header.

NAME AND PURPOSE OF PRODUCT

This is the Web site for the NASA Hubble Space Telescope. The purpose of the Web site is to provide information to the public.

CONTEXT OF USE

Assess the Web site of the NASA Hubble Space Telescope for use by astronomers. These scientists search for information, photos, and illustrations regarding astronomical phenomena and scientific instruments. They publish to scientific journals. Their ages are between 25 and 65, and there are both men and women. All are familiar with how to use a Web browser. Their academic credentials include a Masters or Ph.D. in physics or astronomy.

INSTRUCTIONS

Base your assessment on five categories of usability: efficiency, navigability, flexibility, satisfaction, and effectiveness. In your assessment, always record the number of the Web page at the top of the form. As you discover issues (in a particular category of usability) with a Web page, rate the severity of the issue in that category. A rating of 1 means a very minor issue, whereas a rating of 5 is a severe issue. If there is no issue in a category, write a zero.

Also remember to always rate a page for the overall quality of usability. A rating of 1 is given to pages with very lowest quality, and a rating of 10 is for pages with very highest quality.

1

CONSENT FORM
Objectives Methods of Applying Subjective Analyses to Usability Testing of Web Sites

Introduction

The purpose of this form is to provide you information that may affect your decision as to whether or not to participate in this research study. If you decide to participate in this study, this form will also be used to record your consent.

You have been asked to participate in a research project that is a study into identifying usability problems in Web sites. The purpose of this study is to determine if it is possible to match measurements about characteristics of Web pages to what people say are usability problems with those pages. You were selected to be a possible participant because you have indicated that you have knowledge of using or evaluating Web sites.

What will I be asked to do?

If you agree to participate in this study, you will be asked to do the following tasks. There is a pre-survey interview where demographic information is collected. This means recording your age within an age range, your gender, your major, years of secondary education, and years of experience with browsers. If you have experience designing, making, or evaluating Web sites, it is recorded as well. Then you will sit at a desk before a computer monitor, view a sequence of Web pages, and answer questions on paper forms. The forms have check boxes as well as text areas. The questions are about the kind and importance of usability issues that you might find in these Web pages. You may write down your own observations or comments at any time as well. The study ends when you have analyzed all Web pages or when time runs out. There will be a post-study interview, which you may decline. The time needed in the study is no more than two hours. If you wish, you may return for a second survey of Web pages, but this is optional. The procedure of this second survey is identical to the first except for the Web pages to review. Your participation may be audio or video recorded.

What are the risks involved in this study?

The risks associated in this study are minimal, and are not greater than risks ordinarily encountered in daily life.

What are the possible benefits of this study?

You will receive a one-time offer of a gift card valued at $25.00 (U.S.) for participating in this study. You will receive no additional offer of a gift card even if you participate again. If the study determines that it is possible to identify pages with a higher likelihood of usability problems, a benefit to software professionals and companies is reduced time to finding such problems in Web sites.

Do I have to participate?

No. Your participation is voluntary. You may decide not to participate or to withdraw at any time without your current or future relations with Texas A&M University being affected.

2
Who will know about my participation in this research study?
This study is confidential. The records of this study will be kept private. No identifiers linking you to this study will be included in any sort of report that might be published. Research records will be stored securely and only Paul A. Davis will have access to the records. If you choose to participate in this study, you may choose to be audio [/video] recorded. Any audio [/video] recordings will be stored securely and only Paul A. Davis or Frank Shipman will have access to the recordings. Any recordings will be kept for no more than two years and then erased.

Is there anything else I should consider?
You may withdraw from the study at any time.

Whom do I contact with questions about the research?
If you have questions regarding this study, you may contact the Principal Investigator Paul A. Davis at 979-574-6244 or p-davis@tamu.edu or, alternatively, Frank Shipman at 979-862-3216 or shipman@cs.tamu.edu.

Whom do I contact about my rights as a research participant?
This research study has been reviewed by the Human Subjects' Protection Program and/or the Institutional Review Board at Texas A&M University. For research-related problems or questions regarding your rights as a research participant, you can contact these offices at (979)458-4067 or irb@tamu.edu.

Signature
Please be sure you have read the above information, asked questions and received answers to your satisfaction. You will be given a copy of the consent form for your records. By signing this document, you consent to participate in this study.
______ I agree to be audio [/video] recorded.
______ I do not want to be audio [/video] recorded.

Signature of Participant: ___________________________________
Date: ______________

Printed Name:
____

Signature of Person Obtaining Consent: ________________________
Date: ______________

Printed Name:
____

What are the dimensions of the Hubble Space Telescope?

What is the latest news about the telescope and the project itself?

Of what in space has the telescope taken images most recently?

Where can you listen for news about the project?

What is visible in tonight's sky?

Where is the Space Science Education Resource Directory?

Where can news agencies and the press find information about the telescope?

Where can you find calculators to find temperature, distance, and redshift?

Where can you find the latest news from NASA?

How do you go about building your own Hubble telescope?

APPENDIX J

```
/**
 * Program started:  5/16/08.
 * Program completed:  9/22/08.
 * Preliminary tests with atest suite:  9/26/2008
 * Program tested with 30 file mtest suite:  9/30/2008
 * Added check for CSS files by @import:  11/12/2008
 * This program is written to parse hubblesite.org, dated August 20, 2007.
 * It does not use DOM.  A copy was made available to Frank M. Shipman,
 * Professor, Dept. of Computer Science and Engineering, Texas A&M University.
 *
 * Designer & Programmer:  Paul A. Davis
 * Purpose:  Web Page Metric Analyzer tool for research for dissertation.
 *
 *  History:
 *  10/28/08.  Added flag to track elements not used in measurements.
 *
 *
 * This version of the webPageMetricAnalyzer was created in 05/16/08. A previous version
 * was called webmetric and was created 11/5/2007.
 *
 *
 * 1. 06/10/08. Word count -- count of total number of words on a page, which
 * includes headers in cells of tables, so include contents of tables.
 *         06/10/08 Exclude scripts.
 *
 *         NOTE:  characters formatted by tags but not separated by spaces are counted
 * as separate words even if they are adjacent and _appear_ as one word.
 *
 * 2. 09/10/08 Body text % -- percent of words in non-body versus body words.
 *
 * 3. 05/16/08. Emphasized body text.
 *         a. Exclamation point
 *         b. Words with all characters capitalized but not consisting of all digits
 *         c. Words capitalized by an inline style uppercase but not consisting of all digits
 *         d. Words in strong or bold face
 *
 * 4. 05/25/08. Emphasized body text % -- count of words in bold, capitalized,
 * or adjacent to '!' versus count of body text
 *
 * 5. 04/27/08. Count words made bold by deprecated tag.
 * Count worlds in all caps and bold by deprecated symbol "<b>".
 *
 * 6. 09/06/08. Text positioning count -- count of number of times that text position
 * wasn't flush left with the margin in body.
 *         a.  Do not count text appearing in a table cell.  This is counted in text cluster count.
 *         b.  Text wrapped to right and around an image.  This is not done until to do so is found
 *               in documentation about the Ivory and Hearst, 2001 tool.
 *         c.  Blocks of text.  BLOCKQUOTE, PRE.
*       e.  Code.  1) attribute/value of "align=center" and "align=right" for non-style and
*                     2) inline-style "text-align:center" and "text-align:right"
*
* 7. 09/05/08. Text cluster count -- "Text areas highlighted with color, bordered regions,
*   rules or lists."
*       To do this, count of number of times of change in text color,
*       number of horizontal rules, and number of lists.  A region is a part of the
*       page with a background colored differently from the rest of the page.  Within
*       this region, text must appear.
*       Specifics, all of which are arbitrary and made by Paul A. Davis:
*             a. Text with a different color of background than rest of document.
*                 1)  08/20/08, by default. Exclude bgcolor in <BODY> because this attribute
changes the
*                      background color of all of the page.
*                 2)  08/20/08.  Inline style of "background-color" works.  Exclude attributes
"bgcolor" and
*                      "background-color" in other elements.   See line 830
*                 3)  08/20/08, by default.  Exclude background color of table cells.
*             b. Text surrounded by border includes the following:
*                 1)  08/25/08.  For inline style, "border" will draw a line around text.
*                 2)  09/04/08. Text within a TABLE
*                      a) Ignore "border".
*                      b) Ignore "bgcolor".
*                      c) Ignore "frame" which is considered for whole documents and not text
within
*                           the same document.
*                      d) Ignore "rules" because rules applies between cells.
*                      e) Consider any table with any number of cells.
*                      f) Count a cluster of text as three or more words in the same cell
*                           and at most one of the words is a number.
*                           NOTE: this does not consider text separated by titles or more than one
*                           blank line within the same cell.
*             c. 09/05/08. Horizontal rule <HR> or <HR />
*             d. 09/05/08. Any block starting with <UL>, <OL>, or <DL>
*
* 8. 05/18/08. Link count -- count of all hypertext links on page.
*
* 9. 08/15/08. Page, i.e. file, size -- total bytes for the page as well as for images.
*       Include CSS file sizes in total if those files can be read.
*       08/11/08.  Class FileCharacteristics.java creates linked list of FileNode objects,
*       and these have path and name of image files.  Successful test of storing names
*       of image files and calling a method to sum their file sizes.
*       NEXT.  Include path and name of HTTP page.  Create method to sum file sizes
*       of HTTP and image files together and a different method for image files only.
*
*       9/22/08 - CSS files included in total count.  This includes those files stored
*             locally as well as URLs to those CSS files.  The fact that the CSS file is
*             at a URL is noted.  The program doesn't try to retrieve it but will report
*               the size as 0.
*
* 10. 08/14/08. Graphic % -- percent of bytes of graphic images to page size.
*         Method made to calculate percentage of file size of image files to sum of sizes
*         of image and HTTP files.
*         10/156/08.  In test against NASA Hubble site, some image files were found to have
*         "file:///" as prefix.  Modified readWord() to continue reading characters
*         from the input buffer if the image flag is set and a colon is encountered.
*
* 11. 05/25/08. Graphics count -- count of number of images on a page.
*
* 12. 08/06/08. Color count -- count of total number of colors used in fonts.
* This is not a count of the number of words in that color.
*
* ASSUMPTION.  The program will count only one color as default for a hypertext link.
* Process the page with Private Data and History cleared in the browser so that visual inspection
* of hypertext links appear as they would before a person had visited them.
* Also, if the program encounters a hypertext link, it will add the default color
* to the list of text colors.  It will also mark that color as different from
* any other color that might the program might find in the page.
*
* 13. 05/31/08 Deprecated font -- color of font as set in style element.
*
* Color of hypertext link does not supercede color set by enclosing style or deprecated tag.
* It can supercede color as set in body tag.
*
*
* 14. 09/21/08 Font count -- count of total number of font face and size
* combinations used.
*
* Approach:  set global flags for font-face and size, just as was done with color.
* When font face is found, store in nested tag node.  Same for font size.
* Report results from contents of nested tag node list.  Use bFontFace and bFontSize
* as the global flags.
*
* 15. 06/09/08 Javadoc. Converting comment headings of methods into
* Javadoc.
*
* A method prefixed by "test" checks if flags are set and does not inspect the
* word itself. If flags are set, it signals increment of a counter.
*
* A method prefixed by "check" detects if chars of the word match a pattern,
* and, if it does, signals increment of a counter.
*
*
*/
```

VITA

Paul Arnold Davis has received three Bachelor of Science, a Masters, and a Ph.D. in several subjects. The Bachelor degrees are Bachelor of Science in Zoology from The University of Texas at Austin in 1978, Bachelor of Science in Biochemistry from The University of Texas at Arlington in 1983, and Bachelor of Science in Computer Science from Texas A&M University in 1990. Texas A&M University awarded him a Master of Computer Science in 1994. He was a software developer in industry and at a research institute (TTI) before returning for a doctorate.

He received a Ph.D. in Computer Science from Texas A&M University in 2010. While a doctoral student, he held consecutive, full time positions as Manager and Lead Developer within the Texas A&M University System. For two years, he also taught courses about multimedia instruction for the College of Education & Human Development. His research interests include human-computer interaction, information retrieval, interactive design, machine learning, and software engineering. He plans to enter administration, academia, industry, or research in the future.

Mr. Davis may be reached through the Department of Computer Science and Engineering by (979) 845-5534 at TAMU 3112, College Station, TX 77845-3112. His email is p-davis@tamu.edu.