AN STRUCTURAL METRIC MODEL TO PREDICT THE COMPLEXITY OF WEB INTERFACES

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Abdulaziz Ahmad Attaallah

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Program:
Software Engineering

April 2017

Fargo, North Dakota

North Dakota State University
Graduate School

Title

An Structural Metric Model to Predict the Complexity of Web Interfaces

By

Abdulaziz Ahmad Attaallah

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHLOSIPHY

SUPERVISORY COMMITTEE:

Kenneth Magel
Chair

Kendall Nygard

Jun Kong

Benjamin Balas

Approved:

April 4, 2017
Date

Brian M. Slator
Department Chair

ABSTRACT

The complexity of web pages has been widely investigated. Many experimental studies used several metrics to measure certain aspects of the users, tasks or GUIs. In this research, we focusing on the visual structure of web pages and how different users look at them regarding complexity. Several important measures and design elements have rarely been addressed together to study the complex nature of the visual structure. Therefore, we promoted a metric model to clarify this issue by conducting several experiments on groups of participants and using several websites from different genres. The goal is to form a metric model that can assist developers to measure more precisely the complexity of web interfaces under development. From the first experiment, we could draw the guidelines of the major entities in the metric model, and the focus was on two most important aspects of the web interfaces, which are the structural factors and elements. Thus, four main factors and three main elements were more representatives to the concept of complexity. The four factors are size, density, grouping and alignment, and the three elements are text graphics and links. Based on them we developed a structural metric model that relates these factors and elements together, and the results of the metric model are compared to the web interface users' ratings by using statistical analysis to predict the overall complexity of web interfaces. The results of that study are very promising where they show our metric model is capable of predicting the complex nature of web interfaces with high confidence.

ACKNOWLEDGEMENTS

First, this work is dedicated to Allah (God) for all the favors he has given me, to finish this job. Second, I love to thank my Ph.D. advisors, Dr. Kenneth Magel, Dr. Kendall Nygard, Dr. Jun Kong and Dr. Benjamin Balas for supporting me during these past two years. Dr. Magel, he is the humble advisor who knows everything about everything and one of the smartest people I know. I hope I could be as lively, enthusiastic, and energetic as Professor Magel. Dr. Nygard has been supportive and has given me the freedom and time to make my mind in doing the statistical analysis. I am also very thankful to Dr. Kong and Dr. Benjamin for his scientific advice and knowledge and many insightful discussions and suggestions. Also, I cannot forget to thank Hassan Alsuhabi, a Ph.D. student in the Department of Statistics who was very interested in the research and did his best to help me out throughout the period of the study. Moreover, I would like to thank all the students who participated in the research's surveys.

I especially thank my wife, mom, children, and brother. My hard-working parents have sacrificed their lives for my wife and myself and provided unconditional love and care. I love them so much, and I would not have made it this far without them. My wife has been my best friend all my life, and I love her dearly and thank her for all her advice and support. I know I always have my family to count on when times are rough.

The best outcome from these past two years is having supportive people like my beautiful wife. She is the only person who can appreciate my quirkiness and sense of humor. There are no words to convey how much I love her. She has been a true and great supporter and has unconditionally loved me during my good and bad times. She has been non-judgmental of me and instrumental in instilling confidence. She has faith in me and my intellect even when I felt like digging a hole and crawling into one because I didn't have faith in myself. These past

several years have not been a smooth ride, both academically and personally. I sincerely thank her for sticking by my side, even when I was irritable and depressed. I feel what we both learned a lot about life and strengthened our commitment and determination to each other and to live life to the fullest.

I dedicate this project to my family, my wife and my children for their constant support and unconditional love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SLOC & LOC ....Lines of Code

LLOC ....Logical Lines of Code

GUI ....Graphical User Interface

AECL ....Atomic Energy of Canada Limited

IEEE..Institute of Electrical and Electronics Engineers

PCA..Principal Component Analysis

MS..Model Screens

IVO ....Identified Visual Objects

DB ....Dialogue Boxes

LA ....Layout Appropriateness

SC...Size Complexity

TSC ....Textual-Size Complexity

GSC..Graphical-size complexity

WLZC ....Wrapped-link-size complexity

DC ....Density Complexity

TDC..Textual-Density Complexity

GDC ....Graphical-Density Complexity

LDC..Link-Density Complexity

GC ....Grouping Complexity

GDC ....Textual-Grouping Complexity

GGC ....Graphical-Grouping Complexity

GDC ....Link-Grouping Complexity

AC ....Alignment Complexity

TAC..Textual-Alignment Complexity

GAC..Graphical-Alignment Complexity

LAC..Link-Alignment Complexity

UWIC..Unweighted Web Interface Complexity

WIC..Weighted Web Interface Complexity

IRB...Institutional Review Board

IADAS..International Academy of Digital Arts and Sciences

CP...Concurrent Probing

RP...Retrospective Probing

SEO..Search Engine Optimization

DE..Development Experiment

EE...Evaluation Experiment

LIST OF APPENDIX TABLES

LIST OF APPENDIX FIGURES

CHAPTER 1. INTRODUCTION

Background

Today, with a wide spread of the technology, and the use of computerized systems everywhere we can think of, we can quickly notice the invasion of screens everywhere. Many paper-based systems have been converted to computer based systems. Furthermore, to increase the accessibility of these systems, they tend more to exist in different forms that can be accessed via the internet such as phone applications and websites. Also, a vital feature that distinct this era, is the high competitions and the high demands of technological adaptations, therefore, the proper usability of Graphical User Interface (GUI) is not only a user satisfaction attribute, but it is a quality factor. A user may prefer to purchase a product from a web store over other types of stores due to the clear and straightforward process, and the convenience of finding almost anything out of reach. Consequently, we observe the grow and emerge of computer specializations that handle and takes care of many aspects that are related to GUIs and user interactions.

Numerous studies have been done to mitigate the negative impacts that can happen due to failures in designing proper and quality GUIs, and it is not only essential to the success of the software product but also has a very high value of benefits that software project stakeholders can gain out of it. For example, the ability to predict the success of specific tasks, determining the time completion of specific tasks, measuring the satisfactory level of users at early stages, and computing the performance to achieve the targeted goals [1]. These are very significant benefits that can cause either the failure or the success of software projects. Moreover, neglecting usability testing in some critical systems can lead to disasters, for example, using only visual emergency light indicators on systems that blind users are a potential group of users for it, is a tragic mistake that may lead to deaths in some situations.

Nevertheless, usability testing has a dark side. It requires various things to be set up to gain real output such as facilities, equipment and staff time. Also, the sample size is usually not enough most of the time, and it is hard to maintain the commitment of the participants, also, analyzing the data that they generate is a time-consuming and complicated task to do. Therefore, many solutions have been introduced to moderate the flaws that may occur by software testing, and software metrics are one of them, which are measures of some property of a piece of software or its specifications. The objective of software metrics is finding reproducible and quantifiable measurements, which may have several valued applications in the schedule and budget planning, cost estimation, quality assurance testing, software debugging, software performance optimization, and optimal personnel task assignments [2].

Problem Definition

It is known that testing codes are one of most rushed phases in software development due to many reasons, and that puts the usability testing in a worse situation because development teams usually tend to rank functionalities over the look and feel of the systems that are in preparation. Especially, if the deadlines are very tight, so delivery becomes one of the project's priorities. Some people are ignorant of the importance of GUI particularly in life critical systems in which flaws of GUIs can be disastrous. For instance, The Therac-25 was a radiation therapy machine formed by Atomic Energy of Canada Limited (AECL) in 1982. At least six accidents between 1985 and 1987 occurred, in which patients were given massive overdoses of radiation. Resulting in death or serious injury, and according to many software experts, the complexity of the GUI was a major contributing factor in that catastrophe [3]. The cost is not only human lives, but also sometimes it can affect their lives very unpleasantly from a different perspective. In a study provided by Adaptive Path, Bank of America directed an investigation into why they were dropping behind their competitors. Approximately 45% of their customers were giving up the online registration due to the hard user experience they had during that process [4]. Another devastating study, which is done by the Institute of Electrical and Electronics Engineers (IEEE), almost all projects that are under development, 5 to 15 % will be abandoned before or soon after delivery due to poor usability, and causing a loss of $150 billion [5].

Subsequently, the need for robust GUIs that can be tested before they become available to their end-users is a demanding need, especially, the web interfaces due to the vast spread of the web applications and systems. Consequently, studying and analyzing the structure of web interfaces will allow us to approach quality web pages, which result in successful software products and avoiding costly failures. Web pages have a variety of content that is identified as hypermedia such as videos, images, text, flashes, links and others. The perception of all these visual elements and their categorization varies from one person to another due to many factors. For example, a website oriented to kids' education, its audience holds several characteristics that differ from the audience of an online banking system. Therefore, not only the reflections to those systems diverge based on human factors, but also, the visual design elements have variable and standard features that influence the browsing experience such as the density of objects on a screen. Another problem faces us in assessing the complexity, is the diversity of methods in the literature to measure various parts of GUIs that causes the sense of complexity, we know each website is designed differently and has a distinctive interaction experience.

Several proposals use many quantitative measures to evaluate the complexity. Additionally, there is a healthy sign that the perceived quality of a GUIs has a positive influence on the ideas that users have about the systems' effectiveness and usability [6]. Many metrics models attempt to address the problem of structural complexity; however, the lack of a crucial feature, which is cohesiveness, make them not entirely applicable and acceptable in the web development world.

Objective

This research is about constructing, building and developing a coherent structural complexity metric model to understand complexity based on critical measures along with important design elements exist in the literature. The objective of this dissertation is to introduce this metric model by which automated tools can be built in the future to predict and present levels of web interface complexity. The nature of software development phases most of the time, especially, in the testing phase does not allow appropriate resources such as money, time, staff and participant to conduct usability testing sessions. Thus, using metrics to quantify the complexity during the development process of software products enhances the overall outcomes of the software projects.

The starting point of our approach is investigating the web users to elicit the foundations of the metric model along with the most popular and reported entities and factors in the literature, which have a correlation with GUI complexity. Therefore, addressing to what extent the structural measures and the elements of complexity existed confirms the views of actual users of different characteristics. Also, this study we take into consideration the variety of website genres, which is a factor that has been ignored in many studies in regards to structural complexity factors and objects within the scope of the visual evaluation. Usually, similar systems are compared together to measure the complexity, where in fact some of the structural factors and elements that may cause complexity are either absent or have less emphasis. Moreover, presenting structural design elements to the users from the perspective of structural complexity measure gives the study a unique dimension. For instance, the users evaluate the complex nature of hyperlinks regarding specified factors such as size, density, grouping, and alignment instead of evaluating with no guidelines.

Research Approach

We assume that developers will build a tool in which our metric model will be implemented. The metric model will be able to predict the complexity of web pages based on certain factors and entities that exist on their layout. The primary goal of this tool is to decrease the chances of conducting usability testing session at early stages of look and feel design. Hence, the research has the four broad phases: complexity metrics identification, complexity metric model formation, test case design, and comparison effectiveness.

Complexity metric identification

In this phase, we survey the reach complexity metric literature, and we study web users to establish the complexity metric model. Based on the statistical analysis of the data collected in this phase, a complexity metric model is formed.

Complexity metric model formation

This phase includes considering the literature to find metrics that have been used to address the complexity of GUIs from a structural point of view, and the use of these metrics have become frequent and continued over the years. Then, we select the most efficient and reliable metrics based on the results they produce. In addition, a survey of web users is conducted to investigate their understanding of web interface complexity. By combining the results from both surveys, a list of factors and elements is driven to form the general frame of our metric model. The hypotheses of this research also will be determined by this process, and then, the design for test cases will take place to exam the assumptions' correctness. An illustration for theses assumptions will be like this:

1. A set of factors and entities on web pages are better predictors of web interface complexity than other ones.
2. A set of elements' characteristics which are driven by factors provides a better prediction for web interface complexity.
3. Based on the statistical analysis of users' views on web interface complexity, the combination of structural metrics can work as a predictor of web interface complexity.

Test case design

In this phase, we select a set of web sites that consist of five from different genres to be used in a laboratory experiment. Tasks, setup, participant recruitment, and equipment are established and implemented in this phase.

Evaluation, calculation and data analysis

In the last phase, our calculations and measurements for the complexity of web interfaces of the selected websites are assessed against the users' evaluations. The evaluation will include advanced statistical tests to determine the level of correspondence between the data produced by the two methods of the assessment for the web interfaces such the T-test, Analysis of Variance (ANOVA), Linear Regression and Principal Component Analysis (PCA).

Results and Significance

The result of this dissertation research includes:

1. Designing and implementing structural complexity metrics for web interface by extracting web users' views of structural complexity and by our modifications on Model Screens (MS) technique, which is Identified Visual Objects (IVO). The amendments that we did on the MS is the innovation of our approach.

2. The set of structural complexity metrics is based on the distinction between the structural factors and the structural element, which in theory the users' evaluations match its general frame. Four structural factors and three structural elements were selected to build the metric model as shown in Figure 1. Each structural factor of the four measures the complexity of each structural element of the three. Consequently, by mixing and matching the four factors with three elements, the result is twelve structural metrics.
3. Forty-one textual, graphical, and links attributes, as shown in Figure 2, had been put to the test to select only twelve in each factor category, which were ranked by experiments' participants as shown in Figure 3, to calculate the overall complexity of a web interface.
4. Each attribute of structural elements has been calculated for ten websites' homages per four structural factors. Based on the statistical analysis, we selected twelve metrics, and they are strongly related and carried considerable weight to represent the overall structural complexity.
5. The conjunction of the real screen, screen models and our addition of identifying the structural elements on the screen model resulted in a more compact and comprehensive complexity metric that takes into consideration the most compelling influences of web interfaces. Our research is the first of its kind that measures the complexity of element attributes according to the complexity structural factors as one unit instead of focusing on either one of them separately. Therefore, we believe the outcomes of this approach is much more efficient than the other methods, in some of which the concentration is only on one side, either

factors or elements. By applying our metric on two empirical studies, the results were very promising in predicting the complexity of web interfaces.

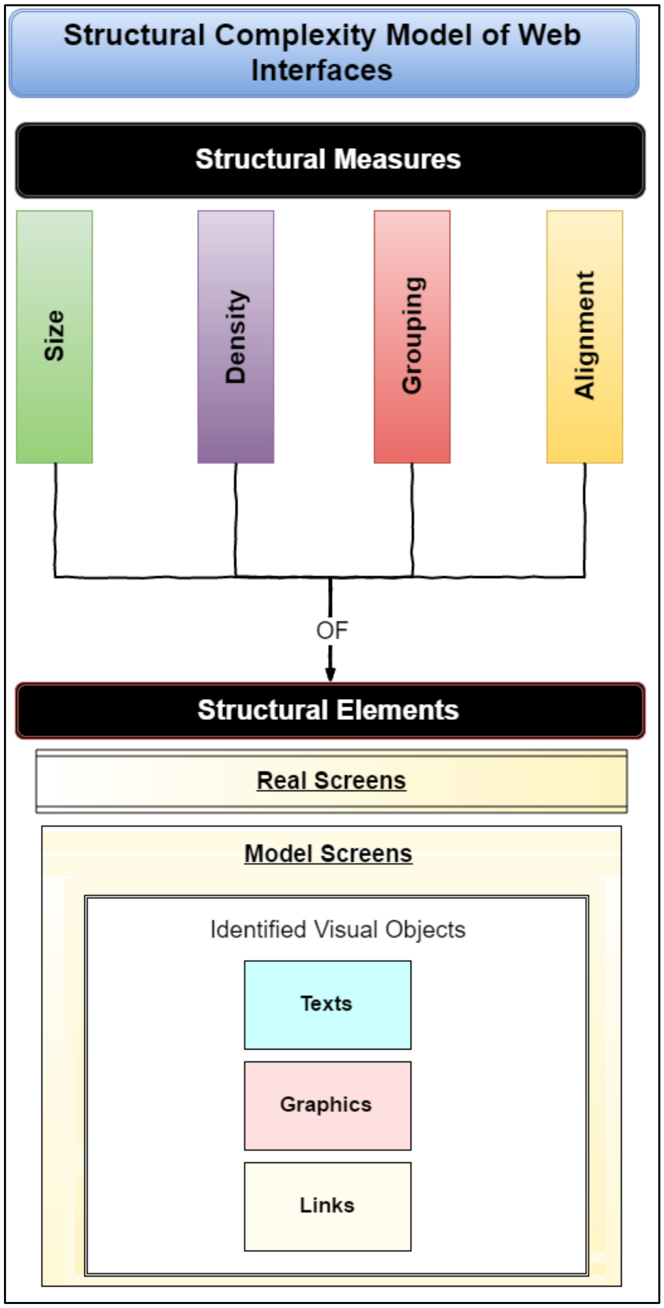


Figure 1. Preliminary structural complexity metric model for web interfaces

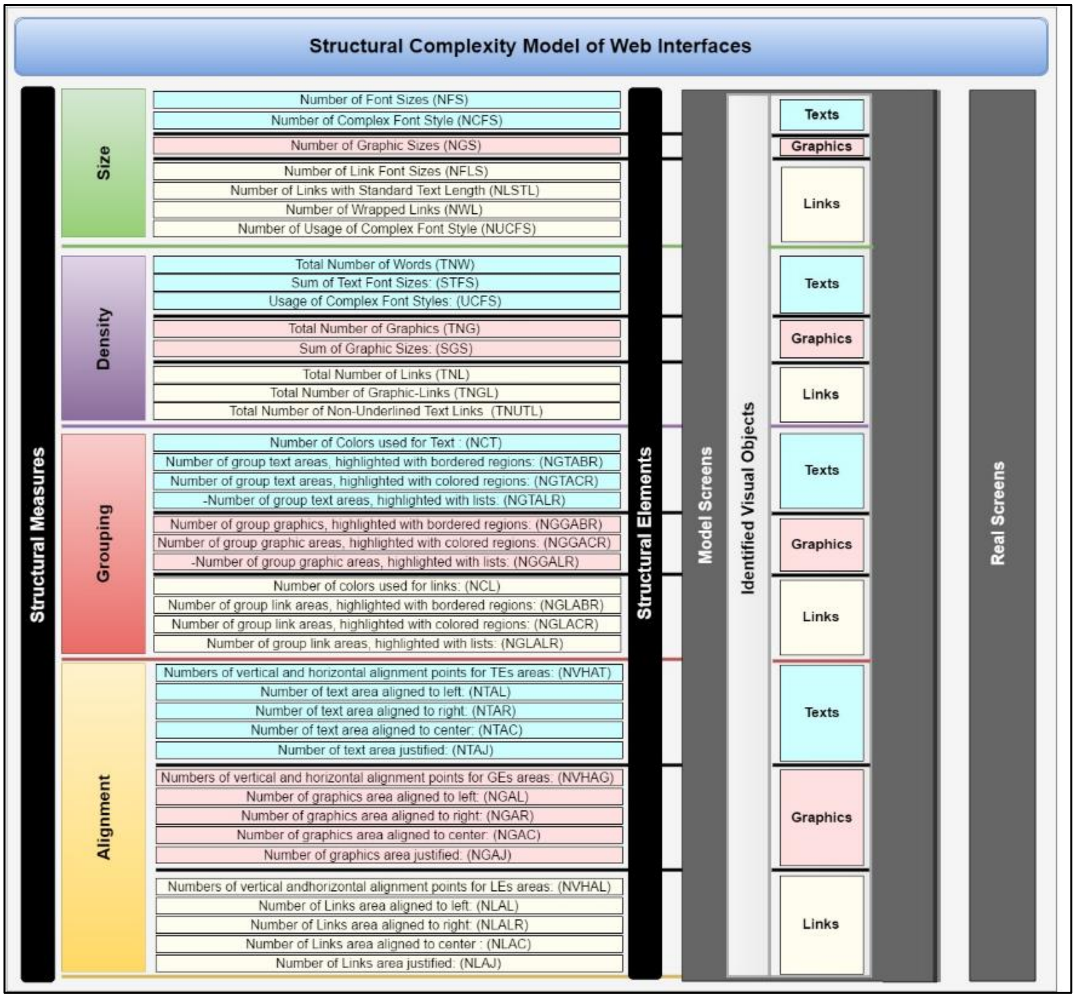


Figure 2. Detailed preliminary structural metric model with forty-one attributes of elements

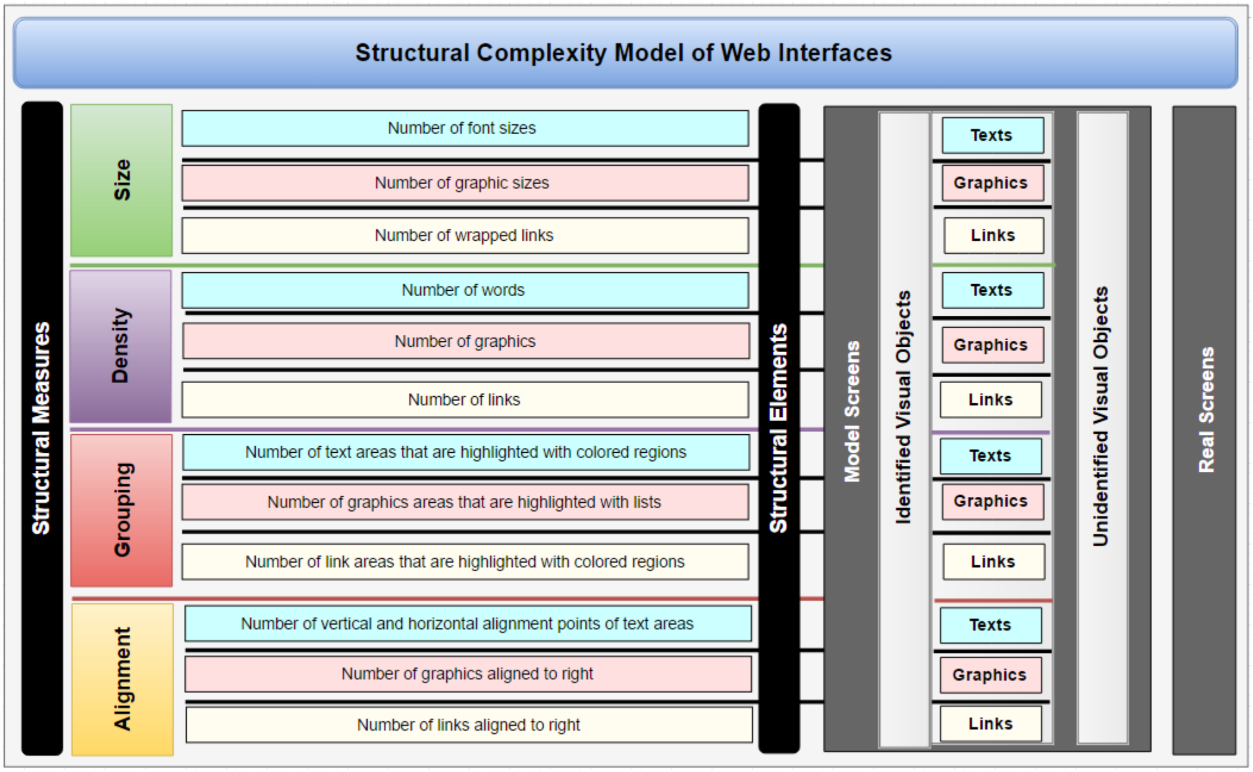


Figure 3. Structural metric model with the twelve highly ranked elements

Structure of the Dissertation

The rest of the dissertation is organized as follows.

Chapter 2 presents the related work. Chapter 3 displays the development of structural complexity metric model and the test plan design. Chapter 4 describe the empirical study and data analysis. Chapter 5 explains the calculations, the analysis of the results and the interpretation of the study’ outcomes. Chapter 6 summarizes the contribution of the dissertation research and the discussion of the future work.

CHAPTER 2. LITERATURE REVIEW

Many subjects are related to this complexity metrics. However, the focus in this chapter is only on complexity metrics that are in a relationship with GUIs, and other complexity metrics are not covered in this chapter.

Software Complexity Metrics

To start off, we need a precise definition of measurements that can enlighten the path to construct one. According to [2], A "measurement is the process by which numbers or symbols are assigned to attributes of entities in the real world in such a way as to describe them according to clearly defined unambiguous rules." Thus, a measurement capture some critical information about attributes exist on entities that we care about. For example, let's imagine that a person wants to choose a room in a house for sleeping, and an essential requirement or characteristic of this room that he/she cares about is the room temperature, which must be 25 Celsius. So, from this description, we can infer the terms used in the definition of the measurement. Consequently, if we apply the definition in many cases in our lives, the significance of measurements will immediately rise. We do measurements for many reasons, for example [7]:

- Predictions: many aspects of software are fatal for the success of the products such as performance, effectiveness, and reliability.
- Evaluation: how do we know if we are doing a magnificent job or a disaster? The ability to evaluate the work and its outcomes continuously, assess very much in the destination of the entire software projects.
- Prioritization: doing the right job, is not always acceptable because it might be it was not the most outstanding job on a task list. Also, it makes the question, what is next? Easier to answer.

GUI Structural Metrics

Lines of Code (LOC) (SLOC) is a very famous complexity metrics that is used to measure the size of a computer program by counting the number of lines in the scripts of source codes of programs. However, sources codes have very distinct nature and characteristics by which measuring the complexity of GUIs is not feasible. There are two key categories of SLOC measures: physical SLOC and logical (LLOC). Precise descriptions of these two measures differ, nonetheless, the most mutual meaning of physical SLOC is a sum of lines in the script of the program's source code without comment lines [8]. An example of it is shown in Figure 4.

```
for (i = 0; i < 100; i++) printf("hello"); /* How many lines of code is this? */
```

In this example we have:

- 1 Physical Line of Code (LOC)
- 2 Logical Lines of Code (LLOC) (for statement and printf statement)
- 1 comment line

Depending on the programmer and coding standards, the above "line of code" could be written on many separate lines:

```
/* Now how many lines of code is this? */
for (i = 0; i < 100; i++)
{
    printf("hello");
}
```

In this example we have:

- 5 Physical Lines of Code (LOC): is placing braces work to be estimated?
- 2 Logical Lines of Code (LLOC): what about all the work writing non-statement lines?
- 1 comment line: tools must account for all code and comments regardless of comment placement.

Figure 4. SLOC (LOC) and logical SLOC (LLOC) examples

Mainly, GUIs have been studied based on the usability concept because it covers several aspects of GUIs characteristics. There are several different ways of evaluating a GUI that include formal, heuristic, and manual testing. Other taxonomies of user evaluation methods comprise; predictive and experimental. Contrasting distinctive software, some of those assessment procedures may hang on exclusively on users and may never be computerized or considered mathematically [9]. For example, in a heuristic evaluation, usability specialists review site's interface and associate it with known usability values. The analysis results in a list of possible usability problems. As any assessment techniques, it has advantages and disadvantage as shown in Table 1, and it does not mean to abandon the usability testing.

Table 1. Heuristic evaluation's advantages and disadvantage

Advantages	Disadvantages
• It can deliver some rapid and cheap advice to engineers. • It can find reaction early in the design development. • Assigning the precise heuristic can aid propose the finest practical measures to designers. • It can be used together with other usability testing methods. • Usability testing can be conducted to additional examine possible problems.	• It needs information and knowledge to apply the heuristics successfully. • Skilled usability experts are sometimes hard to find and can be expensive. • Multiple experts must be used and total their results. • The assessment may classify more trivial problems and rarer main problems.

Calculation methods are mainly divided into two categories. In both types, methods include a mental walk-through and heuristic evaluations that depend on skilled engineers and experts to distinguish problems based on guidelines and human performance criteria. Other methods are more experimental and data-collection oriented, and they are achieved either in workshops or the work fields [10].

Typography and Information Theory

Bonseippe [11] pointed to the Information Theory and the Concept of Complexity that Shannon and Weaver developed a formula for to measure the complexity of typographically designed pages, and Figure 5 shows an example of an experiment conducted by Twyman [12] to check for content effects and to simplify reading of the data analysis. Many studies afterward adopted this method to measure the visual screen complexity of Windows applications. Recently, it has been taken to measure the visual complexity of web interfaces. This technique includes calculating the number of components on the screen and the number of horizontal and vertical alignment lines connecting these components. In relation to the information-theory formula, as these numbers rise, so does the level of visual complexity [13].

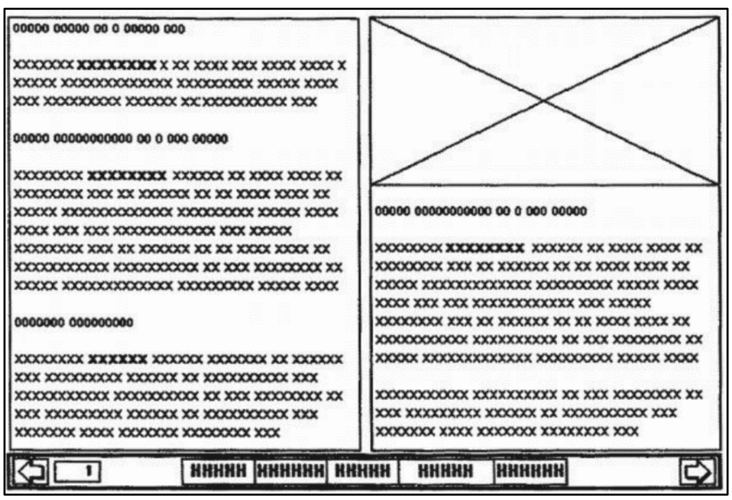


Figure 5. Measuring the complexity of typographical designed pages

Evaluation of Screen Formats Using Structural Factors

Tulis [14] built on top the information theory formula, to become more applicable in measuring the Dialogue Boxes (DB) complexity, in which he presented the model of complexity that consist of the total complexity score, and his approach based on four measurements of complexity (size, density, alignment and grouping). These four measures have been widely used in several different ways of measuring the complexity of screen competent, for example, it was utilized in conjunction with the probabilities of the information theory as Miyoshi and Murata [10,13,15] studied. Figure 6 shows an example of the equations used to calculate the complexity and density complexity, and Figure 8 shows a case of the hypothetical DBs employed in an experiment to measure the same four factors

One of two measures about size based on the information theory using on probability[8] was defined as the sum of the complexity of each element type such as a button, a radio-button, or an edit box, as follows:

$$CS = \sum_i CS_i$$

$$CS_i = -N_i \sum_s p_{is} \log_2 p_{is}$$

where CS_i is the measure of complexity with respect to the size of the ith type element on the screen, N_i is the number of the ith type elements and the p_{is} is the probability of selecting elements with the sth size in the i th type elements. The subscript s indicates the size order.

This measure of density complexity depends on the grouping factor because the measure is calculated using density within the group. We defined the independent measure of density along with other factors and calculated it using the proximity of each element on the screen as follows:

$$e_{ij} = 1 - l_{ij}$$

$$ID = \frac{\sum_{i<j} e_{ij}}{N(N-1)/2}$$

where e_{ij} and l_{ij} are the proximity and the distance between the ith and jth elements on the screen and the ID (Index of Density) is the average of the proximity of the other elements.

Figure 6. Complexity of size of each element type and complexity of density with other factors

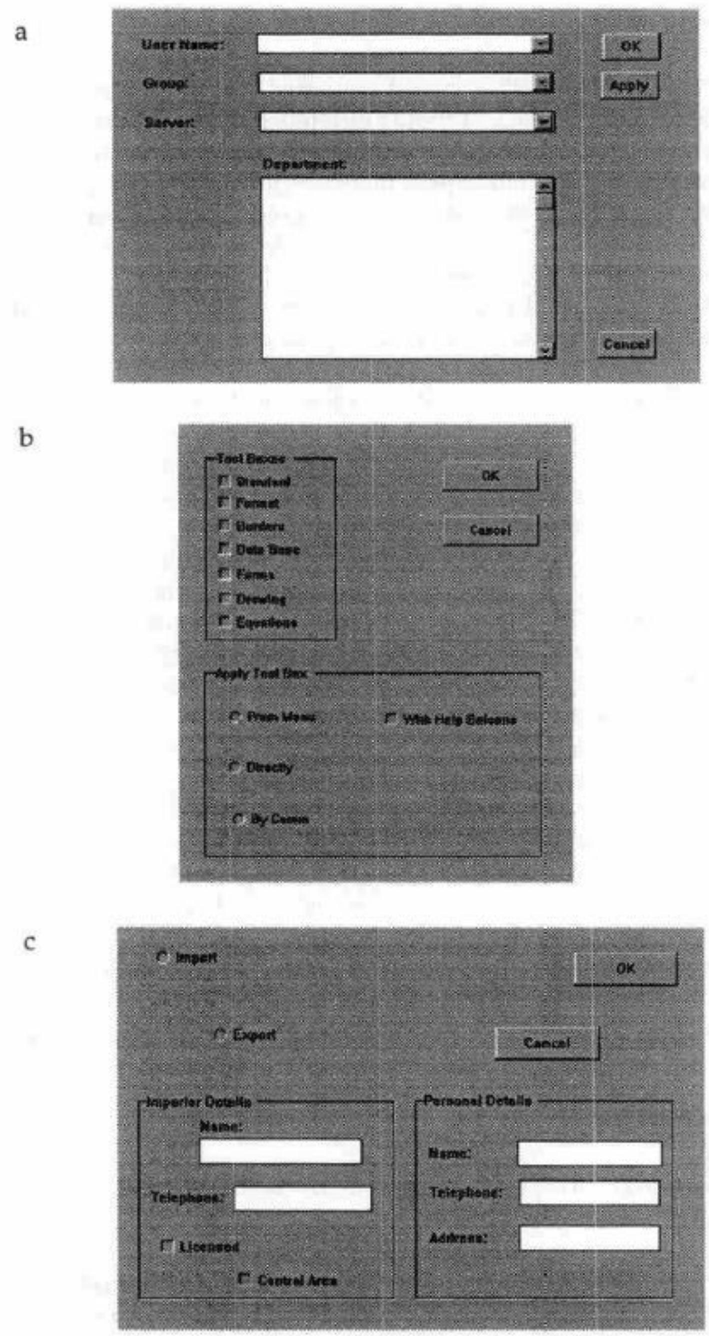


FIGURE 1 Hypothetical DB examples: (a) shows size-variation factor, (b) shows grouping and local-density factors, (c) shows alignment factor.

Figure 7. Hypothetical DBs used in an experiment to measure the same four factors

Source of Figure 7: [10] A. Parush, R. Nadir, and A. Shtub, "Evaluating the Layout of Graphical User Interface Screens: Validation of a Numerical Computerized Model," Int. J. Hum.-Comput. Interact.

Also, Sears [16] industrialized a layout metric named Layout Appropriateness (LA), in which each task requires a series of activities and the metric tries to compute the costs of each series of activities. Xing [17] established metrics which apply three features related to complexity: numeric size, a variety of elements, and the relation among components. Moreover, Parush developed a numerical model [10] consists of four screen factors: element size, grouping, alignment, and local density to assess the GUIs.

Automation of Screen Evaluations

The before mentioned studies of the metric models did not have automated tools to do the calculation; thus, the following studies took chances to develop automated tools to do so. Sears [18] presented a metric-based tool, AIDE, which to some extent it computerizes the assessment process of user interface layouts, and Figure 8 displays the tool's main interface.

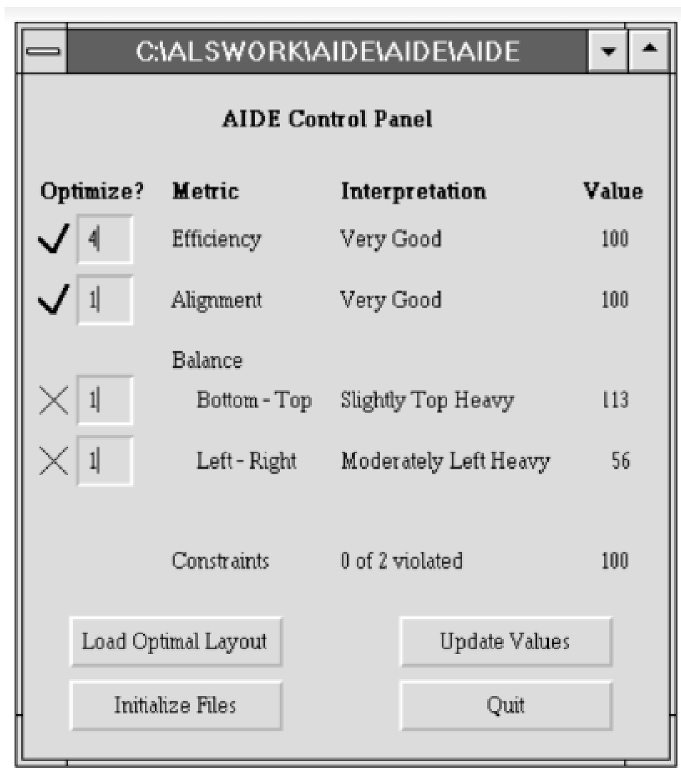


Figure 8. The AIDE Control Panel. Efficiency and Alignment have weights of 4 and 1 respectively

Source of Figure 8: [18] A. Sears, AIDE: A tool to assist in the design and evaluation of user interfaces. 2001.

Fu et al. [19] developed a mathematical tool to assess the screen complexity of web pages utilizing four measurements: size complexity, local density, grouping, and alignment. In addition, they applied the screen molding mechanize in their experiments in which only the structure outlines of web pages are drawn on the screen to measure the different aspects of the components without any contents inside them. Figure 10 displays an example of model screens and real screens used in Fu et al. [19] research. Alemerien and Magel [20] developed a metric-based tool, GUIEvaluator, which evaluate the complexity of the user interface based on five modified structural measures of complexity: alignment, grouping, size, density, and balance. Apparently, all these measurements were taken for visual elements. Thus, there are many studies conducted to discover the unique features and attributes of these large elements, such as text, videos, flashes, pictures, menus, lists, buttons, and data entry boxes.

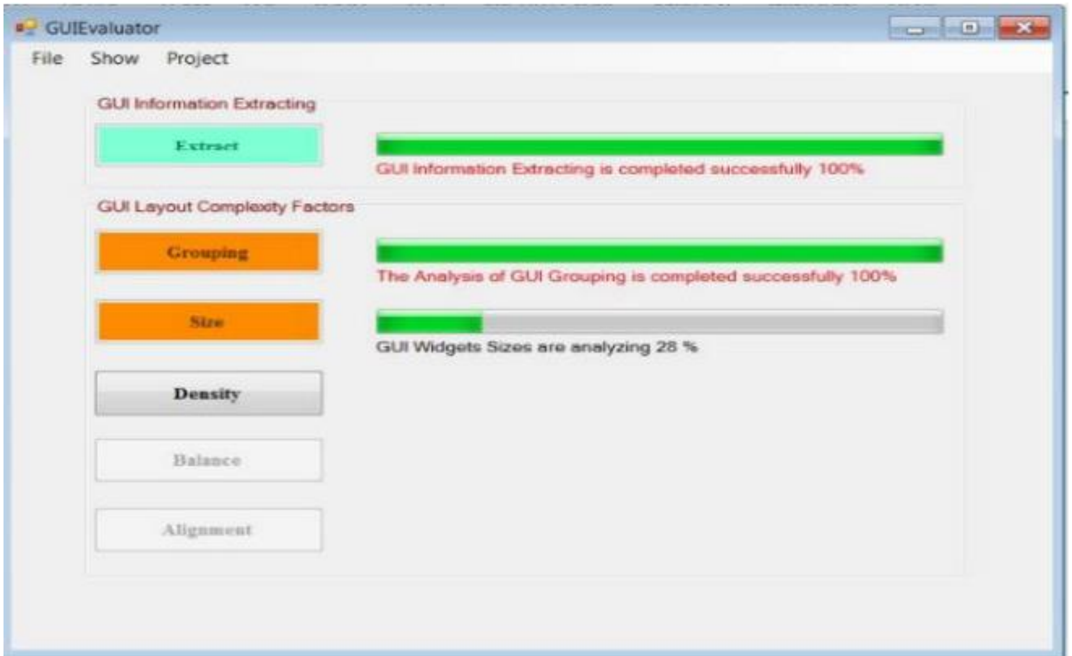


Figure 9. The extraction and analysis window of the GUIEvaluator

Source of Figure 9: [20] K. Alemerien and K. Magel, "GUIEvaluator: A Metric-tool for Evaluating the Complexity of Graphical User Interfaces.," in SEKE, 2014, pp. 13–18

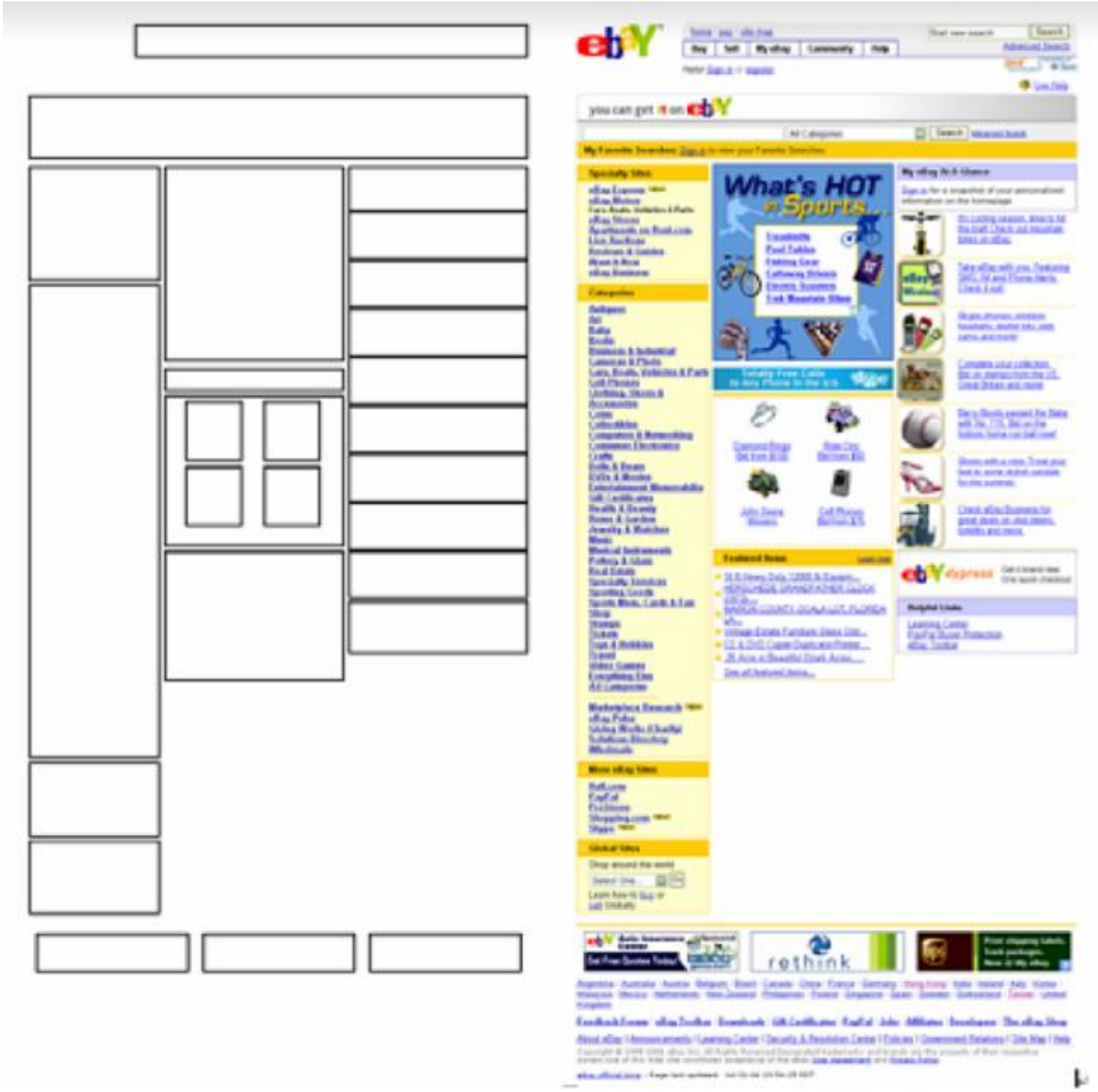


Figure 10. Model and Real Screen of Ebay USA

Source of Figure 10: [19] F. Fu, S.-Y. Chiu, and C. H. Su, “Measuring the Screen Complexity of Web Pages,” in Human Interface and the Management of Information. Interacting in Information Environments, M. J. Smith and G. Salvendy, Eds. Sp

Building Blocks of Web Interfaces

Nonetheless, Ivory [21] suggested that the building blocks of web interfaces are texts, links, and graphics. Furthermore, other studies demonstrate the significance of these three elements such as Nielsen [22] mentioned in his book that text recommended being kept short; by applying 50% less text in print publications. Spool et al. [23] state that a huge number of links obstructs navigation. Schalles [24] says graphics complexity impact beginners much more than specialists, as they requisite to deliberately preserve a sense of signs in working memory.

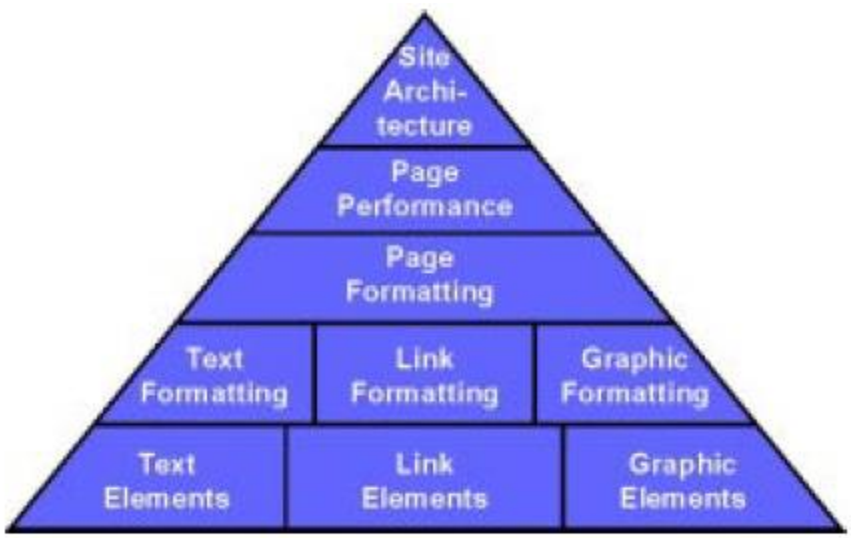


Figure 11. Aspects associated with Web interface structure

Source of Figure 11: [21] M. Y. Ivory, "An Empirical Foundation for Automated Web Interface Evaluation," University of California, Berkeley, 2001.

In our approach to this research, we are attempting to develop a metric model based on the structural measures and elements to extend the work mentioned in the previous work. As the first step to our approach, we decided to understand more how users of websites look at them regarding complexity. Even though this may result in subjective output, it is very significant to have a solid base for the metric model that we are developing. In the previous work, the emphasis on the structural complexity issue was either from an aesthetic point of view or with attention to major structural complexity factors of web pages. Secondly, based on the result of the previous work, which shows a strong relationship between complexity structural measure and the visual objects on the web pages, simplifying and identifying key visual objects is essential to be addressed in a metric model. The distinction of factors, characteristics, and entities we assume it will give a better reading to the overall complexity of the GUIs. In the research, we want to examine to what extent users of websites; realize the connection between the concept of structural complexity and the actual attributes that affecting the concept itself. In addition, we perceive the previous work somehow did not cover most of the aspect that we put together as one unit. Subsequently, this will contribute to design decisions in the software development world.

CHAPTER 3. CONSTRUCTION OF STRUCTURAL COMPLEXITY METRIC MODEL

To simplify the big picture of the software metrics world, we need to explain and clarify some important concepts and phenomena.[25] Linda Westfall is the President of the Westfall Team, which provides Software Metrics and Software Quality Engineering training and consulting services, and she has put together useful guidelines for software engineers who want to build their own software metrics. The instructions consist of twelve steps, which are covered briefly along with some terminologies, and we are going to apply them in our case to build our complexity metric.

Software Metrics

According to Goodman [26], software metrics are "The continuous application of measurement-based techniques to the software development process and its products to supply meaningful and timely management information, together with the use of those techniques to improve that process and its products." This definition can be illustrated in Figure 12.

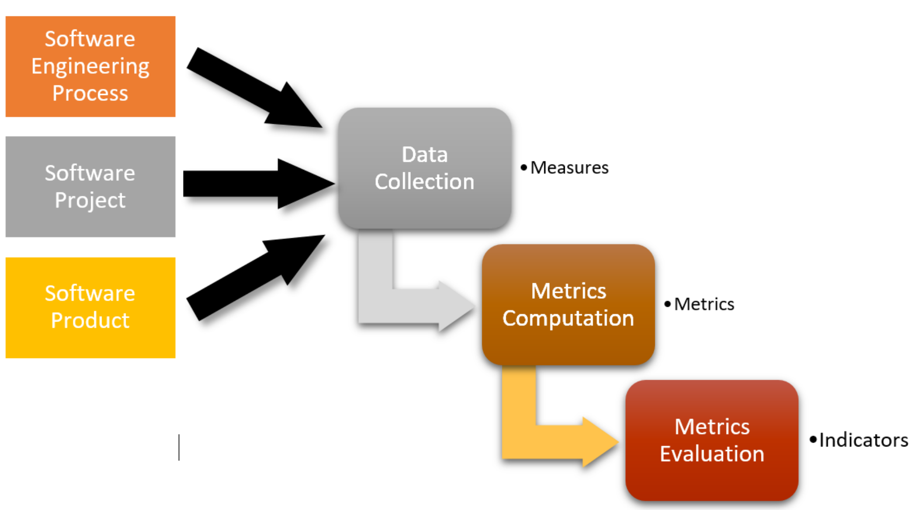


Figure 12. An illustration of the major components of software metrics and how they interact

To measure, we should primary uncover the entities, for example, we might choose a car as our entity. When we chose an entity, we must select the characteristic of that entity that we need to describe. The car's speed or the pressure in its tires would be two attributes of a car. Finally, we must possess a distinct and known planning scheme. It is pointless to say that the car's speed is 65 or its tire pressure is75 except we recognize that we are speaking about miles per hour and pounds per square inch, respectively. Software entities are products of the software process. These embrace all the artifacts, deliverables, and documents that are produced. Examples of software output entities contain requirements documentation, design specifications, code (source, object & executable), test documentation (plans, scripts, specifications, cases, reports), project plans, status reports, budgets, problem reports, and software metrics. Each of these software entities has many properties or features that we could want to quantify. We ought to inspect a computer's price, performance, or usability.

Software metric construction

The twelve steps that Linda [25] has introduced to build a software metric are utilized and summarized into five steps that we think they help in drawing the big picture of software metrics in general and applied them on our approach.

Step one: metrics customers

Description: The customer of the metric is the person (or people) who will be making decisions or taking action based upon the metric.

Application:

- Software Engineers/Programmers: The people that essentially do the software development. Interested in making informed decisions about their work and work products. These people are accountable for gathering the substantial amount of the facts vital to the metrics program.

- Test Managers/Testers: The people responsible for executing the verification and validation activities. Interested in discovering as many new flaws as possible in the time allocated to testing and in gaining assurance that the software works as quantified. These people are also responsible for gathering the substantial amount of the vital data.
- Specialists: Individuals executing focused functions (e.g., Marketing, Software Quality Assurance, Process Engineering, Customer Technical Assistance). Interested in measurable information upon which they can base their conclusions, outcomes, and approvals.

Step two: target goals

Description: There are two types of goals, the high-level goals such as strategic goals like being the low-cost provider and keeping a high level of client satisfaction, and there are the low-level goals such reducing the number of error messages on login process. Most importantly, finding measurable and questionable goals.

Application:

- Usability Testers: Want quantifiable indications about the level of complexity of the web interfaces at early stages of software development.
- Utility Developers: Want robust mathematical model that can predict the complexity level of web interfaces to build tools with which other developers can use.

Step three: define the questions

Description: Questions needed to be answered to guarantee that each goal is being found.

Application: The following question are only examples; a complete list of questions can be found in the dissertation's surveys appendix.

- What reasons or factors that make web pages complex (confusing or less understandable)?
- What objects or elements that make web pages complex (confusing or less understandable)?
- Rate the factors based on the negative impact on a web page?
  - Balance of objects
  - Alignment of objects
  - Fonts of objects
  - Symmetry of objects
  - Spacing of objects
  - Unity of objects
  - Colors of objects
  - Sequence of objects
  - Size of objects
  - Regularity of objects
  - Density of objects
  - Grouping of objects

Step four: metric selection

Description: metrics that deliver the information desired to answer these questions. Each nominated metric has an explicit purpose of solving one or more of the questions that need to be answered to determine if the goals are met. Metrics objective template can be used to exemplify the goals and the selection of metrics based on them, as shown in Figure 13.

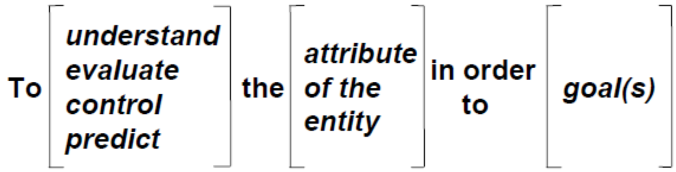


Figure 13. Metrics Objective Template

Source of Figure 13: [25] L. Westfall, "12 Steps to Useful Software Metrics." 2005.

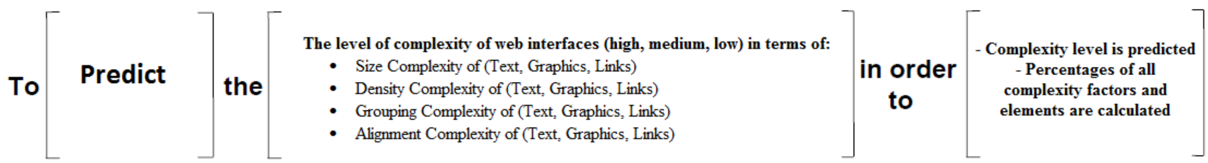


Figure 14. Application of Metrics Objective Template

Step five: standardize definitions

Description: To standard meanings for the entities and their measured characteristics. When people use terms like a defect, problem report, size, and even project, other people will interpret these words in their own background with meanings that can vary from our proposed description.

Application: Unfortunately, there is little standardization in the business of the definitions for most software attributes. However, definitions from the IEEE Glossary of Software Engineering Terminology [IEEE-610] or those found in software engineering and metrics literature.

Step six: measurement function selection

Description: Some metrics, called metric primitives, are measured directly, and their measurement function contains only one variable. Examples of primitives measures that embrace the number of lines of code are reviewed during an inspection or the hours spent making an inspection meeting. There are two approaches for choosing a model to use as measurement function: use an existing model or make a new one. In many cases, there is no need to "re-invent the wheel." Thus, we are have selected several measurement functions and made our own amendments and modifications meet our model goals.

Application: In our approach, we have twelve functions, which are the result of multiplying the four factors and the three elements, and because each element has various attributes, only one attribute is measured by four factors. All factors, elements, attributes and their functions will we explained extensively in the upcoming sections of the dissertation, and here is an example.

$$\text{Textual-size complexity (TZC)}= \frac{\sum_{i=1}^{typef}(nf_{size}-1)}{tntz} \epsilon\ [0,1]$$

Where:

i: counter of types

typef: number of font types

nf_{size} : the number of font sizes

$tntz$: the total number of text sizes.

Structural Metric Model Components and Mechanisms

Web interface structure

A web interface is a combination of diverse elements (text, links, and graphics), arranging of these elements, and other features that disturb the general interface value. Web interface design involves a complex set of events for addressing these various aspects. For example, Ivory [21] surveyed 157 measures of web interfaces by which many aspects of web interfaces quality can be addressed such as consistency. Nevertheless, the research settled on the three elements mentioned above because they match the results of interviews with professional designers, as well as, its compliance with Venn diagram which regulates the information construction with classifying and grouping content objects and developing class labels to replicate the information structure. It is tough to inspect all visual elements of web interfaces without having some web interface quality factors in mind because the factors and the visual elements are abundant. For example, most of the websites today have rich controls and components that most average users can quickly name such as search boxes, videos, animations, pictures, pop-ups, ads, etc. In addition, the factors that have an influence on the visual appearance of the web interfaces can be easily identified such the balance, the unity, the grouping, the regularity, the density, and the alignment, etc. Therefore, we decided to specify the scope of the visual elements and factors into categories, to simplify the process of evaluation and measurement.

Web interface complexity

Our investigation of web interface structure was taken a step further by focusing on one aspect of it, which is the complexity of visual elements and their factors. Moreover, the investigation took into consideration the variety of classes of the users. Thus, many categories of web users were surveyed to get a better understanding of the web interface complexity. In this research, we only inspect the complexity of web interfaces from a visual point of view. Consequently, navigation, information structure, and task performance are not involved, however, we only study how the objects based on people's relative perception. Various studies have been done to examine the navigation complexity, for example, a group of researchers [25] presented in their survey, twelve metrics that concentrated on navigability factors in most of which the focus was on the optimization of navigability.

Structural complexity metric model

The structural metric model consists of three primary parts, and the reason for partitioning the model into these sections is that we want to perform an in-depth studying for each section. Therefore, we are not only selecting attributes of elements since they exist on them such as the number of words in a text, but we are categorizing the attributes according to the structural factors. For example, the font style of the text element is classified as a density attribute.

The structural measures

It has the four core measures mentioned in the previous work: size, density, alignment, and grouping. These measures are used to calculate aspects of web interface object as visual elements. Each measure has a different interpretation depending on the type of the element which is being measured. Thus, for instance, the size of text has a different reading in the context of the graphics because the attributes under measurement are diverse.

Measure of size complexity

According to Fu et al. [19] Size complexity in relation to web interfaces employs the classification of elements into clusters according to real physical size and deviation in those sizes. The size measure can have an enormous impact on users' views of visual objects as shown in Figure 15.

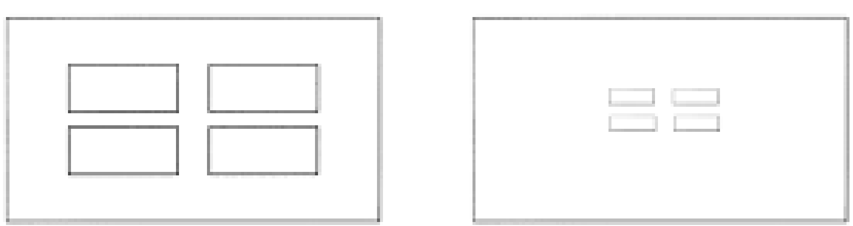


Figure 15. An example of size measure on screen visual objects

Measure of density complexity

According to Fu et al. [19], Density complexity is the degree to which the screen is occupied with objects, and it is accomplished by limiting screen density levels to the ideal percentage of 50% for graphic screens. The density measure can quickly change the desire of exploring as shown in Figure 16.

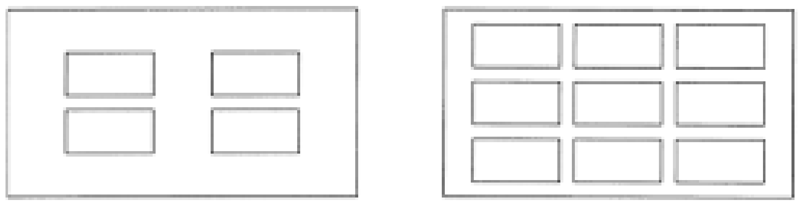


Figure 16. An example of density measure on screen visual objects

Measure of alignment complexity

According to Fu et al. [19], Alignment complexity is computing the level of alignment of a graphic screen that includes counting the number of different rows and columns on the screen that is employed as starting locations of objects. The alignment measure can give a bad time for readers of web pages' contents as shown in Figure 17.

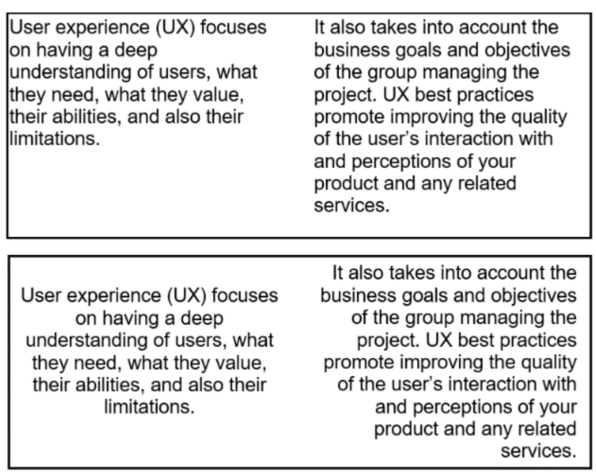


Figure 17. An example of alignment measure on screen visual objects

Measure of grouping complexity

According to Fu et al. [19] Grouping complexity reflects consistency: the amount of all elements matching together and look visually as one section and quantity. Elements with identical functions or information are fenced and enclosed by a frame or border by line, background color or space. The grouping measure can enhance the understandability of the visual contents as shown in Figure 18.

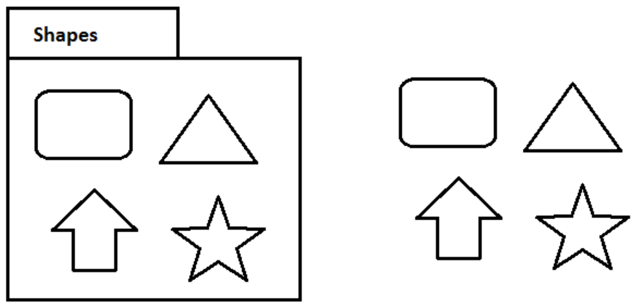


Figure 18. An example of grouping measure on screen visual objects

The structural elements

The essential elements mentioned in the previous work: text, links, and graphics. These elements are a source of most of the other elements that may exist on the screen of web pages. There are many things correlated with these elements' nature and/or properties that can be measured, and they play a significant role in the scene of complexity and usability of web interfaces.

The text element

The majority of the web content is reading material since the invention of the internet. Today, many companies tend to provide services to digitize old books such as Google and Amazon. The meaning of the text in this context is that any types of text or script which do not have the characteristics and features of links such as clickability, navigability and color changing. Ivory [21] states some of the aspects that can be measured or manipulated about texts like Word Count, Body Text Percentage, Emphasized Body Text Percentage, Text Positioning Count, Text Cluster Count.

<u>The link element</u>

Even though links share some characteristics with the text elements, they have unique features that make them different. Links are one of the main pillars of browseable content on the web. The meaning of links in this context is all visual objects that have the following characteristics and feature: clickability, navigability and color changing. Also, Ivory [21] states some of the aspects that can be measured or manipulated about links as well as Page Link Number, Internal Link Count and Redundant count.

<u>The graphic element</u>

They are all visual images or designs on web pages that may or may not deliver a factual information, in some cases, they are utilized for aesthetic purposes or organizational purposes. Additionally, Ivory [12] identified some unique attributes about graphics and aspects that can be manipulated like Number Animated Images, Graphic Ad Number Count, and Graphics Size.

Model screens

As reported by Ngo et al. [28–30] very high correlations were discovered among perceived and calculated aesthetics of the interface. The use of model screen allowed them to control properties of contents, and to simplify the explanation of the data analyses. Moreover, in agreement with Grabinger [31] findings, the features recognized by model screens are assessed when viewers judge the readability of choices of real screens from actual programs. However, part of the findings was that no evidence to specify whether people’s observations of interface aesthetics would vary if the real screens were used rather than the screen models. Model screens are a structural representation of the real screens, in other words, they are a structure with no content as shown in Figure 19.

Considering the previous studies, we approached the model screen with some amendments. Mainly, we decided to make a hybrid model screen in which a screen has the content of the real screens, but covered and highlighted with identified and labeled rectangles. The identification of these boxes depends on the structural element under the test and the observation of users. Consequently, for instance, if a user is making a judgment on text elements on a web page, then only text elements are highlighted by labeled rectangles. Figure 20 displays an example of our modified model screen. Additionally, we apply the same procedural on the other structural elements. Subsequently, the goal is to increase the accuracy of the users' judgments by emphasizing on certain aspects, features, factors and elements.

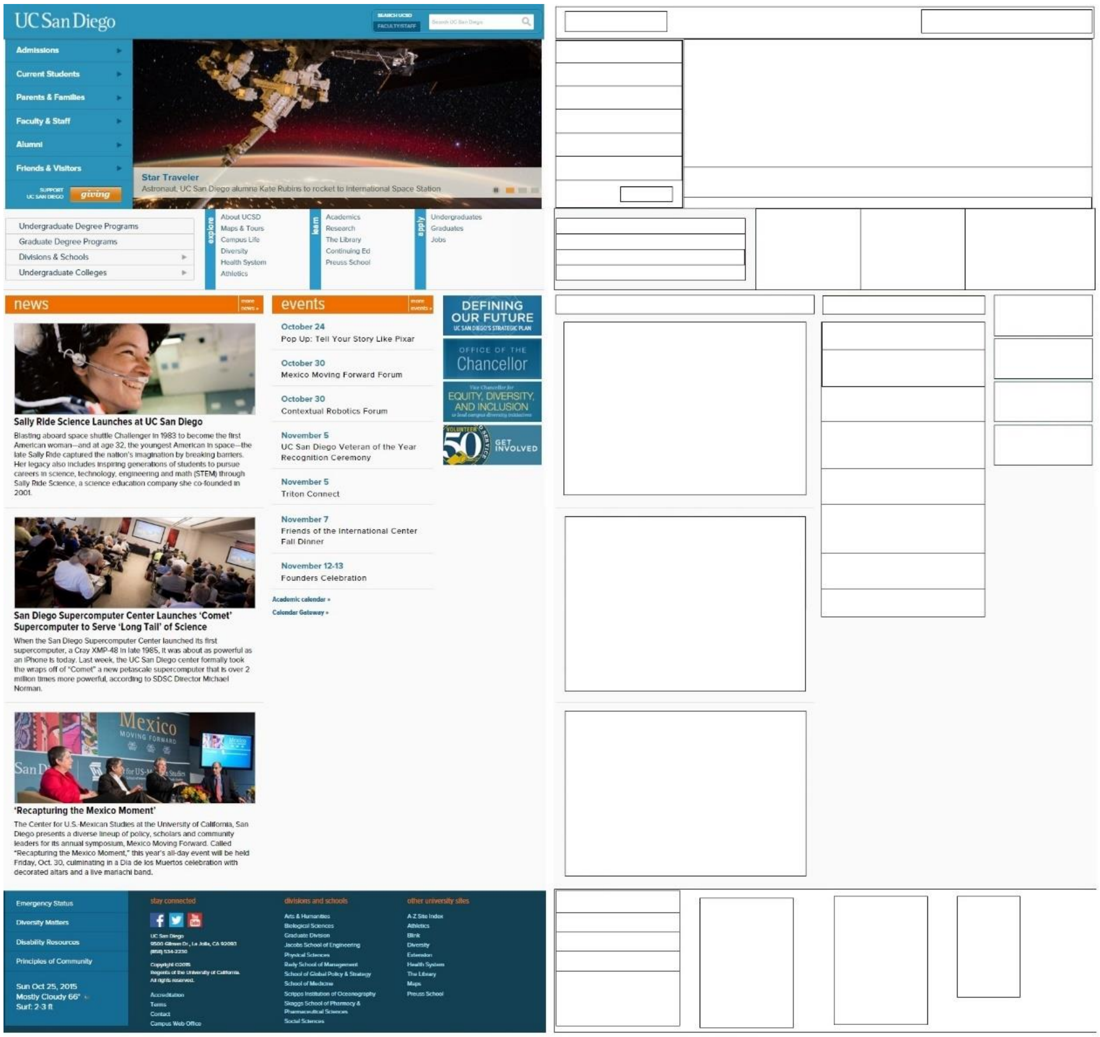


Figure 19. A demonstration of model screen on the right vs real screen on the left

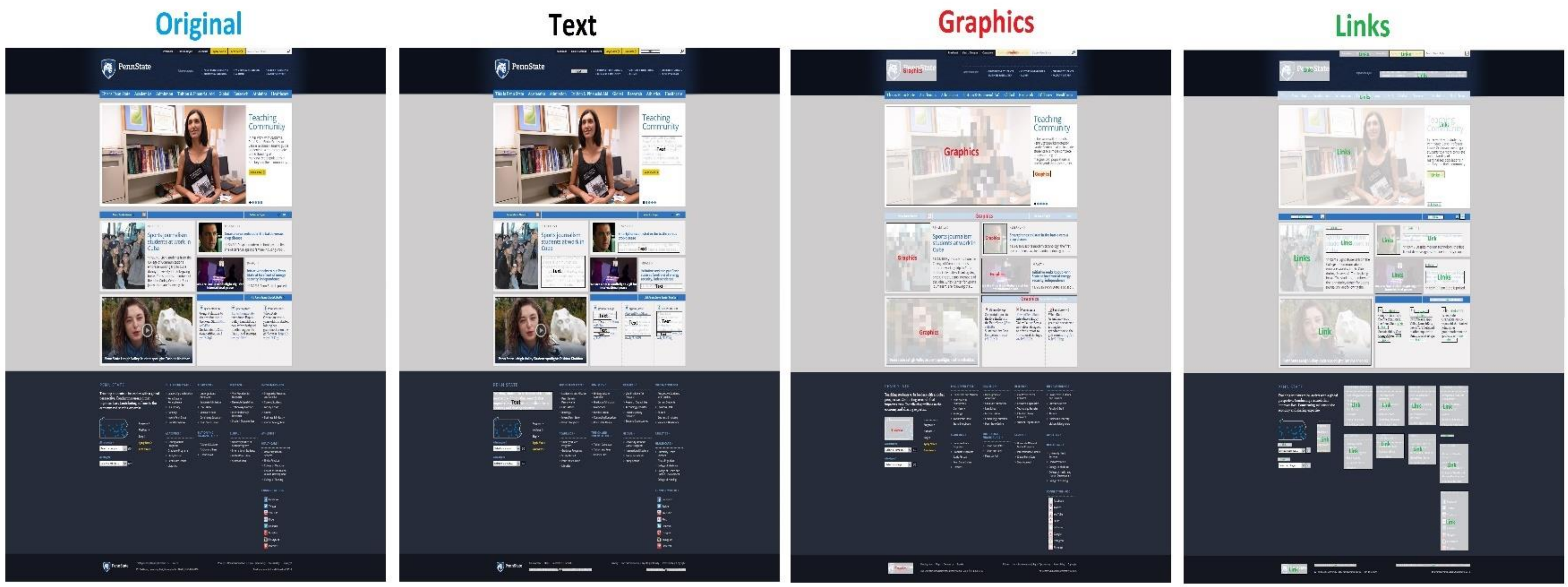


Figure 20. An example of the modified model screen

Measurement attribute selections

The massive number of attributes of the structural elements have been excessively studied in the literature. For example, In Ivory [21] research 157 quantifiable measures functioned to evaluate 62 structural features of web interfaces. Examples of the measures are: Word Count, Body Text Percentage, Emphasized Body Text Percentage, Text Positioning Count, Text Cluster Count, Link Count, Page Size, Graphic Percentage, Graphics Count, Color Count, Font Count, and Reading Complexity.

Our approach in selecting the most appropriate the attributes of the three chosen elements, which are Texts, Graphics, and Links, was based on their fitness under the four structural factors, which are Size, Density, Alignment, and Grouping. Subsequently, a set of measures categorized and selected as shown in Table 2. There were many reasons to disqualify groups of measures, sometimes because of its nature such as the Number of Animated Graphics, which cannot be represented on model screens to be judged by viewers.

Table 2. A list table of all measures, elements, and attributes used in the first experiment

#	Factors	Elements	Attribute Measures
1	Size	Text	1. Font Sizes 2. Usage of complex font style predominately sans serif, serif, or undetermined
		Links	1. Link Text Length: Use 2-4 words in text links 2. Wrapped Links: links traversing multiple lines 3. Font Size 4. Usage of font style predominately sans serif, serif, or undetermined
		Graphics	1. Number of graphic sizes
2	Density	Text	1. Word Count 2. Font Sizes 3. Usage of font style predominately sans serif, serif, or undetermined

Table 2. A list table of all measures, elements, and attributes used in the first experiment (continued)

#	Factors	Elements	Attribute Measures
		Links	1. Total number of Links 2. Number of graphic links 3. Non-Underlined Text Links count
		Graphics	1. Graphics Total Number 2. Total Graphic Sizes
3	Grouping	Text	1. Number of colors used for text 2. Number of text areas that are highlighted with bordered regions 3. Number of text areas that are highlighted with colored regions 4. Number of text areas that are highlighted with lists
		Links	1. Number of colors used for links 2. Number of link areas that are highlighted with bordered regions 3. Number of link areas that are highlighted with colored regions 4. Number of link areas that are highlighted with lists
		Graphics	1. Number of graphics areas that are highlighted with bordered regions 2. Number of graphics areas that are highlighted with colored regions 3. Number of graphics areas that are highlighted with lists
4	Alignment	Text	1. Number of vertical and horizontal alignment points 2. Number of text aligned to left 3. Number of text aligned to right 4. Number of text aligned to center 5. Number of text justified
		Links	1. Number of vertical and horizontal alignment points 2. Number of links aligned to left 3. Number of links aligned to right 4. Number of links aligned to center 5. Number of text justified
		Graphics	1. Number of vertical and horizontal alignment points 2. Number of graphics aligned to left 3. Number of graphics aligned to right 4. Number of graphics aligned to center 5. Number of graphics justified (stretched)

Measurement attribute reduction

One of the goals of this research is to construct a metric model as simple and illustrative as it could be for the web interface complexity. Therefore, after the profound and detailed analysis of the data that produced the first metric model and the viewers' judgments, we decided to take it a step further by concentrating on only the top-ranked measurement in each category. The outcome of that progression has twelve measures to compute the overall complexity of web interfaces. Table 3 shows the most qualified measures over the other 41 measures.

Table 3. A list table of the nominated twelve elements' attributes

#	Factors	Elements	Attribute Measures
1	Size	Textual-size	Number of font sizes
		Link-size	Number of wrapped links
		Graphical-size	Number of graphic sizes
2	Density	Textual-density	Number of words
		Link-density	Number of links
		Graphical-density	Number of graphics
3	Grouping	Textual-grouping	Number of text areas that are highlighted with colored regions
		Link-grouping	Number of link areas that are highlighted with colored regions
		Graphical-grouping	Number of graphics areas that are highlighted with lists
4	Alignment	Textual-alignment	Number of vertical and horizontal alignment points of text areas
		Link-alignment	Number of links aligned to right
		Graphical-alignment	Number of graphics aligned to right

Measurement attribute functions

There is a reasonable amount of formulas to measure different structural factors in the literature. For example, Ngo et al. [30] applied the structural factors on the visual objects of the web interfaces as shown in Figure 21. Since we measure both the three structural elements enclosed within rectangular frames, we are borrowing ideas for measuring the frames and the highlighted areas on the screen models.

$$DM = 1 - 2\left|0.5 - \frac{\sum_i^n a_i}{a_{\text{frame}}}\right| \in [0, 1]$$

where a_i and a_{frame} are the areas of object i and the frame; and n is the number of objects on the frame.

Figure 21. Measuring the density of visual objects on screens

Source of Figure 21: [30] D. C. L. Ngo, L. S. Teo, and J. G. Byrne, "Modelling interface aesthetics," Inf. Sci., vol. 152, pp. 25–46, Jun. 2003.

Likewise, in Ivory [21] research the attribute measures were used and called formatting measures, and since we are employing some of these measures, the primitive metrics which consist of one number will be borrowed as well such as the number of words. However, these measures will be calculated in the formulas based on their existence on the highlighted areas. Accordingly, each complexity factor is measured by computing the attributes of the three structural element, which means each complexity factor is represented numerically. The formulas of the twelve measures are described as follows:

Size complexity (sc)

<u>Textual-size complexity (tzc)</u>

$$\frac{\sum_{i=1}^{typef}(nf_{size}-1)}{tntz} \in [0,1]$$

Where i: counter of types, typef: number of font types, nf_{size} : the number of font sizes and $tntz$: the total number of text sizes.

<u>Graphical-size complexity (gzc)</u>

$$\frac{\sum_{i=1}^{typeg}(ng_{size}-1)}{tngz} \in [0,1]$$

Where i: counter of types, typeg: number of graphic types, ng_{size} : the number of graphic sizes and $tngz$: the total number of graphic sizes.

Wrapped-link-size complexity (wlzc)

$$\frac{\sum_{i=1}^{whl} nwhl_{size}-1}{tnhlz} \in [0,1]$$

Where i: counter of types, whl: number of types of wrapped links, $nwhl_{size}$: the number of wrapped-link sizes and $tnhlz$: the total number of links.

Density complexity (dc)

Textual-density complexity (tdc)

$$1 - 2 \text{ x } \left|\frac{1}{2} - \frac{\sum_{i=1}^{areat} tanw}{nw+ng+nhl}\right| \ \in [0,1]$$

Where: i: counter of text areas, tanw: number of words in each text area, areat: total number of text areas, nw: total number of words, ng: total number of graphics and nhl: total number of links

Graphical-density complexity (gdc)

$$1 - 2 \text{ x } \left|\frac{1}{2} - \frac{\sum_{i=1}^{areag} gan}{nw+ng+nhl}\right| \ \in [0,1]$$

Where: i: counter of graphic areas, gan: number of graphics in each graphic area, areag: total number of graphic areas, nw: total number of words, ng: total number of graphics and nhl: total number of links

Link-density complexity (ldc)

$$1 - 2 \text{ x } \left|\frac{1}{2} - \frac{\sum_{i=1}^{areahl} hlan}{nw+ng+nhl}\right| \ \in [0,1]$$

Where: i: counter of hyperlink areas, hlan: number of links in each link area, areahl: the total number of link areas, nw: the total number of words, ng: the total number of graphics, nhl: the total number of links.

Grouping complexity (gc)

Textual-grouping complexity (tgc)

$$\frac{\sum_{i=1}^{areat} tac}{areat} \ \epsilon \ [0,1]$$

Where: i: counter of text areas, tac: is the text area highlighted with colors, 1 if highlighted, and 0 if not. areat: the total number of text areas

Graphical-grouping complexity (ggc)

$$\frac{\sum_{i=1}^{areag} gal}{areag} \ \epsilon \ [0,1]$$

Where: i: counter of text areas, gal: is the graphic area highlighted as lists, 1 if highlighted, and 0 if not. areag: the total number of graphic areas

Link-grouping complexity (lgc)

$$\frac{\sum_{i=1}^{areahl} hlac}{areahl} \ \epsilon \ [0,1]$$

Where: i: counter of graphic areas, hlac: is the link area that its items highlighted with colors, 1 if highlighted, and 0 if not. areahl: the total number of link areas

Alignment complexity (ac)

Textual-alignment complexity (tac)

$$\frac{3}{(tan_{vp} + tan_{hp} + \tan)} \ \epsilon \ [0,1]$$

Where: tan_{vp}: number of text areas of vertical alignment points, tan_{hp}: number of text areas of horizontal alignment points and tan: number of text areas

Graphical-alignment complexity (gac)

$$\frac{\sum_{i=1}^{areag} gar}{areag} \ \epsilon \ [0,1]$$

Where: i: counter of graphic areas, gar: number of graphic areas aligned to the right and gan: total the number of graphic areas.

<u>Link-alignment complexity (lac)</u>

$$\frac{\sum_{i=1}^{areahl} hlar}{areahl} \;\epsilon\; [0,1]$$

Where: i: counter of hyperlink areas, hlar: number of link areas aligned to the right and $areahl$: total number of link areas

Unweighted Web Interface Complexity (UWIC)

Size Complexity (SC)= TSC+GSC+LSC

Density Complexity (DC)= TDC+GDC+LDC

Grouping Complexity (GC)= TGC+GGC+LGC

Alignment Complexity (AC)=TAC+GAC+LAC

UWIC= SC+DC+GC+AC

CHAPTER 4. EMPIRICAL STUDY

In the interest of our approach, we have conducted a sequence of experiments by which we can judge the solidity of our approach. Since web interfaces are the main component on in the research, we are putting several homepages of websites under the test of our metric model. Additionally, web users play a very important role in our investigation, which is confirming the rigidity of our approach.

In this chapter, we:

- Present the overall lab environment, recruitments, and participants
- Shortly describe the setup, the software tools, and the equipment of the experiment
- Present the tasks of the experiments, and the design of the surveys
- Discuss the calculation process of the metric variables
- Present the data collection formed by the metric model, and the users' judgment
- Evaluate our metric model by comparing it to the users' views on complexity

Laboratory Environment

Our lab was equipped with three desks and:

- Three personal computer with Microsoft Windows professional installed on them
- Each computer had 23.6-inch screen with 1920 x 1080 Full HD Resolution
- Each computer had Chrome browser and Microsoft Internet Explorer 8 browser
- Microsoft Windows Photo Viewer, Microsoft Paint, and Adobe Photoshop CC
- FireShot, a Chrome browser extension to capture screen shots of websites [32]

- Instant Wireframe, a Chrome browser extension to view any web page with a wireframe overlay [33]
- Each computer had a keyboard and mouse
- Google Forms

Study Preparation

An Institutional Review Board (IRB) form was submitted to the representatives of the Research and Creative Activity center to obtain approval to conduct our experiment that involves human subject. The complete form and approval letter are included in the appendixes of this document. A public recruitment letter was sent out to North Dakota State University Student Research Participant Listserv in which a detailed explanation of the research, participation requirements, compensations, and a link to the researcher's Doodle account. The account is an online scheduling tool that can be used easily to locate a date and time to meet with multiple participants. The complete recruitment letter is included in the appendixes of this document.

A formal script is verbally communicated to the participants demonstrating the tasks as well as, a formal consent form presented to the participant to be signed which are included in the appendixes of this document. Pre-evaluation questionnaires were filled up by participants to make sure participants match the required profile, and to check whether any effects observed are dependent on demographic attributes. The complete pre-evaluation questionnaires are included in the appendixes of this document as well.

Two main experiment were conducted:

- The first experiment was conducted to understand and confirm the fundamental units of our metric model, which have been the center pillars of various studies in the literature.

- The second experiment was conducted to discover the probability of finding a compliance between the metric model numbers and the users' evaluations of the units, factors, elements, and attributes of web interface complexity.

Each experiment was directed twice, for each time we maintained the same sample of participants and the same sets of websites. The reason behind that is to examine the correctness and the robustness of the experiment design and the metric model outcomes.

Websites

Websites selection criteria

As mentioned the two experiment were conducted twice during the study period. Thus, we adjusted the criteria slightly different for each set of experiments:

Table 4. Experiments' websites differences and criteria

Websites' Criteria of the DE	Websites' Criteria of the EE
• Unranked, ungraded and un-awarded websites • Variety and multiplicity of visual objects • Shrinkage and shortage of visual objects • Random genres	• Ranked, graded and awarded by The International Academy of Digital Arts and Sciences (IADAS), which its judging figure consisted of over one thousand industry experts and technology innovators [34]. • Variety and multiplicity of visual objects • Shrinkage and shortage of visual objects • Random genres

Websites genres

Diversification of web genres adds some strength points to the experiments, and to research in general because, with this variety, the metric model can be tested relatively under different circumstances and conditions which have a huge impact on metric variables. We selected multiple genres for each version of the two experiments. Consequently, there were five websites for each and there are as follows:

Table 5. Experiments' websites genres

Web sites of the Development Experiment		Web sites of the Evaluation Experiment	
Genres	URL Title	Genres	URL Title
Educational	www.psu.edu	Educational	www.hampshire.edu
Celebrity	www.graceland.com	Personal Blogs	www.thoughteconomics.com
Shopping	www.dujour.com	Art	www.guggenheim.org
News	www.forbes.com	Companies	www.oracle.com
	www.msn.com	News	www.theguardian.com

Websites figures

The home pages of all selected websites were screenshotted using the FireShot [32] tool, which allowed us to acquire the screenshots with no compromises in the actual resolution. All figures of real and model screens are included in the appendixes of the document.

Modeling the screens of web interfaces

In order to model the screens of the web interfaces as we describe in chapter 3, we needed to utilize several tools to identify, wireframe and label the targeted structural element, Texts, Graphics, and Links. First, we made three identical copies of each screenshot of the real screen, and then we positioned them on one image frame for each screenshot as well, using Microsoft Paint and Adobe Photoshop CC. Second, we wireframed them to identify each structural object on the real screen in order to obtain the actual dimensions and locations, using Instant Wireframe [33]. Third, we titled each screenshot with an element name. Thus, we had three model screens plus the real screen, and then we highlighted all elements that correspond to the title of the screenshots, using Microsoft Paint and Adobe Photoshop CC.

Participants

The total number of participants for both experiments at both courses of execution is eighty-seven people, and Table 6 gives a demonstration of the different demographic attributes of the participants. As mentioned before the study was executed in an academic enlistment, at North Dakota State University, therefore, the participants were all students and all demographics information is based on users' entries.

Table 6. Participants demographics of the experiments

Demographics	Participants of Development Experiments	Participants of Evaluation Experiments
Genders	24 males, 16 females	35 males, 12 females
Ages	18-25 were 23 26-39 were 14 40-59 were 3	18-25 were 27 26-39 were 20 40-59 were 0
Education Levels	28 graduates, 12 undergraduates	29 graduates, 18 undergraduates
Computer Related Fields	33 computer related fields, 7 unrelated fields	34 computer related fields,13 unrelated fields
Web Surfing Hours	1-3 were 15 4-6 were 16 7 and more were 9	1-3 were 17 4-6 were 19 7 and more were 11
Web Surfing Levels	Experts 16 very good 23 moderate 1	Experts 20 very good 18 moderate 8

Surveys and Questionnaires

One questionnaire and three surveys were designed to satisfy the need of the experiments, and the complete copy of the questionnaire and the surveys are attached in the appendixes of this documents.

Questionnaire design

A mutual pre-evaluation questionnaire, which was used in the first and the second experiments, was presented to the participants to collect all required demographics information, and the first page of it, had the consent form.

First experiment survey design

The first experiment had only one major survey:

<u>The first survey</u>

It had two types of questions: 1) open-ended and 2) close-ended questions. On the first type, we wanted the participants to express their generic views on complexity freely. Therefore, we allowed them to type five statements to have their input about the structural factors, reasons, elements, and objects of complexity. Thus, we collected overall 200 statements about the structural complexity from the point view of the participants. In the second type, we presented our fundamental structural measures and elements with others in the form of lists, and they had to evaluate them based on 5-point Likert scale.

Second experiment survey designs

The second experiment had two one major surveys:

<u>The first survey</u>

The main goal of this survey is acquiring participants' judgments on the four structural factors because they are the base of our metric model. Thus, they were asked to compare the regular model screens of all web interfaces against the real screens, using a 5-point Likert scale to express their views of complexity.

<u>The second survey</u>

In this survey, the goal was to gather participants' views and judgments on the web interfaces using our modified version of the model screen. Consequently, this form had to be filled up for each website. However, in this evaluation, they had to use Microsoft Windows Photo Viewer to look, observe and answer what had been asked on the survey's form. Also, the 5-point Likert scale was utilized to express their views of complexity. Moreover, before the second time of the execution of this survey, we improved it slightly by adding some visual aids that explain some technical terminologies in the context of GUIs such as Balance and Unity of objects. We also changed the moderating techniques, subsequently, instead of the adopting the Concurrent Probing (CP) that needs interactions from the moderators at every time participants have questions, we switched to Retrospective Probing (RP) that does the opposite by waiting until the session is complete and then they do the interactions.

Tasks of Experiments

As described in the previous sections, two techniques were employed in the evaluation sessions, which are CP and RP. The two experiments were conducted at one session for all surveys, and we had two sessions as we explained before that, and the second session had a different sample of participants. In addition, it had slight changes in the websites selections, survey designs, and moderation techniques. The study took approximately 35-45 minutes. Two browsers were open on the screen for each participant, the one on the left of the screen was Microsoft Internet Explorer 8, and the second had Chrome, and it was occupied with the questionnaire and the three surveys on three other tabs. A sample example of the task scenario is like the following:

- The moderator presents and verbally deliver the participants rights and the study's references and representatives.
- The moderator presents the actual setup of the experiment such as the purpose of the two browsers, the number of surveys, and their essence

- The moderator explains the rules of evaluation process such as
  - Each user had to switch between the tabs of the websites
  - No navigation activities allowed except for scrolling, and reflect his/her opinions to associated questions on the other browser.
  - Mouse usage was to zoom in and out of websites screenshots
- The moderator interacts with participants to give overall input on the experiment
- Participants were rewarded with $10 cash as a compensation for their time.

Calculating the Structural Complexity

To calculate the structural complexity based on the participants' judgment was an easy task because we only needed to export the data from Google forms as spreadsheets. In contrast, calculating it based on the metrics model was a hard and lengthy task. We had to count and distinguish every different variable that exists in the functions of the metric model, which was discussed in chapter 3, for all web interfaces. Next, we had to apply and the functions to output the results.

CHAPTER 5. RESULTS, ANALYSIS, AND EVALUATIONS

In this chapter, we present, analyze, evaluate and predict the data that was produced by the calculation of empirical study's data. To infer conclusions, statistical analysis, which encompasses assembling and studying every data sample, must be implemented. A sample, in statistics, is a demonstrative range drawn from a total of population. The mechanism that can be followed to execute the statistical analysis can be summarized in five phases:

- Specifying types of the data to be examined.
- Discovering the relation of the data to the root population.
- Generating a model to review understanding of how the data relates to the original population.
- Proving or disproving the legitimacy of the model.
- Using indicative analytics to operate states that will support and guide future activities.

The aim of statistical analysis is to recognize trends and patterns. The metric model, for example, might employ statistical analysis to discover patterns in sets of structural measures and elements over some other sets [35].

As stated in [36] there are two main branches in statistics which must be used to permit us to depict the big picture of the data under examination. The two branches are descriptive statistics which provide a concise summary of data and give information that describes the data. For example, the percentage of the direct and indirect statements, which the experiments' participants entered, points to our fundamental components of the metric model. The second branch is inferential statistics, which enable us to make inferences about populations using data drawn from the population. Subsequently, we can use hypothesis testing, correlation testing, and regression analysis to generalize a concept, and that what we demand to test the correctness of the research hypotheses. In addition, it is important to note that we use both branches on the data formed by the metric calculations and users' data input. Moreover, depending on the data samples and their types, certain statistical tests can be used. According to [37, 38], nonparametric statistics refer to a statistical technique where the data is not required to embrace a normal distribution. Nonparametric statistics employ data that is frequently ordinal, meaning it does not depend on numbers, but rather a ranking, order of sorts, or a number of occurrences. In our data, such style of data exists, which is the count of statements, references, mentions, and naming of factors and elements in the open-end questions that asked in the surveys. Also, there are the parametric statistics, which undertake that sample data originates from a population that belongs to a probability distribution based on a static set of parameters. The following table displays sets of statistical tests, metrics, and data types:

Table 7. Selecting the correct statistics for diverse data types and usability metrics

Data Type	Popular Metrics	Statistical Test
Nominal (classes)	Task success (0 or 1)	Frequencies, crosstabs, Chi-square
Ordinal (ranks)	Severity ratings, rankings (designs)	Frequencies, crosstabs, chi-square, Wilcoxon rank sum tests
Interval	Likert scale data, SUS scores	All descriptive statistics, t-tests, ANOVAs, correlation, regression analysis
Ratio	Completion time, average task success	All descriptive statistics, t-tests, ANOVAs, correlation, regression analysis

First Experiment

As already stated, the goals of conducting the first experiment were:

- To understand the meaning of complexity from web interface users' perspectives

- To perceive how the users of web interfaces connect the dots between complexity and visual factors and objects
- To detect to what extent the users of web interfaces can express the concept of web interface complexity
- To confirm the legitimacy and validity of the frequently reported claims of the four factors of complexity
- To verify the effectiveness of exposing the three structural elements to the users of web interfaces as the building blocks of the web content.

Therefore, the two branches of statistics mentioned above performed on users' data only, and based on which the metric model was developed. Also, we briefly outline the research questions and the hypotheses that have driven the first experiment of the study.

Hypotheses

The generic frame of the metric model was directed based on questions and hypotheses in this section. Consequently, the survey questions were designed and phrased as well based on the hypotheses. Moreover, questions and hypotheses were examined by analyzing the data which produced in the first experiment; we summarize the research questions and the hypotheses related to it as follows:

- Q1. Which can structural factors be a better predictor for web interfaces' complexity to users?
- Q2. Which can structural objects be a better predictor for web interfaces' complexity to users?
- Q3. Which have structural factors high level of importance to the users of web interfaces?

- Q4. Which have structural objects high level of importance to the users of web interfaces?

Considering these question, we formulated two main hypotheses for the first experiment. The first hypothesis addresses the first two questions, and the second hypothesis addresses the last two questions.

- H1. The sum of the structural factors: Size, Density, Grouping, and Alignment is a better predictor than the sum of Spacing, Balance, Regularity, Sequence, Unity and Color for measuring the web interface complexity.
- H2. The sum of structural object, which are text, graphics, and hyperlinks, is a better predictor than the sum of Buttons, Menus, Audios, Videos and Search Boxes for measuring the web interface complexity.

These main hypotheses will be broken down into several hypotheses in order to test the validity of each one. The central idea of approaching the hypotheses this way is to simplify the work and makes much more quantifiable.

Data collection and analysis

The following list describes the types of data and the questions of the first experiment:

- <u>Open-ended questions and frequency count:</u> we are able to collect naming, references, and mentions, as displayed on the first survey. After the statement analysis procedure, we counted the direct and indirect mentions of the four and three structural complexity measures and elements of our model.
- <u>Close-ended questions and ratings' sum:</u> the third type of the data collection is the second type, but we sum the ratings of the four measures and the three elements in all cases.

- with other specific measures and elements,
- with no other measures and elements,
- the ratings of the four factors against each other.
- the ratings of the three elements against each other.

Factors of complexity

There are two main types of data collected from the surveys. The first type of data collection and results is the result of analyzing the statements, which was a total of 603 naming, references and mentions of different factors, and the data was nominal data. We counted the references to structural factors, and we used the Wilcoxon signed-rank test which is a nonparametric test to perform the analysis. The detailed description and analysis are as follows:

- Each individual participant entered five factors, he/she thinks each factor can cause complexity to web interfaces
- For each data entry of each factor, we counted the frequencies
- Then we calculated the expected frequencies using Wilcoxon signed rank test to obtain the p-value
- The null hypothesis is H_0: the frequencies of mentions for the four structural factors: size, density, grouping and alignment compared to the other factors are equal in both data sets.

As shown in Table 8, the p-values are 0. 011 which is dramatically less than 0.05, which means the null hypothesis cannot be accepted, and there is a huge difference between the two sets of data, references for the 4 structural factors and other factors' references, in both experiments.

Table 8. Examining the significance level of the selected four factors of the metric against the other factors by using Wilcoxon test on frequencies of users' inputs

Frequencies of factors in users' inputs					
Total of References for the 4 Structural Factors		Other Factors' References		Total of All Factors' Types of References	
DE	EE	DE	EE	DE	EE
62	39	10	15	72	54
51	41	18	15	69	56
43	38	27	15	70	53
33	34	27	16	60	50
42	37	22	18	64	55
The p-value of Wilcoxon signed rank test (Development Experiment)				0.01193	
The p-value of Wilcoxon signed rank test (Evaluation Experiment)				0.01116	

The data indicates that users tend to mention the selected four factors of our model as factors of complexity more than the other six factors. That grants us a good suggestion to depend more on these factors to be used for the analysis and calculation of complexity of web interfaces. To strengthen this approach, we need to do the second type of data collection.

The descriptive statistical analysis is needed here to calculate the medians and means. In fact, this type of statistics allows us to describe and summarize data in ways that are meaningful and useful for other calculations and statistical tests, by using the measures of central tendency and measures of variability, or dispersion such as the mean, median, range, and variance [38]. Depending on the data type collected, some suitable statistical procedures will be implemented.

The second data type of the first experiment is analyzed as shown in Table 9.
The result of the data collection in this type, which has close-ended questions, is rating on a scale of five, Likert Scale, for each factor of the four with other factors, and then the four factors alone, the detailed description is as follows:

- Each individual participant rate the five factors, he/she thinks each factor can cause complexity to web interface
- For each factor, we calculate the descriptive measures
- Then we calculate the t-test to obtain the p-values

The rating data are included in the appendixes section of this document.

There are several cases or hypotheses that we need to develop to explain the different rating of factors:

A. The ratings of the four factors along with the other six factors, which its null hypothesis is H_0: the means of ratings of the four structural factors and the means of the six structural factors are equal in both experiments.
B. The ratings of the four factors against each other in both experiments, which its null hypothesis is H_0: the means of ratings of the four structural factors in the development experiment are equal to the evaluation experiment.
C. The ratings of the four factors in each experiment shows one or more factors have higher importance levels than the others

Table 9. Examining the equivalence relationship of the users' rating data between the selected four factors and the other six factors by conducting the t-test

Statistical Measures	DE		EE	
	Structural Four Factors	Structural Six Factors	Structural Four Factors	Structural Six Factors
Mean	3.5666	3.066	3.6032	3.2608
Variance	1.6859	1.483	1.278	1.298
Observations	240	360	184	276
t Stat	4.736		3.1712	
P(T<=t) one-tail	0.00000142		0.000818	
t Critical one-tail	1.64797		1.648730	
P(T<=t) two-tail	0.0000028		0.0016366	
t Critical two-tail	1.96482		1.96600	

The p-values in Table 9 addresses the first null hypothesis which equals to 0.00000142 and 0.000818 for the development and the evaluation experiment respectively. Hence, the values are exceptionally less than the Alpha value, 0.05. Consequently, we reject the null hypothesis because there is enough evidence to conclude that the difference between the two sets of data in both experiments is massive. That means the users tend to rate the four factors of our model more than the other six as factors of complexity for web interfaces. The following figures show the convergence of ratings for all factors and the level of ratings.

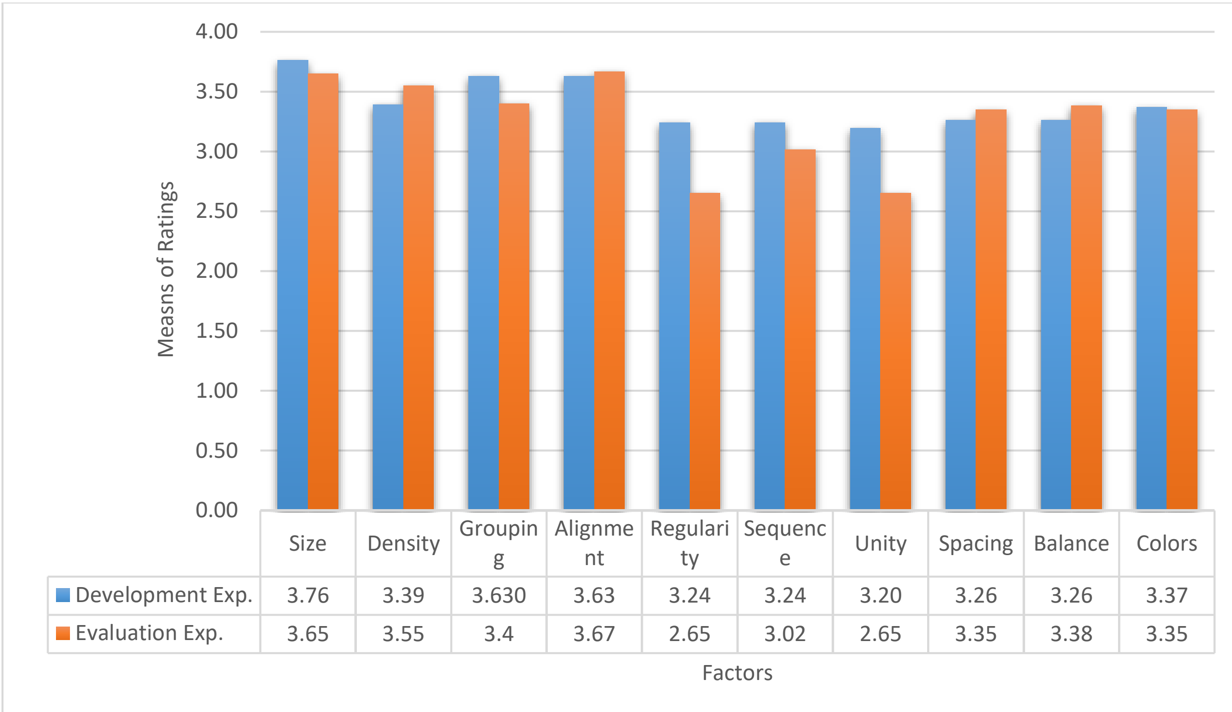


Figure 22. Means of users' ratings of complexity factors on a scale of five

This previous figure shows the means of ratings for all factors, and it illustrates that the four factors of our complexity model have the highest means among the other factors, but as we can notice some of the other factors have high numbers of means which also indicates significance.

Moving to the second case of our analysis, which is inspecting the users' rating of the four factors against each other in both experiments. We hypothesize that the means of users' ratings in the development and the evaluation experiment going to be equal. Therefore, we implemented the t-test, and the result was 0.216 for the p-value as shown in Table 4. That means it is enormously greater than the level of significance 0.05 which means that we do have enough evidence to reject the null hypothesis. Thus, we have enough evidence to conclude that there is a huge difference between the two sets of data in both experiments.

Table 10. Examining the equivalence relationship of the users' rating data between the selected four factors by conducting the t-test

t-Test: Two-Sample Assuming Unequal Variances		
Statistical Measures	Four Factors of DE	Four Factors of EE
Mean	3.5125	3.429
Variance	1.1881	1.153
Observations	240	184
t Stat	0.7850	
P(T<=t) one-tail	0.21644	
t Critical one-tail	1.64870	
P(T<=t) two-tail	0.43289	
t Critical two-tail	1.96595	

Consequently, even though there are few changes between the two experiments, the users tend to confirm that these factors have a high level of importance to affect the complexity of web interfaces.

To add more clarification to our understanding of the users' rating data, we needed to look at the factors and point to the factors that have a higher level of significance over the other among the four, since they have a higher level of importance over the other six. Subsequently, we applied the PCA on the ratings of the four factors, and the results were as shown in Table 5 and 6. Whereas, two components represent 65 to 69 percent of web interface complexity in both experiments based on the cumulative proportion.

Table 11. PCA - Importance of components of the four factors based on users' ratings

	DE				EE			
Importance of components	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 1	Comp. 2	Comp. 3	Comp. 4
Standard deviation	1.217	1.074	0.940	0.690	1.297	1.043	0.838	0.723
Proportion of Variance	0.370	0.288	0.221	0.119	0.42	0.27	0.17	0.13
Cumulative Proportion	0.370	0.659	0.880	1.000	0.42	0.69	0.86	1.00

Moreover, the data of the development experiment shows size and density factors have lower loadings on the first component, but a higher loading on the second component. In addition, the grouping and alignment factors behave inversely on the two components. Loadings of factors data on the components of the evaluation experiment are slightly different, whereas, all factors have high lodgings on the first components except the density factor, but it has a high loading on the second. Also, the size and grouping factors have low loadings on the second component and high loadings on the first. From the behavior of the loadings' data of the factors, we can recognize a pattern that helps us to understand and explain such behavior, which is all factors have high loadings on either significant components, and none of them do not have high loadings on at least one of the significant components. Thus, we conclude that all factors have relatively the same level of importance.

Table 12. PCA - Loadings of the four factors based on users' ratings

	DE				EE			
Loadings	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 1	Comp. 2	Comp. 3	Comp. 4
Size	-0.39	-0.55	0.63	0.38	0.50	0.42	0.73	0.22
Density	-0.30	-0.67	-0.54	-0.39	0.33	0.70	-0.63	
Grouping	-0.61	0.30	-0.45	0.57	0.60	-0.30		-0.74
Alignment	-0.61	0.37	0.31	0.61	0.52	-0.50	-0.25	0.63

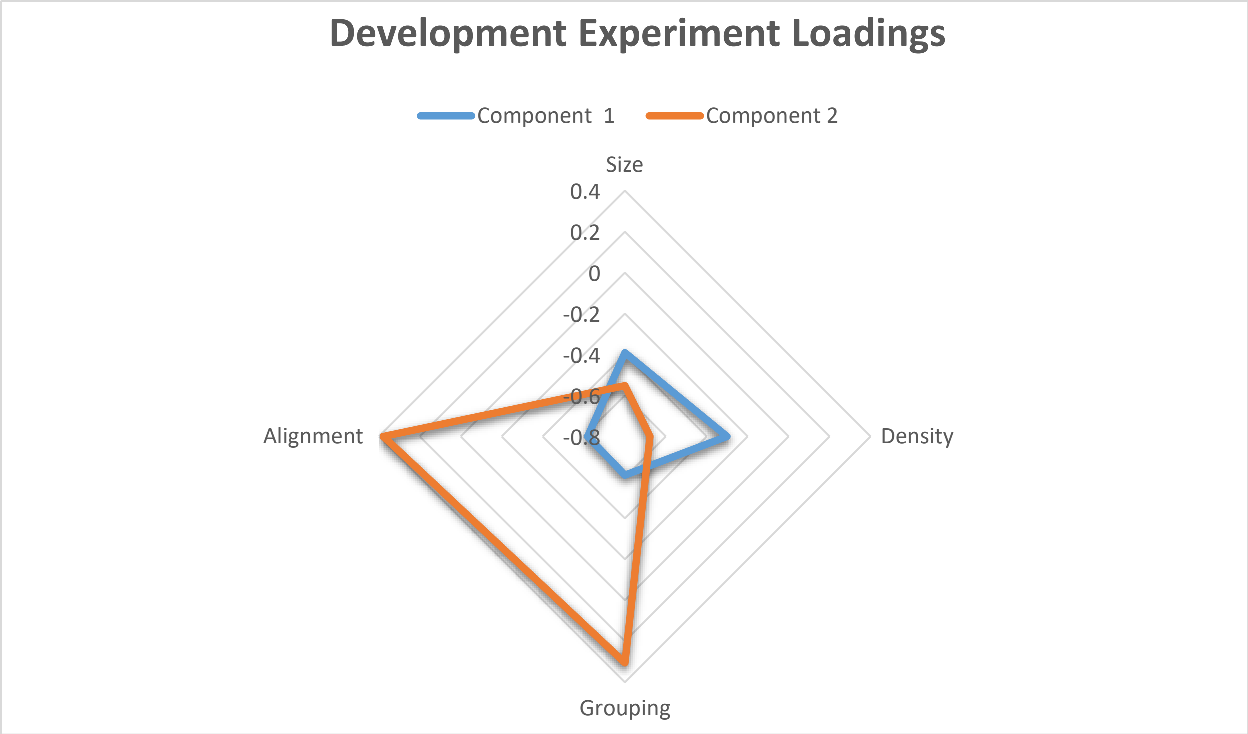


Figure 23. Radar chart shows the loadings of the users' ratings of the four factors in the development experiment

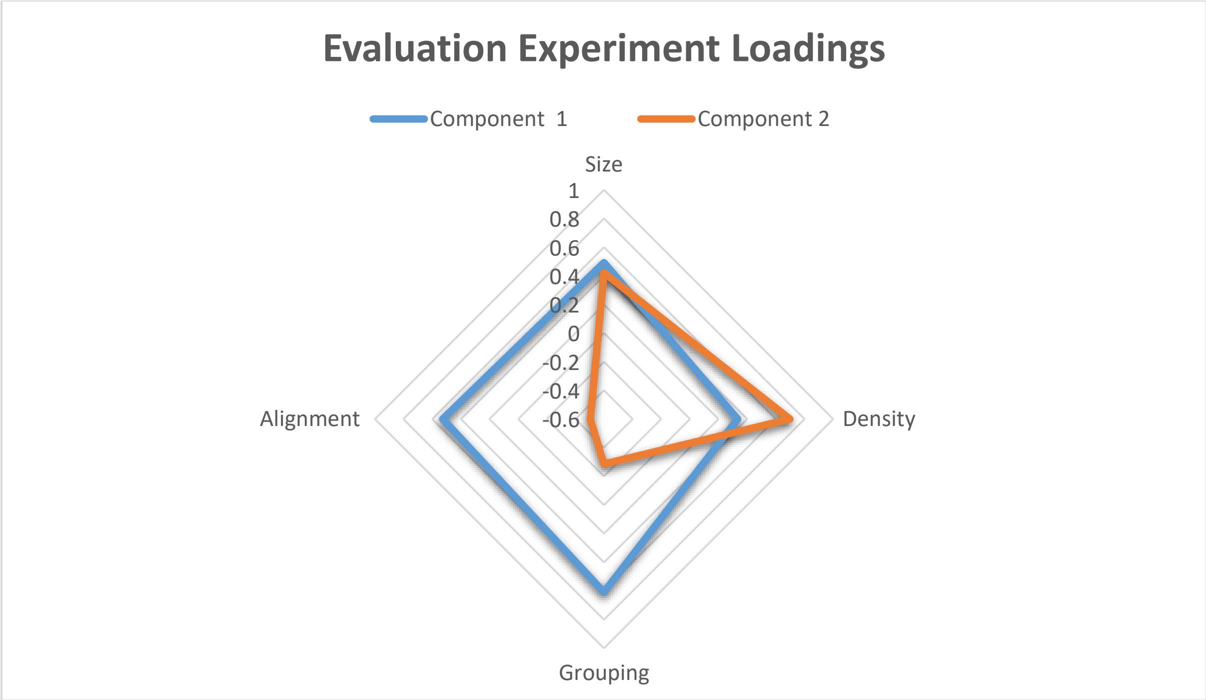


Figure 24. Radar chart shows the loadings of the users' ratings of the four factors in the evaluation experiment

Elements of complexity

Respectively, the elements of complexity are analyzed and evaluated following the same methods occurred to the factors. Therefore, the first type of data collection and results consists of The result of analyzing the statements, which was a total of 575 naming, references and mentions of different elements and their detailed description is as follows:

- Each individual participant entered five factors, he/she thinks each element can cause complexity to web interface
- For each data entry of each element, we counted the frequencies as shown in Table 13
- Then we calculated the expected frequencies using Wilcoxon signed rank test to obtain the p-value

- The null hypothesis is H_0: the frequencies of mentions for the three structural elements: texts, graphics, and links compared to the other elements are equal in both data sets.

As shown the p-values are 0. 011 which is dramatically less than 0.05 which means the null hypothesis cannot be accepted, and there is a huge difference between the two sets of data, references for the 4 structural factors and other factors' references, in both experiments.

Table 13. Examining the significance level of the selected three elements of the metric against the other elements by using Wilcoxon test on frequencies of users' inputs

Total of References for the 3 Structural Elements		Other Elements' References		Total of All Elements' Types of References	
DE	EE	DE	EE	DE	EE
60	63	12	6	72	69
57	38	12	17	69	55
58	47	19	14	77	61
39	24	18	17	57	41
41	27	19	12	60	39
The p-value of Wilcoxon signed rank test (Development Experiment)				0.01167	
The p-value of Wilcoxon signed rank test (Evaluation Experiment)				0.01193	

The data indicates that users tend to mention the selected three elements of our model as elements of complexity more than the other five elements. That grants us a good suggestion to depend more on these elements to be used for the analysis and calculation of complexity of web interfaces. To support this approach, we need to do the second type of data collection.

Repeatedly, descriptive and inferential statistical analysis is needed here to calculate the medians, means, and t-test. The following table presents these statistical measures implemented on the data of the complexity elements. The result of the data collection in this type, which has close-ended questions, is rating on a scale of five, Likert Scale, for each element of the three with other elements and then the three elements alone, the detailed description is as follows:

- Each individual participant rate the five elements, he/she thinks each element can cause complexity to web interface
- For each element, we calculate the descriptive measures
- Then we calculate the t-test to obtain the p-values

The rating data are included in the appendixes section of this document.

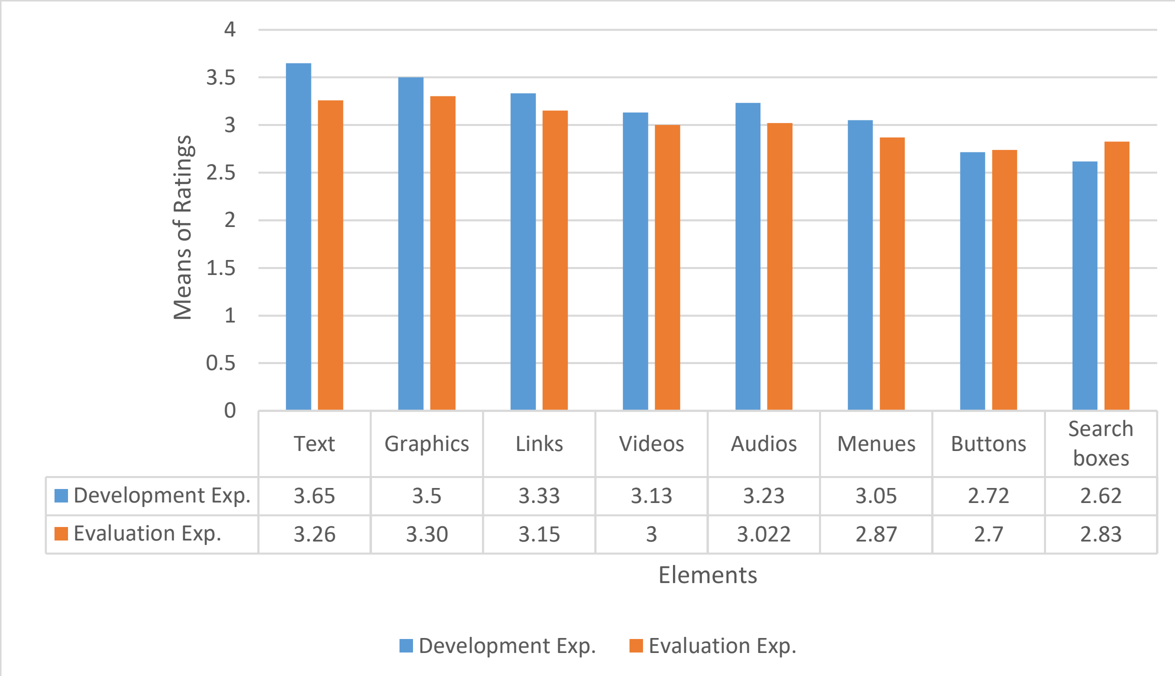
There are several cases or hypotheses that we need to develop to explain the different rating of factors:

A. The ratings of the three elements along with the other five elements, which its null hypothesis is H_0: the means of ratings of the three structural elements and the means of the five structural elements are equal in both experiments.
B. The ratings of the three elements against each other in both experiments, which its null hypothesis is H_0: the means of ratings of the three structural elements in the development experiment are equal to the evaluation experiment.
C. The ratings of the three elements in each experiment shows one or more elements have higher importance levels than the others

Table 14. Examining the equivalence relationship of the users' rating data between the selected three elements and the other six factors by conducting the t-test

Statistical Measures	DE		EE	
	Structural Three Elements	Structural Five Elements	Structural Three Elements	Structural Five Elements
Mean	3.4944	2.95	3.2391	2.8913
Variance	1.2904	1.5326	1.6723	1.721
Observations	180	300	138	230
t Stat	4.91348		2.48420	
P(T<=t) one-tail	6.51783E-07		0.0067724	
t Critical one-tail	1.64864		1.65008	
P(T<=t) two-tail	1.30357E-06		0.013544709	
t Critical two-tail	1.96586		1.96812	

The p-values in Table 8 addresses the first null hypothesis which equals to 6.51783E-07 and 0.0067724 for the development and the evaluation experiment respectively. Hence, the values are exceptionally less than the Alpha value, 0.05, especially for the development experiment. Accordingly, we reject the null hypothesis because there is enough evidence to conclude that the difference between the two sets of data in both experiments is enormous. That means the users tend to rate the three elements of our model more than the other five as elements of complexity for web interfaces. The following figures show the convergence of ratings for all elements and the level of ratings.



	Text	Graphics	Links	Videos	Audios	Menues	Buttons	Search boxes
Development Exp.	3.65	3.5	3.33	3.13	3.23	3.05	2.72	2.62
Evaluation Exp.	3.26	3.30	3.15	3	3.022	2.87	2.7	2.83

Figure 25. Means of users' ratings of complexity elements on a scale of five

This previous figure shows means of ratings for all elements, and it demonstrates that the three elements of our complexity model have the highest means among the other factors, nonetheless as we can notice some of the other elements have high numbers of means which also indicates significance.

Moving to the second case of our analysis, which is inspecting the users' rating of the three elements against each other in both experiments. We hypothesize that the means of users' ratings in the development and the evaluation experiment going to be equivalent. Therefore, we implemented the t-test, and the result was 0.393 for the p-value as shown in Table 4. That means it is enormously greater than the level of significance 0.05 which means that we do have enough evidence to reject the null hypothesis. Thus, we have enough evidence to conclude that there is a huge difference between the two sets of data in both experiments.

Table 15. Examining the equivalence relationship of the users' rating data between the selected three elements by conducting the t-test

t-Test: Two-Sample Assuming Unequal Variances		
Statistical Measures	Three Element of the DE	Three Elements of the EE
Mean	3.3111	3.2753
Variance	1.5451	1.1936
Observations	180	138
t Stat	0.272	
P(T<=t) one-tail	0.39278	
t Critical one-tail	1.64978	
P(T<=t) two-tail	0.7855	
t Critical two-tail	1.96764	

Consequently, even though there are few changes between the two experiments, the users tend to confirm that these elements have a high level of importance to affect the complexity of web interfaces.

To add more explanation to our understanding of the users' rating data, we needed to consider the elements and specify the elements that have a higher level of significance over the other among the three, since they have a higher level of importance over the other five. Afterward, we applied the PCA on the ratings of the three elements, and the results were as shown in Table 10 and 11. Whereas two components represent 75 percent of web interface complexity in the development experiment, and one component represents 59 percent of web interface complexity in the evaluation experiment based on the cumulative proportion.

Table 16. PCA - Importance of components of the three elements based on users' ratings

	DE			EE		
Importance of components	Comp. 1	Comp. 2	Comp. 3	Comp. 1	Comp. 2	Comp. 3
Standard deviation	1.110	1.019	0.852	1.333	0.881	0.667
Proportion of Variance	0.411	0.346	0.242	0.592	0.259	0.148
Cumulative Proportion	0.411	0.757	1.000	0.592	0.851	1.000

Moreover, the data of the development experiment shows graphics element have lower loadings on the first component, but the highest loading on the second component. In addition, the texts and links elements behave inversely on the two components. Loadings of elements data on the components of the evaluation experiment are slightly different, whereas, all elements have high lodgings on the first components except the links element. We can recognize the links have low loading on the second component of the development experiment and slightly below the average value of loadings on the significant component of the evaluation experiment. Even though links have a low value of loadings in the evaluation experiment, links have the highest loading value on the first component of the development experiment. We conclude that all three elements have relatively high level of importance based on the collective indicators driven from the t-test that show the three elements strongly related.

Table 17. PCA - Loadings of the three elements based on users' ratings

	DE			EE		
Loadings	Comp. 1	Comp. 2	Comp. 3	Comp. 1	Comp. 2	Comp. 3
Texts	0.655	-0.423	-0.627	0.612	-0.410	0.676
Graphics	0.206	0.898	-0.390	0.632	-0.260	-0.730
Links	0.727	0.126	0.675	0.475	0.874	

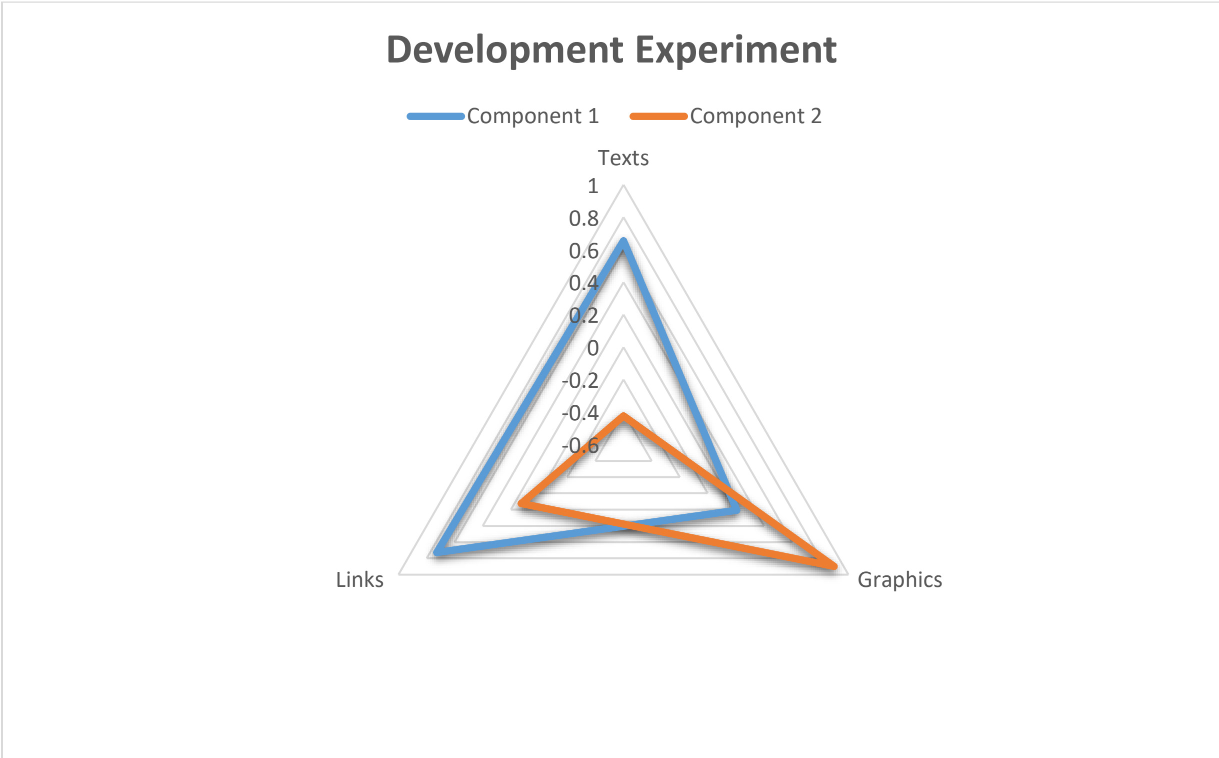


Figure 26. Radar chart shows the loadings of the users' ratings of the three elements in the development experiment

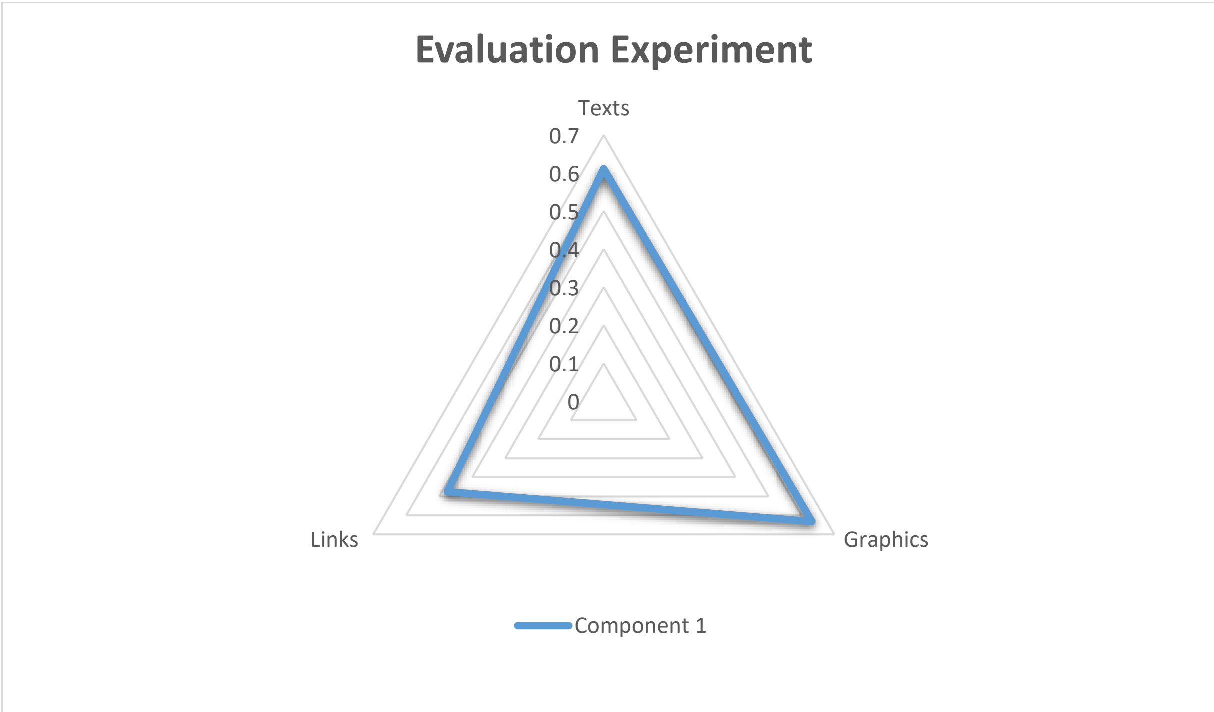


Figure 27. Radar chart shows the loadings of the users' ratings of the three elements in the evaluation experiment

Complexity factors of elements

In this section, we analyze the data of ratings that users produced in answering close-ending questions that reflects their understanding of each structural factor of the four applied on each structural element of the three. By carrying out this inquiry, the result allows us to build a hypothesis by which we can draw a picture of which the elements and factors have implication over the other. The complete tables of PCA for this approach are included in the appendixes. This inquiry contains the users' ratings and evaluations of each factor in the context of each element, accordingly, we have four factors multiplied by three elements which equal twelve. In both experiments, the development and the evaluation, four components have a stander deviations equal or greater than one. The four components, which were driven by the data of the PCA of the developments and the evaluation experiments, represents 73 to 75 percent of the data

based on cumulative proportion. Furthermore, in the loadings data of the PCA of the development experiment, all items have loadings values below the average which are 0.5 except the density of graphics has a loading value above the average on the fourth component, 0.614. The may indicate that all items have the same level of loadings and importance and the density of graphics is the highest. Almost the same happens with the loadings data of the PCA of the evaluation experiment, where all items have loadings values below the average which are 0.5 except the size of texts has loading values slightly above the average on the third and fourth components, which are 0.538 and 0.51 respectively. Similarly, this indicates that all items have the same level of loadings and importance and the size of text is the highest. The following two radar charts shows the loadings on the components data on the different items of complexity.
In consequence, we need to conduct a second experiment to answer more questions as described in the second experiment section.

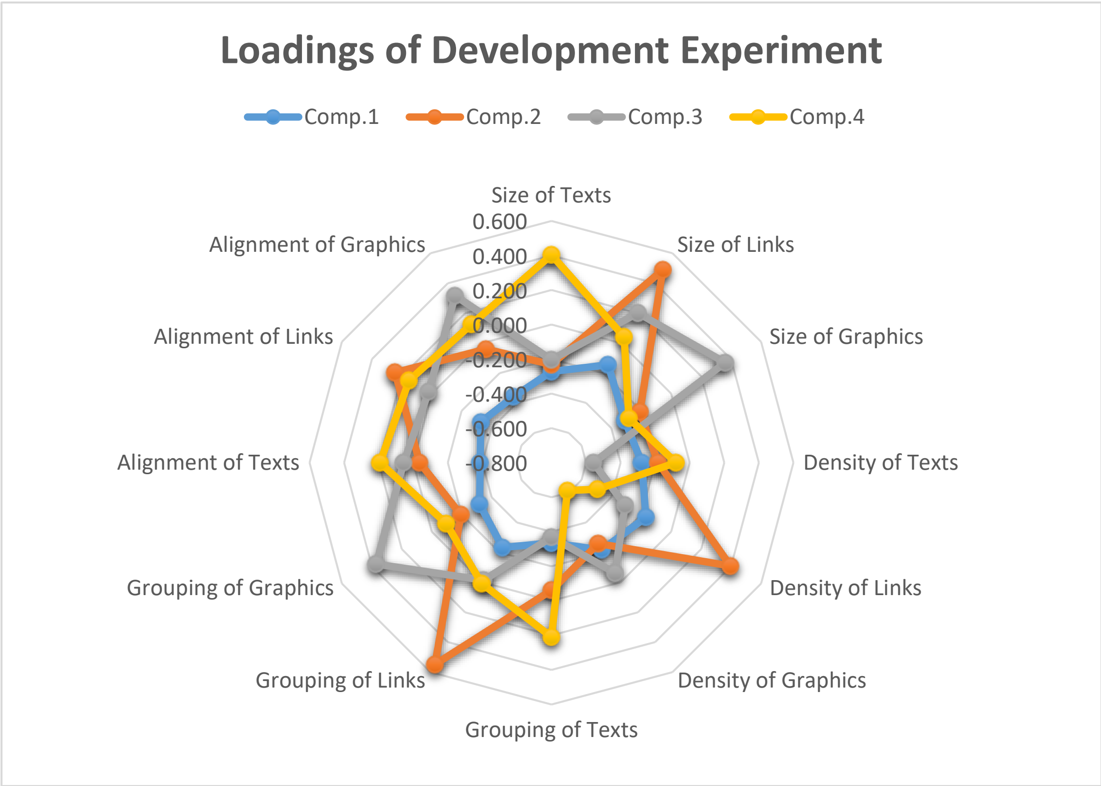


Figure 28. Radar chart shows the loadings of the users' ratings of the twelve-complexity items in the development experiment

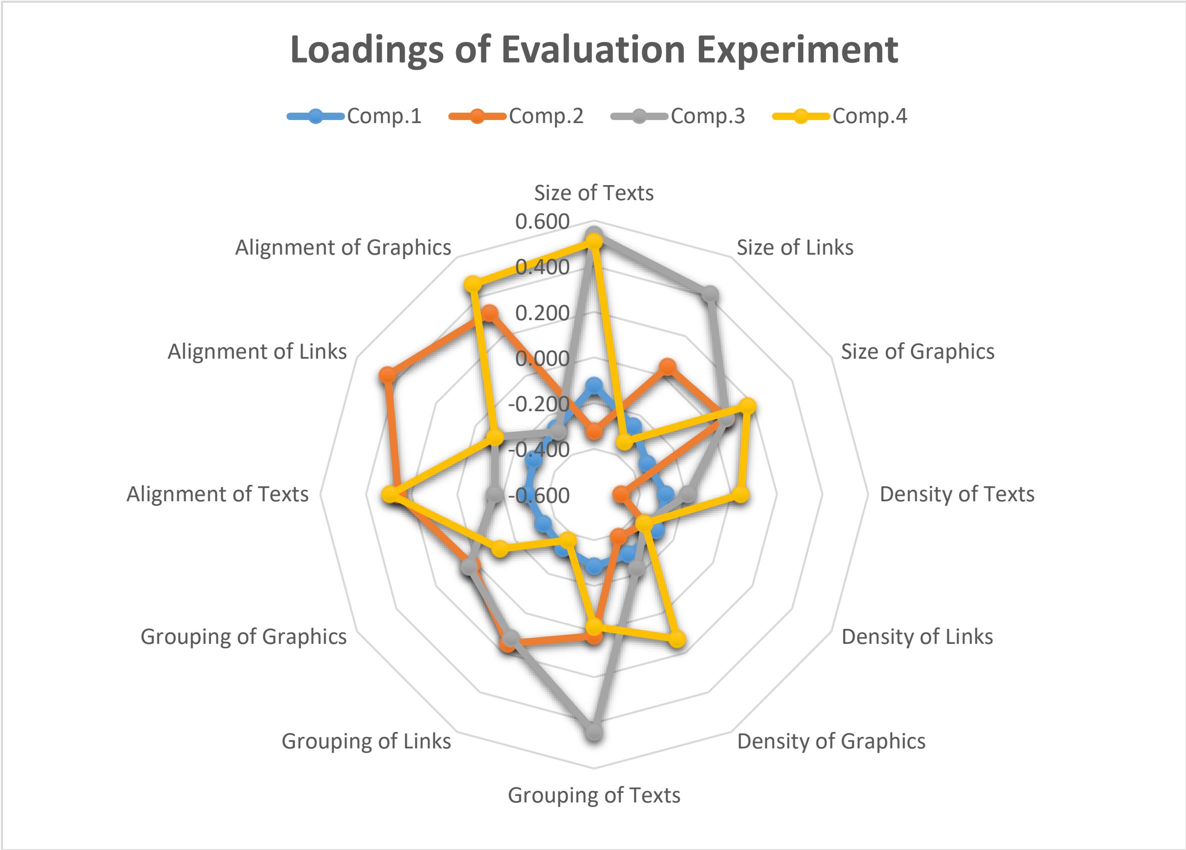


Figure 29. Radar chart shows the loadings of the users' ratings of the twelve-complexity items in the evaluation experiment

Second Experiment

As already stated, the goals of conducting the first experiment were:

- To test the measurability of the web interfaces complexity
- To investigate the types of attributes that cause complexity from web interface users perspectives
- To test the benefits and accuracy of using Model Screens
- To validate the metric model under development
- To detect to what extend our model can be generalized
- To quantify the complexity of web interfaces using certain equations

Therefore, the two branches of statistics mentioned above are performed on users' data and results of the complexity equations. Also, as we did with the first experiment, we briefly outline the research questions and the hypotheses that have driven the second experiment of the study.

Hypotheses

The basic border of the metric model was focused based on questions and hypotheses in section. Consequently, the survey questions and metric equations were designed and phrased as well based on them. Moreover, questions and hypotheses were inspected by evaluating the data which shaped in the second experiment; we review the research questions and the hypotheses related to it as follows:

- Q1. What characteristics of Size factor can be used to measure the level of complexity, which is determined by the users' opinions from the first study of Texts, Graphics, and Hyperlinks on web interfaces?

- Q2. What characteristics of Density factor can be used to measure the level of complexity, which is determined by the users' opinions from the first study of Texts, Graphics, and Hyperlinks on web interfaces?
- Q3. What characteristics of Grouping factor can be used to measure the level of complexity, which is determined by the users' opinions from the first study of Texts, Graphics, and Hyperlinks on web interfaces?
- Q4. What characteristics of Alignment factor can be used to measure the level of complexity, which is determined by the users' opinions from the first study of Texts, Graphics, and Hyperlinks on web interfaces?

The questions obviously address the structural factors and elements of our metric model which were selected based on the results of the first experiment. Since we have we have four factors and three elements, and each factor will be implemented on three elements, there will be three hypotheses for each question. Likewise, each hypothesis addresses measures of elements' attributes that represent each factor as follows:

- H1. The Number of Different Font Sizes is the best indicator for measuring the complexity of the textual-size on web interfaces based on users' opinions.
- H2. The Number of Different Graphic Sizes is the best indicator for measuring the complexity of the graphical size on web interfaces based on users' opinions.
- H3. The Number of Wrapped Hyperlinks is the best indicator for measuring the complexity of hyperlink-size of web interfaces based on users' opinions.
- H4. The Number of Words is the best indicator for measuring the complexity of the textual density on web interfaces based on users' opinions.

- H5. The Number Graphics is the best indicator for measuring the complexity of the graphical-density on web interfaces based on users' opinions.
- H6. The Number of Hyperlinks is the best indicator for measuring the complexity of the hyperlink-density on web interfaces based on users' opinions.
- H7. The Number of Text Areas Grouped by Colors is the best indicator for measuring the complexity of the textual-grouping on web interfaces based on users' opinions.
- H8. The Number of Graphics Grouped as Lists is the best indicator for the measuring the complexity of the graphical-grouping on web interfaces based on users' opinions.
- H9. The Number of Links Grouped by Colors is the best indicator for measuring the complexity of the textual-grouping on web interfaces based on users' opinions.
- H10. The Number of Text Areas Vertically and Horizontally Aligned is the best indicator for measuring the complexity of the textual-alignment on web interfaces based on users' opinions.
- H11. The Number of Graphics Aligned to Right is the best indicator for the measuring the complexity of the graphical alignment on web interfaces based on users' opinions.
- H12. The Number of Links Aligned to Right is the best indicator for measuring the complexity of the textual-grouping on web interfaces based on users' opinions.

These main hypotheses will be broken down into some hypotheses in order to exam the strength of each one. The essential impression of approaching the hypotheses by this method is to simplify the work and make much more measurable. The results of this experiment allow us to ask the following question:

- Do the twelve measures of four structural factors: size, density, grouping, and alignment work in conjunction as a better predictor than each individually for web interface complexity?

To answer this question, we formulated the following hypothesis:

- H1. The users' perceptions of the relative complexity of different websites using screen models are explained by conjunction of the twelve measures of the four main structural factors for major three objects categories

Accordingly, we can calculate the weighted overall complexity of the web interfaces used in the experiments by utilizing the following formulas:

Size Complexity of Web Interface (SCWI) = Size ($Text$)* Weight+ Size ($Graphics$) * Weight + Size ($Links$) * Weight

Density Complexity of Web Interface (DCWI) = Density ($Text$) * Weight + Density ($Graphics$) * Weight + Density ($Links$) * Weight

Grouping Complexity of Web Interface (GCWI) = Grouping ($Text$) * Weight + Grouping ($Graphics$) * Weight + Grouping ($Links$) * Weight

Alignment Complexity of Web Interface (ACWI) = Alignment ($Text$) * Weight + Alignment ($Graphics$) * Weight + Alignment ($Links$) * Weight

Weighted Web Interface Complexity (WIC)= SCWI+ DCWI+ GCWI+ ACWI

Data collection and analysis

As explained before, the second experiment has been conducted twice. The first time was for the development purposes and the second time for the evaluation purposes. The goals of conducting the experiment for the first time were:

- Collecting and analyzing the sets of elements' attributes that can represent by complexity factors. For instance, the size of texts, most users believe that in terms of size complexity, texts can be represented by a total number of font sizes exist on a page rather than the total number of complex font families or styles.
- Reducing the number of elements attributes for each complexity factor to one or two at maximum.
- Using the regular model screen for further analysis and comparative study
- Formulating the equations of the complexity factors

On the second execution of the second experiment, which is the evaluation experiment, the metric model was refined, reduced and the complexity factors of the element's attributes were calculated. Consequently, the data collection and analysis of the development experiment were exclusively performed on the users' data entered into the surveys, and initially, no metric data existed as explained above. There are two phases in analyzing the data of the evaluation experiment, the first one is collecting the data produced by the calculation of the metrics and the second one is users' data. In short, these are the data that will be collected and analyzed in the second experiment:

- Users' views data of development experiment
- Users' views data of the evaluation experiment
- Metric's data of the calculations

Two essential surveys were presented to users, the firs one is about evaluating the model screen of the five selected websites against the real screens using the four factors of complexity on five Likert-scale. The second one is about evaluating our modified model screen in which there are three screenshots each one titled with a structural element and each one has correspondent elements that are highlighted and identified. Additionally, each structural element is measured according to an attribute reflects that factor of the four, and users rates them using the five Likert-scale.

Model screens versus real screen

First, we collect and analyze the data of the development experiment. The data that we collected represents ratings on a scale of five for each factor. The null hypothesis that should be tested is that the variances of the means of ratings for these four factors are equal, H_0. Also, it includes that there is no correlation between these factors. We implemented the ANOVA test to obtain the p-vale which gives us a reading to judge the hypothesis. In the following table, the p-value equals 5.17E-09 which is a very small value compared to the alpha value, 0.05. Consequently, there is enough evidence to show a huge difference between the means of factors, thus, the null hypothesis is rejected. Moreover, the correlation Table, display a moderate correlation between the size and density factors, which its value is 0.5. However, the grouping and alignment factors have a low correlation values with the size factor, which are 0.34 and 0.36. As well as, the density factor and the alignment factor with only 0.35 correlation value. In contrast, a strong correlation between the grouping and the alignment factors, which equals to 0.64. Overall, these numbers collectively show the association between these factors, therefore, we can infer that the factors are reasonably correlated. The ratings data are included in the appendixes section of this document.

Moving to the data collection and analysis of the evaluation experiment, the differences between the two experiments was explained in the empirical study chapter. We apply the same tests and data collection to develop the same null hypothesis which in its summary, the means of the factors' ratings are equal and there is a correlation between the factors based on the ratings. However, the results of ANOVA test are opposite in this experiment, whereas, the p-value equals 0.069 and it is slightly greater that the alpha value, 0.05.

Table 18. ANOVA test results of users' ratings of the four factors using the regular model screens

DE		EE	
Source of Variation	Between Groups	Source of Variation	Between Groups
F Statistical	14.0809	F Statistical	2.365261
p-value	5.17E-09	p-value	0.069764
F Critical	2.612877	F Critical	2.616437

In addition, the results of the correlation in this experiment are also varied with development experiment. Furthermore, the correlation results in this one are much higher and greater than the previous one. Whereas, the density, grouping, and alignment factors have above average correlation values with the size factor, which is 0.53-0.59. Also, the density and the and grouping factors have correlation value located in the same range, which is 0.59. Moreover, the alignment factor and the grouping factor repeatedly have the highest correlation value among the others, the same as the development experiment, which is 0.69. Nonetheless, the density and the alignment factors have a below average value for their correlation, which is 0.46.

Table 19. Correlations results of users' ratings of the four factors using the regular model screens

	DE				EE			
	Size	Density	Grouping	Alignment	Size	Density	Grouping	Alignment
Size	1.00				1.00			
Density	0.50	1.00			0.59	1.00		
Grouping	0.35	0.36	1.00		0.58	0.59	1.00	
Alignment	0.36	0.34	0.64	1.00	0.53	0.46	0.69	1.00

Our overall justification of these diverse results is that:

- The changes made on the components of the experiments such as:
    - The terminology explanations with visual aids
    - The technique of moderating the sessions, Retrospective Probing (RP)
    - Using highly ranked websites by independent source
- The results of the development experiment indicate that all factors have the same level of significance, on the other hand, the evaluation experiment indicate that some factors have higher level of significance than the others, and the difference in both readings lead us to estimate that the regular model screens are not adequate to judge the complexity of web interfaces. Therefore, we were not sure since the beginning about the adequacy of the model screens to be used as they are.
- The results of the first surveys in the first experiment adds more support to our argument because they show that during the development and the evaluation experiments the users’ ratings means are not equal and the factors have different levels of importance and their p-values back that 0.6 and 0.7 are respectively higher than the alpha value.

Complexity of elements' attributes

In the development experiment, we wanted to explore what attributes or characteristics of structural elements within the frame of each factor that can affect the complexity of the web interfaces. Hence, we presented sets of attributes to the users, and then we allowed them to rate them on a scale of five. Moreover, the rating procedure had our modified model screens as we demonstrated in the previous chapters. Only the highly rated attributes of complexity metrics were calculated. Afterward, two sets of values for the complexity were computed, the first one came from the total ratings and the second one came from the calculations of the complexity metrics. By implementing a statistical comparison using a t-test, we could observe the strength of the relationship between the sets of values which were given by the complexity metric and the values of the users' ratings. Subsequently, we could drive a hypothesis which is the means of the complexity of elemental attributes of users' ratings are equal to the calculated means of our metric model for the same attributes, H_0. The evaluation experiment was conducted in the same manner to confirm the numbers of the metric model, the users' ratings and the validity of the hypothesis. All actual data produced by the users' ratings or by the calculation of the metrics are available in the appendixes of this document.

As shown in the tables, in the development experiment after implementing the t-test on the means of the users' ratings and the means of the calculated metrics, we obtained a p-value equals to 0.44 which is greater than the alpha value, 0.05. That indicates the divergences between the means of the two methods of evaluating the complexity of the elements attributes complexity are very limited. Consequently, there is no reasonable significance to reject the null hypothesis, and it must be accepted. Similarly, the two sets of data that came from the evaluation experiment lead to a comparable conclusion, in which the p-value equals to 0.07 which is greater than the alpha value, 0.05. Hence, the outcomes of the two experiments suggest that metric model is very close to predicting the complexity of the web interfaces, we proposed to take it a step further and calculate the overall complexity of the web interfaces.

Table 20. T-test: the means of the users' ratings and the metric values

Statistical Measure	DE		EE	
	Metric Means	Ratings' Means	Metric Means	Ratings' Means
Mean	0.36	0.35	0.3392	0.2431
Variance	0.033	0.35	0.0366	0.0092
t Stat	0.142089		1.5569	
P(T<=t) one-tail	0.4442		0.0695	
t Critical one-tail	1.72472		1.7459	
P(T<=t) two-tail	0.88843		0.1390	
t Critical two-tail	2.08596		2.1199	

To understand the users' views more, we implemented the Principal Component Analysis (PCA) on their ratings, especially, the highly rated attributes. The result of the PCA in the development experiment is that four attributes have a stander deviation greater and equal to one and they represent a cumulative proportion of 59 percent, which means they have a better representation of the complexity over the other. In other words, four major components exemplify the attributes of complexity. However, the attributes that have heavier loadings on the first three components, somewhat have values less than the average, which means the weight of these attributes on the first three components is somehow equally distributed on them. The size of links has the highest loading among the other factors on the fourth component with 0.53. Likewise, in the evaluation experiment, the weight of these attributes on the first three components is somehow equitably assigned on them. The density of texts has the highest loading among the other factors on the fourth component with 0.51. The following figures demonstrate the PCA loadings of the users' ratings for the twelve factors of complexity on the first four components.

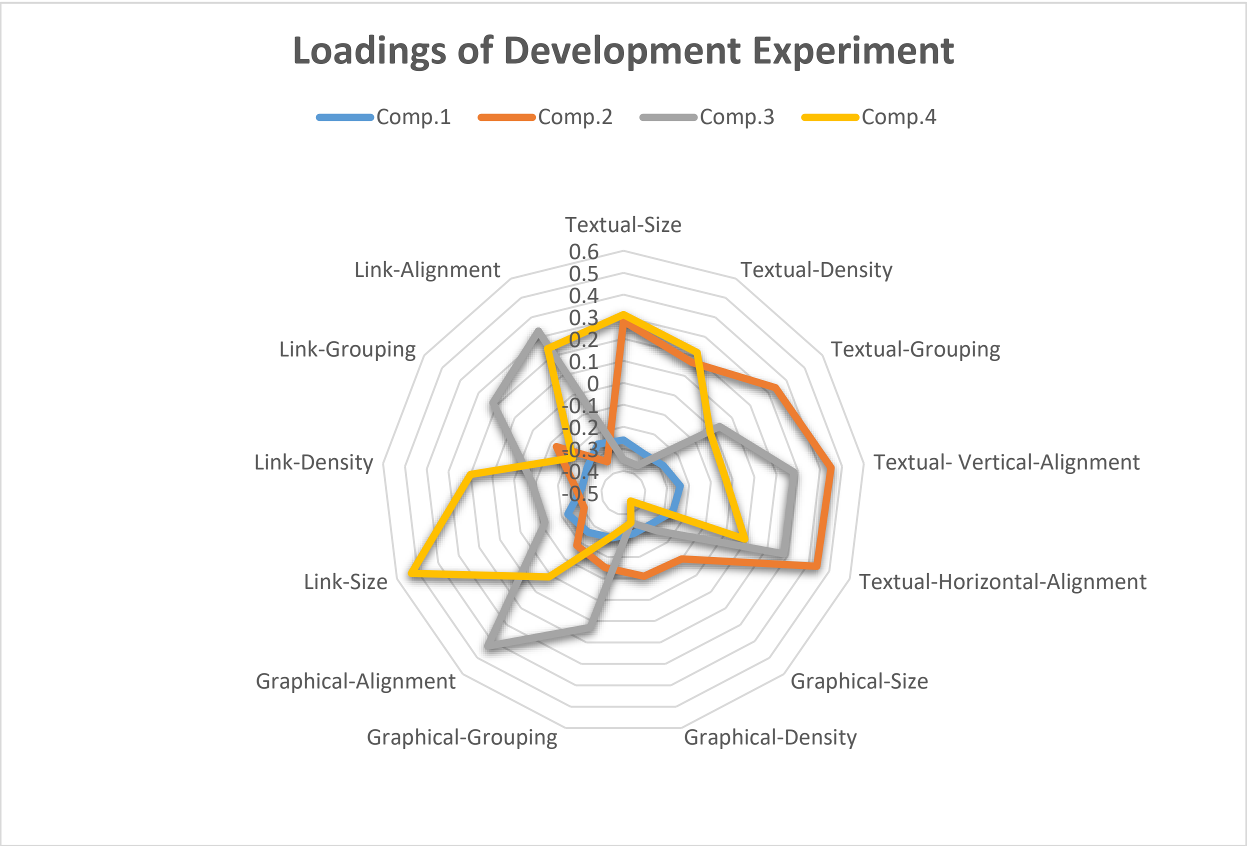


Figure 30. PCA - Loadings of users' ratings for the twelve factors using the modified model screens of the development experiment

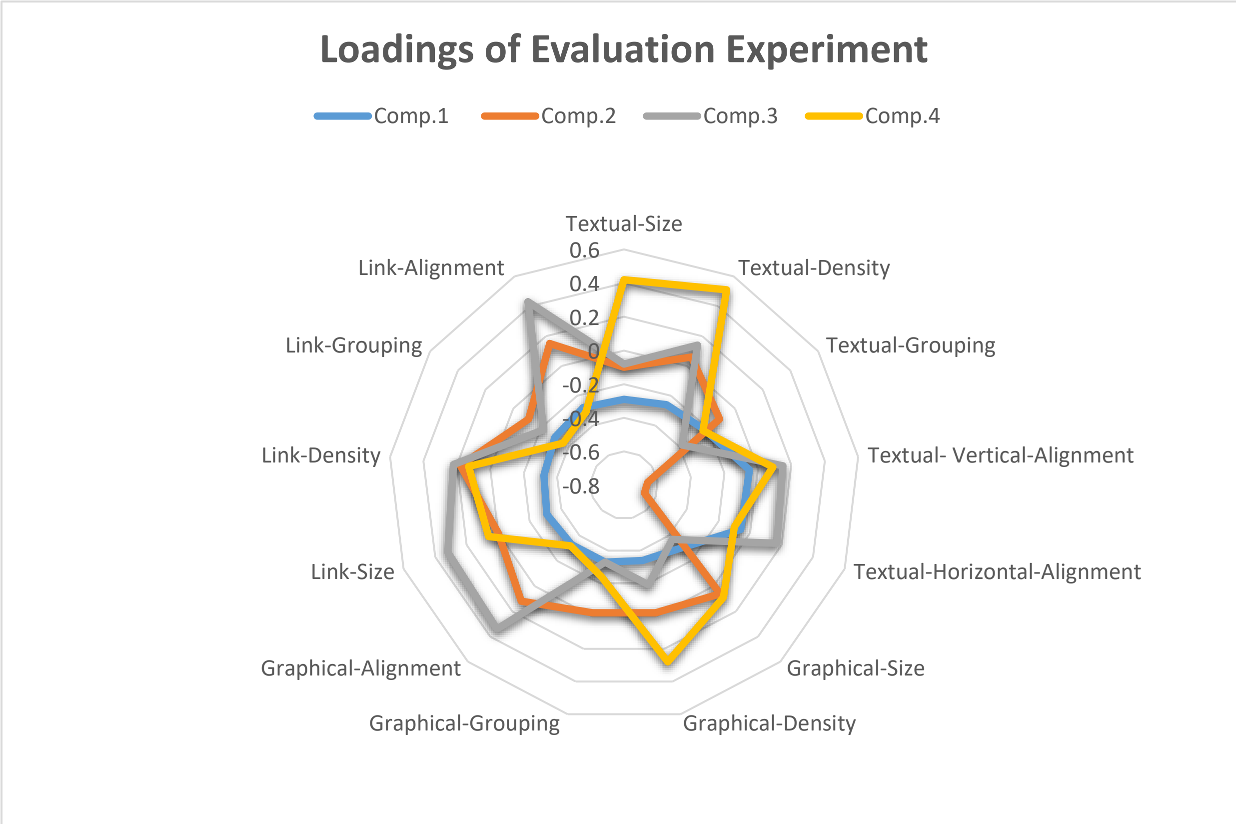


Figure 31. PCA - Loadings of users' ratings for the twelve factors using the modified model screens of the evaluation experiment

Calculating the overall weighted complexity of web interfaces

In this section, we calculated the overall weighted complexity of web interfaces for the ten homepages of the websites used in the development and evaluation experiment. The reason for performing this calculation is to answer and validate the last question and hypothesis mentioned in the hypotheses section of the second experiment. In addition, we use the equations and the results of the calculations of the attributes that done previously in this chapter, which can be found in the appendixes of this document. The following tables show the results of calculation. Interpreting the results of the calculation lead us to carry out the PCA on each factor for both sets of results mutually. Although it appears that the size and density factors have the highest averages respectively, and the grouping and alignment factors interchange their ranks in the different sets of results, the PCA gives distinctive and more understandable readings. By combining the two sets of results and applying the PCA which deliver a stander deviation of 1.6 for one component and the rest were less than one with 0.71 cumulative proportion. Where density and size have 0.57 and 0.55 loadings on the first and the most significant component. These values are somewhat above the 0.5 logical average of loadings, and the alignment and grouping have 0.44 and 0.41 loadings on the same component. The description of these number is that the density and size are most influential factors on the complexity of the web interfaces, and the alignment and grouping factors have the easier influence than the other two on the complexity of the web interfaces.

Table 21. Calculation the overall weighted complexity of the development experiment

Homepages	Weighted-Size Complexity	Weighted-Density Complexity	Weighted-Grouping Complexity	Weighted-Alignment Complexity	Overall Weighted Complexity
forbes.com	2.24	1.31	0.32	0.47	4.34
msn.com	3.86	1.97	0.75	0.84	7.43
graceLand.com	2.37	1.93	1.20	0.51	6.01
dujour.com	1.99	0.89	0.16	0.51	3.55
psu.edu	1.52	1.04	0.49	0.70	3.75
Average	2.40	1.43	0.58	0.61	25.07

Table 22. Calculation the overall weighted complexity of the evaluation experiment

Homepages	Weighted-Size Complexity	Weighted-Density Complexity	Weighted-Grouping Complexity	Weighted-Alignment Complexity	Overall Weighted Complexity
thoughteconomics.com	1.44	1.00	0.23	0.18	2.85
theguardian.com	2.31	1.68	0.95	0.26	5.19
www.guggenheim.org	1.71	1.08	0.79	0.23	3.81
www.hampshire.edu	1.05	0.61	0.50	0.12	2.28
www.oracle.com	0.35	0.45	0.27	0.17	1.25
Average	1.37	0.96	0.55	0.19	15.39

Table 23. PCA - Importance of components for the means of the collective data for each factor

Collective PCA for the data of development and evaluation experiments				
Importance of components	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	1.69	0.94	0.48	0.21
Proportion of Variance	0.71	0.22	0.06	0.01
Cumulative Proportion	0.71	0.93	0.99	1.00

Table 24. PCA - Loadings for the means of the collective data of each factor

Collective PCA loadings for the data of development and evaluation experiments				
Loadings	Comp.1	Comp.2	Comp.3	Comp.4
Size Complexity	-0.554	0.209	-0.57	0.57
Density Complexity	-0.573	-0.167	-0.262	-0.759
Grouping Complexity	-0.412	-0.727	0.45	0.316
Alignment Complexity	-0.442	0.633	0.635	

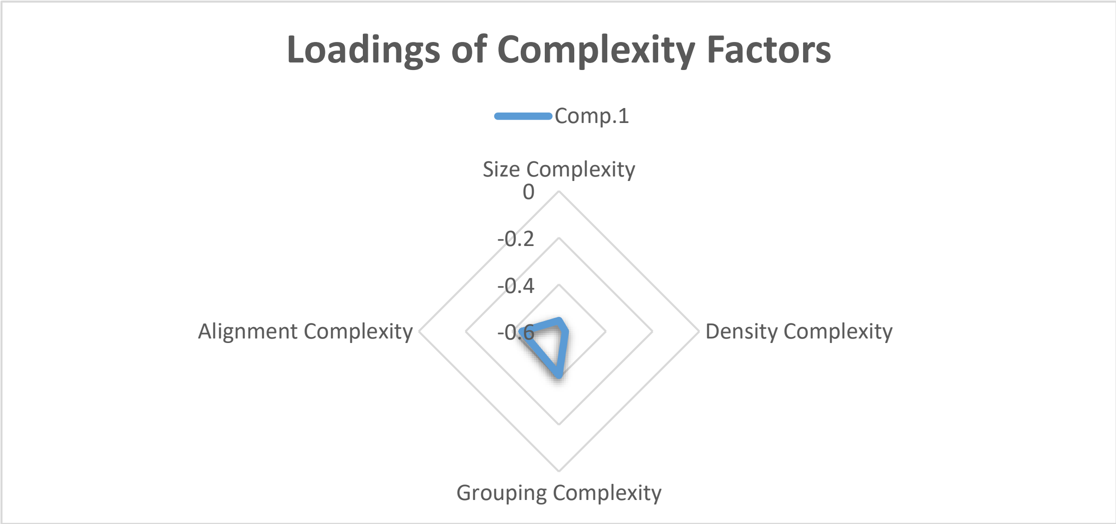


Figure 32. Radar chart of the loadings for the means of the collective data of each factor

CHAPTER 6. CONCLUSION AND FUTURE WORK

While calculating, usability can cost four times as much as directing qualitative studies; metrics are occasionally valuable. Between other things, metrics can aid managers design progress and support decisions about when to release a product. Metrics are the indications that show whether designing approaches are valid. Using metrics is important to chasing changes over time against repetitions and setting goals.

This thesis addresses some of these issues by examining methods to reduce the probability of designing inefficient web interfaces, and consequently, the software design cost reduces as well. In this chapter, we first present a summary of our contribution that is archived by this dissertation research.

Contribution

In this dissertation, we developed a metric model to compute the complexity of web interfaces which results in elevating the design quality of software. This model consists of developing measures of software usability complexity that can be utilized to target fundament components of GUI.

The starting point of our work is to construct surveys to discover the boundaries of the web interface complexity. These surveys were utilized to build two major experiments, and the first one led to the development of the exploratory structural metric model of complexity, which includes:

- Four structural factors of complexity: size, density, grouping, and alignment
- Three structural elements of complexity: texts, graphics, and links
- Model screens and real screen of web interfaces

- Attributes and characters of structural elements

Two key experiments were conducted, and each one conducted twice. The first set of experiments were investigative to specify and outline the metric model. The second set of experiments were confirmative to validate the metric model results. Also, a questioner by which we could collect demographic information of the study's participants. The survey of the first experiment aims to explore the concept of web interface complexity from users' point of view. Thus, we utilized two different types of questions, which are open-ended and close-ended questions. In the open-ended questions, we allowed the participants to type in five factors and five elements of complexity from their perspectives. While in the second type, we presented lists of factors and elements which were reported in previous studies causing the complexity of GUI.

The use of two styles of questions in our surveys allowed us to obtain many accurate results of the complexity concept from users' perspectives where there was a high level of conformity between the results of the two styles. That drove us to draw a clear line around the factors and elements of complexity. Moreover, the form of model screens that exist in the literature seem to be not very positive in illustrating the screens' structure of web interfaces. Therefore, we invented our modified model screens, in which we draw boundaries around the targeted visual objects that occupy areas on the screen, and then we title them with a category name of elements. This version of model screens yielded decent results that came out from users' ratings and the metrics calculations.

The adoption of several different types of statistical tests for various kinds of data enabled us to drive conclusions that can be generalized. In the first experiment, users we able to browse and observe factors and elements that may cause less understandable GUI, and then type in their views. The type of data that was produced by this kind of evaluation is ordinal data, which were ranked and categorized to implement the Wilcoxon signed rank test. As explained before the experiment was conducted twice, and each occurrence the results were almost similar. The interpretation of the results led as to conclude that users perceive our metrics factors and elements have a higher influence on the usability of web interfaces. In addition, in the surveys of the first experiment, which had closed-ended questions by which users rated our metric factors and elements collectively against other factors and elements and individually among each other. The statistical test that we used to analyze the interval data of the ratings was the t-test, by which we could conclude that our metrics factors and elements possess higher impact on the usability of web interfaces as well. In addition, performing the PCA enabled us to recognize that the four-factor and the three elements have approximately the same level of importance.

In the second experiment, we applied the regular model screen, and our modified model screen on the websites and then the users evaluated the four factors and the twelve measures of elements' attributes, and we performed statistical tests to analyze the interval data of the ratings. First, we collected the data of the users' ratings for the four factors and using the regular model screen, and then we implemented the ANOVA test, and we calculated the correlation ship between the four factors. The results slightly from the development and evaluation experiments and in our analysis, we found that the variables and the changes that we did in the evaluation experiment led to the different outcomes, however, the results of the correlations showed most of the factors are connected, and some of them like grouping and alignment factors have strong correlations.

Additionally, we could calculate the overall unweighted and weighted complexity of the websites based on the four factors and the three elements. For both experiments, we compared the data of users' ratings of websites complexity to the data of the calculated complexity by the equations of our metric model using a t-test to observe the difference between the two methods of evaluation. The results showed our method is a very strong in predicting the level of complexity for each web interface where the means of the both methods were very close in the development and the evaluation experiments. Furthermore, we conducted PCA on the data to detect the factors of higher influence on the web interface complexity, and the results of this analysis show that size and density factors have this high rank of influence.

Limitations and Future Work

Two major factors were mentioned explicitly and implicitly in the first experiment at both phases, the development, and the evaluation phases, and they are the colors and the navigation system of web interfaces. The reason for excluding them from our model was that compactness and concentration of the model which is a very important feature of the model, and we wanted to keep for easier representation of web interface complexity. In addition, we wanted to investigate the repetitive claims in the literature about the four structural factors of GUI complexity. Furthermore, we wanted to examine our model without the existence of these two factors and with them to study the difference in future studies. Also, each factor of these can have a standalone and independent study due to the variety of variables and parameters that must take into consideration.

Therefore, we have a future to include the navigation and color as factors of complexity to our model, however, since our model specialized in the structural aspects of the factors. Thus, structural factors of the menus of websites will be studied like the type of menus: depth versus breadth, vertical versus side navigation bars, tabs navigation, breadcrumb navigation, and footer navigation. Similarly, the strategy of selecting a structural aspect of the color factor will be applied such as categorizing the colors into groups based on the hue, chrome, and value. Plus, the number of colors used on web interfaces and their effects on the other factors. These two factors will be added individually and collectively to the model, and then a cooperative study will be conducted to investigate the level of representation of web interface complexity.

The second significant limitation to this study is the disability to build painless tool because there is obvious difficulty in developing a software tool capable of recognizing the structural elements, which are the graphics, text, and links based on our criteria. It requires images processing studies and technologies which are out of the scope of our study. Some SEO tools allowed us to perform some calculations such as the word counts and the links counts, but the graphics need more advanced tools to detect them because it requires more than scanning the web pages' scripts to identify objects.

Nevertheless, there is some abstract model implemented in open source image recognition tools that can be exploited to develop a software tool to fits the criteria of this research. Such approach is also a plan to this research to expand the capabilities and the automation of the calculations of our metric model.

Moreover, we noticed from users comments that websites genres had a huge influence on their ratings, which is something this study could not address since the intention from the beginning was the factors and the elements regardless of the website genres. Also, many previous studies did not focus on this matter. Consequently, our focus was oriented towards incorporating different types of websites instead of one. Hence, another plan is held for performing the same experiments in this research on several genres independently.

REFERENCES

[1] W. Albert, T. Tullis, and D. Tedesco, Beyond the Usability Lab: Conducting Large-scale Online User Experience Studies. Morgan Kaufmann, 2009.

[2] N. Fenton and J. Bieman, Software Metrics: A Rigorous and Practical Approach, Third Edition. CRC Press, 2014.

[3] S. A. Gómez, A. Groza, and C. I. Chesñevar, "An Argumentative Approach to Assessing Safety in Medical Device Software Using Defeasible Logic Programming," in *International Conference on Advancements of Medicine and Health Care through Technology; 5th – 7th June 2014, Cluj-Napoca, Romania*, S. Vlad and R. V. Ciupa, Eds. Springer International Publishing, 2014, pp. 167–172.

[4] S. Hirsch, J. Fraser, and S. Beckman, "Adaptive Path Reports » Leveraging Business Value: How ROI Changes User Experience," *cybermanual.com*. [Online]. Available: http://www.cybermanual.com/adaptive-path-reports-leveraging-business-value-how-roi-changes-user-experience.html. [Accessed: 04-May-2017].

[5] "Why Software Fails - IEEE Spectrum." [Online]. Available: http://spectrum.ieee.org/computing/software/why-software-fails. [Accessed: 02-Feb-2017].

[6] M. Zen and J. Vanderdonckt, "Towards an evaluation of graphical user interfaces aesthetics based on metrics," in *2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS)*, 2014, pp. 1–12.

[7] B. Falah, "Test Case Selection Based on a Spectrum of Complexity Metrics," North Dakota State University, Fargo, ND, USA, 2011.

[8] V. Nguyen, S. Deeds-rubin, T. Tan, and B. Boehm, “A SLOC Counting Standard,” in *The 22nd International Annual Forum on COCOMO and Systems/Software Cost Modeling.*, 2007, vol. 2.

[9] M. Al-Kabi, “GUI Structural Metrics,” *Int. Arab J. Inf. Technol. IAJIT 8 2 124-129*.

[10] A. Parush, R. Nadir, and A. Shtub, “Evaluating the Layout of Graphical User Interface Screens: Validation of a Numerical Computerized Model,” *Int. J. Hum.-Comput. Interact.*, vol. 10, no. 4, pp. 343–360, Dec. 1998.

[11] G. Bonsiepe, “A Method of Quantifying Order in Typographic Design,” *J. Typogr. Res.*, no. 2.3, Jul. 1968.

[12] M. Twyman, “Typography without words,” *Visible Lang.*, vol. 15, no. 1, pp. 5–12, 1981.

[13] C. E. Shannon and W. Weaver, *The mathematical theory of communication*. Urbana: University of Illinois Press, 1963.

[14] T. S. Tullis, *A system for evaluating screen formats: Research and application, Advances in human-computer interaction*. Ablex Publishing Corp., Norwood, NJ, 1988.

[15] T. Miyoshi and A. Murata, “A method to evaluate properness of GUI design based on complexity indexes of size, local density, aliment, and grouping,” in *2001 IEEE International Conference on Systems, Man, and Cybernetics*, 2001, vol. 1, pp. 221–226 vol.1.

[16] A. Sears, “Layout appropriateness: a metric for evaluating user interface widget layout,” *IEEE Trans. Softw. Eng.*, vol. 19, no. 7, pp. 707–719, Jul. 1993.

[17] J. Xing, “Measures of information complexity and the implications for automation design,” DTIC Document, 2004.

[18] A. Sears, "AIDE: A tool to assist in the design and evaluation of user interfaces," *Interact. Syst. Res. Cent. Baltim.*, 1993.

[19] F. Fu, S.-Y. Chiu, and C. H. Su, "Measuring the Screen Complexity of Web Pages," in *Human Interface and the Management of Information. Interacting in Information Environments*, M. J. Smith and G. Salvendy, Eds. Springer Berlin Heidelberg, 2007, pp. 720–729.

[20] K. Alemerien and K. Magel, "GUIEvaluator: A Metric-tool for Evaluating the Complexity of Graphical User Interfaces.," in *SEKE*, 2014, pp. 13–18.

[21] M. Y. Ivory, "An Empirical Foundation for Automated Web Interface Evaluation," University of California, Berkeley, 2001.

[22] J. Nielsen, *Designing Web Usability: The Practice of Simplicity*. Thousand Oaks, CA, USA: New Riders Publishing, 1999.

[23] J. M. Spool, W. Schroeder, T. Scanlon, and C. Snyder, *Web Site Usability: A Designer's Guide*. User Interface Engineering, 1997.

[24] C. Schalles, *Usability Evaluation of Modeling Languages*, 2013 edition. Wiesbaden: Springer Gabler, 2012.

[25] L. Westfall, "12 Steps to Useful Software Metric Presented By Linda Westfall," *http://www.westfallteam.com*, 11-Jan-2013. [Online]. Available: http://www.westfallteam.com/node/114. [Accessed: 04-May-2017].

[26] P. Goodman, *The Practical Implementation of Software Metrics*. McGraw-Hill, Inc., 1993.

[27] M. S. Sumit Chhabra, "A Survey of Metrics for Assessing the Navigational Quality of a Website Based on the Structure of Website.," *Int. J. Comput. Sci. Its Appl.*, vol. 167.

[28] D. C. Ngo, *VISIT: visitor information system implementation tool*. Trinity College, 1995.

[29] D. C. L. Ngo, A. Samsudin, and R. Abdullah, "Aesthetic measures for assessing graphic screens," *J Inf Sci Eng*, vol. 16, no. 1, pp. 97–116, 2000.

[30] D. C. L. Ngo, L. S. Teo, and J. G. Byrne, "Modelling interface aesthetics," *Inf. Sci.*, vol. 152, pp. 25–46, Jun. 2003.

[31] R. S. Grabinger, "Computer screen designs: Viewer judgments," *Educ. Technol. Res. Dev.*, vol. 41, no. 2, pp. 35–73, 1993.

[32] "Quick Screen Capture tool. Take Screenshots, Edit and Share them Effectively! FREE downloads." [Online]. Available: http://screenshot-program.com/. [Accessed: 10-Feb-2017].

[33] "Instant Wireframe." [Online]. Available: https://chrome.google.com/webstore/detail/instant-wireframe/pmpmnmbgidlnoamkpmcikaejhmeldnha. [Accessed: 10-Feb-2017].

[34] "The Webby Awards." [Online]. Available: http://webbyawards.com/. [Accessed: 10-Feb-2017].

[35] "What is statistical analysis? - Definition from WhatIs.com." [Online]. Available: http://whatis.techtarget.com/definition/statistical-analysis. [Accessed: 13-Feb-2017].

[36] M. O. Lacort, *Descriptive and Inferential Statistics - Summaries of theory and Exercises solved*. Lulu.com, 2014.

[37] W. Albert and T. Tullis, *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Newnes, 2013.

[38] N. J. Salkind, *Statistics for People Who (Think They) Hate Statistics: Using Microsoft Excel 2016*. SAGE Publications, 2016.

[39] S. Somyurek, T. Guyer, and B. Atasoy, *The Effects of Individual Differences on Learner's Navigation in a Courseware*, vol. 7. 2008.

[40] D. R. Danielson, "Transitional volatility in web navigation," *Inf. Technol. Soc. 13Special Issue Web Navig.*, pp. 131–158, 2003.

[41] J. S. Ahuja and J. Webster, "Perceived disorientation: an examination of a new measure to assess web design effectiveness," *Interact. Comput.*, vol. 14, no. 1, pp. 15–29, Dec. 2001.

[42] H. G. Kang and P. H. Seong, "An information theory-based approach for quantitative evaluation of user interface complexity," *IEEE Trans. Nucl. Sci.*, vol. 45, no. 6, pp. 3165–3174, Dec. 1998.

[43] G. Geissler, G. Zinkhan, and R. T. Watson, "Web home page complexity and communication effectiveness," *J. Assoc. Inf. Syst.*, vol. 2, no. 1, p. 2, 2001.

[44] J. Kalbach and A. Gustafson, *Designing Web Navigation: Optimizing the User Experience*, 1st edition. Beijing ; Sebastopol: O'Reilly Media, 2007.

[45] D. C. L. Ngo and J. G. Byrne, "Aesthetic measures for screen design," in *Computer Human Interaction Conference, 1998. Proceedings. 1998 Australasian*, 1998, pp. 64–71.

[46] D. C. L. Ngo, L. S. Teo, and J. G. Byrne, "Formalising guidelines for the design of screen layouts," *Displays*, vol. 21, no. 1, pp. 3–15, Mar. 2000.

[47] C. Wei and F. Xing, "The comparison of user-centered design and goal-directed design," in *2010 IEEE 11th International Conference on Computer-Aided Industrial Design Conceptual Design (CAIDCD)*, 2010, vol. 1, pp. 359–360.

[48] P. Zaphiris, B. Shneiderman, and K. L. Norman, “Expandable indexes vs. sequential menus for searching hierarchies on the World Wide Web,” *Behav. Inf. Technol.*, vol. 21, no. 3, pp. 201–207, Jan. 2002.

[49] L. Sánchez-González, F. García, F. Ruiz, and M. Piattini, “Toward a quality framework for business process models,” *Int. J. Coop. Inf. Syst.*, vol. 22, no. 01, p. 1350003, Mar. 2013.

[50] A. Seffah, M. Donyaee, R. B. Kline, and H. K. Padda, “Usability measurement and metrics: A consolidated model,” *Softw. Qual. J.*, vol. 14, no. 2, pp. 159–178, 2006.

[51] B. Indurkhya, H. van Oostendorp, and P. van Schaik, “Cognitive modelling of web navigation,” *Behav. Inf. Technol.*, vol. 31, no. 1, pp. 1–2, Jan. 2012.

[52] J. Xing, “Measures of information complexity and the implications for automation design,” DTIC Document, 2004.

[53] M. C. Puerta Melguizo, U. Vidya, and H. van Oostendorp, “Seeking information online: the influence of menu type, navigation path complexity and spatial ability on information gathering tasks,” *Behav. Inf. Technol.*, vol. 31, no. 1, pp. 59–70, Jan. 2012.

[54] J. C. Speirs, “Website Adaptive Navigation Effects on User Experiences,” *scholarsarchive.byu.edu*, 2012.

[55] J. Sauro and J. R. Lewis, *Quantifying the User Experience: Practical Statistics for User Research*. Morgan Kaufmann, 2016.

[56] J. Nielsen and H. Loranger, *Prioritizing Web Usability*, 1 edition. Berkeley, Calif: New Riders, 2006.

[57] T. Comber and J. Maltby, “Investigating layout complexity,” *Des. Specif. Verification Interact. Syst. 96 Proc. Eurographics Workshop Springer Wien N. Y.*, Jan. 1996.

APPENDIX A. USERS' DATA AND STATISTICAL TABLES

Table A1. The significance level of the selected four factors of the metric against the other factors by using Wilcoxon test on frequencies of users' inputs (First Experiment)

Size		Density		Grouping		Alignment		Total of References for the 4 Structural Factors		Other Factors' References		Total of All Factors' Types of References	
DE	EE	DE	EE	DE	EE	DE	EE	DE	EE	DE	EE	DE	EE
8	13	24	13	20	8	10	5	62	39	10	15	72	54
9	6	20	12	14	15	8	8	51	41	18	15	69	56
10	6	15	12	10	12	8	8	43	38	27	15	70	53
4	5	15	13	8	12	6	4	33	34	27	16	60	50
6	4	12	11	12	11	12	11	42	37	22	18	64	55
37	34	86	61	64	58	44	36	231	189	104	79	335	268

The p-value of Wilcoxon signed rank test (Development Experiment)	0.01193
The p-value of Wilcoxon signed rank test (Evaluation Experiment)	0.01116

Table A2. Descriptive data analysis of the actual frequencies of complexity factors (First Experiment)

Descriptive Statistical Measures	Total of References for the 4 Structural Factors		Other Factors' References	
	DE	EE	DE	EE
Mean	46.2	37.8	20.8	15.8
Standard Error	4.872371086	1.15758369	3.184336666	0.583095
Median	43	38	22	15
Standard Deviation	10.89495296	2.588435821	7.120393248	1.30384
Sample Variance	118.7	6.7	50.7	1.7
Range	29	7	17	3
Minimum	33	34	10	15
Maximum	62	41	27	18
Sum	231	189	104	79
Confidence Level (95.0%)	13.52787085	3.213967571	8.841135949	1.618932

Table A3. Descriptive data analysis for the ratings of all factors against each other (First Experiment)

Descriptive Statistical Measures	Structural Four Factors		Structural Six Factors	
	DE	EE	DE	EE
Mean	3.566666667	3.60326087	3.066666667	3.260869565
Standard Error	0.083813124	0.083369803	0.064183079	0.068603364
Median	4	4	3	3
Standard Deviation	1.298427329	1.130883023	1.217788305	1.139724674
Sample Variance	1.685913529	1.278896412	1.483008357	1.298972332
Range	5	4	5	5
Minimum	0	1	0	0
Maximum	5	5	5	5
Sum	856	663	1104	900
Confidence Level (95.0%)	0.165106775	0.164489614	0.126222054	0.135054494

Table A4. Descriptive data analysis for the users' rating data of the four factors of complexity against each other (First Experiment)

Descriptive Statistical Measures	DE				EE			
	Size	Density	Grouping	Alignment	Size	Density	Grouping	Alignment
Mean	3.43	3.66	3.51	3.43	3.56	3.41	3.30	3.43
Standard Error	0.13	0.13	0.13	0.16	0.13	0.15	0.17	0.16
Median	3	4	4	3.5	4	3	3	4
Standard Deviation	1.01	1.05	1.03	1.25	0.93	1.06	1.20	1.08
Sample Variance	1.02	1.10	1.06	1.57	0.87	1.13	1.46	1.18
Range	4	4	4	4	3	4	4	4
Minimum	1	1	1	1	2	1	1	1
Maximum	5	5	5	5	5	5	5	5
Sum	206	220	211	206	164	157	152	158
Confidence Level (95.0%)	0.26	0.27	0.26	0.32	0.27	0.31	0.35	0.32

Table A5. The significance level of the selected three elements of the metric against the other elements by using Wilcoxon test on frequencies of users' inputs (First Experiment)

Text		Graphics		Links		Total of References for the 3 Structural Elements		Other Elements' References		Total of All Elements' Types of References	
DE	EE	DE	EE	DE	EE	DE	EE	DE	EE	DE	EE
16	2	33	37	11	24	60	63	12	6	72	69
11	4	32	20	14	14	57	38	12	17	69	55
19	7	25	23	14	17	58	47	19	14	77	61
11	11	19	5	9	8	39	24	18	17	57	41
3	5	22	12	16	10	41	27	19	12	60	39
60	29	131	97	64	73	255	199	80	66	335	265
The p-value of Wilcoxon signed rank test (Development Experiment)								0.01167			
The p-value of Wilcoxon signed rank test (Evaluation Experiment)								0.01193			

Table A6. Descriptive data analysis of the actual frequencies of complexity elements (First Experiment)

Descriptive Statistical Measures	Total of References for the 3 Structural Elements		Other Elements’ References	
	DE	EE	DE	EE
Mean	51	16	39.8	13.2
Standard Error	4.5276	1.6431	7.0950	2.0346
Median	57	18	38	14
Standard Deviation	10.1242	3.674	15.865	4.5497
Sample Variance	102.5	13.5	251.7	20.7
Range	21	7	39	11
Minimum	39	12	24	6
Maximum	60	19	63	17
Sum	255	80	199	66
Confidence Level (95.0%)	12.57088987	4.562165	19.69906879	5.64923

Table A7. Descriptive data analysis for the ratings of all elements against each other (First Experiment)

Descriptive Statistical Measures	Structural Three Elements		Structural Five Elements	
	DE	EE	DE	EE
Mean	3.494	3.239	2.95	2.891
Standard Error	0.0846	0.1100	0.0714	0.0865
Median	3	4	3	3
Standard Deviation	1.1359	1.293	1.237	1.312
Sample Variance	1.290	1.672	1.532	1.721
Range	4	4	4	4
Minimum	1	1	1	1
Maximum	5	5	5	5
Sum	629	447	885	665
Confidence Level (95.0%)	0.167	0.217	0.140	0.170

Table A8. Descriptive data analysis for the users' rating data of the three elements of complexity against each other (First Experiment)

Descriptive Statistical Measures	DE			EE		
	Texts	Graphics	Links	Texts	Graphics	Links
Mean	3.68	3.366	2.88	2.97	3.45	3.39
Standard Error	0.162	0.161	0.142	0.156	0.183	0.133
Median	4	3	3	3	3	3
Standard Deviation	1.255	1.248	1.106	1.064	1.241	0.906
Sample Variance	1.57	1.55	1.22	1.13285	1.542512	0.821256
Range	4	4	4	4	4	3
Minimum	1	1	1	1	1	2
Maximum	5	5	5	5	5	5
Sum	221	202	173	137	159	156
Confidence Level (95.0%)	0.32	0.32	0.28	0.31	0.36	0.26

Table A9. PCA - importance of components for users' ratings of twelve complexity items (First Experiment)

Importance of components	DE											
	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7	Comp. 8	Comp. 9	Comp. 10	Comp. 11	Comp. 12
Standard deviation	2.05	1.40	1.16	1.13	0.90	0.70	0.66	0.61	0.54	0.52	0.49	0.42
Proportion of Variance	0.35	0.16	0.11	0.10	0.06	0.04	0.03	0.03	0.02	0.02	0.02	0.01
Cumulative Proportion	0.35	0.51	0.62	0.73	0.80	0.84	0.88	0.91	0.94	0.96	0.98	1.00

Table A10. PCA – loadings for users' ratings of the twelve complexity items (First Experiment)

DE												
Loadings	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7	Comp. 8	Comp. 9	Comp. 10	Comp. 11	Comp. 12
Size of Texts	-0.273	-0.232	-0.203	0.404	0.397	0.427	0.120	-0.185	0.263	-0.196	0.396	0.121
Size of Link	-0.145	0.490	0.201		0.578	0.107		0.333			-0.400	0.260
Size of Graphics	-0.310	-0.209	0.361	-0.283	-0.117		-0.247	0.534	0.255	-0.407	0.239	
Density of Texts	-0.277	-0.183	-0.558		0.160		-0.237	0.391		0.381	-0.101	-0.426
Density of Links	-0.169	0.396	-0.308	-0.492			0.219		-0.120	0.176	0.506	0.340
Density of Graphics	-0.218	-0.259		-0.614	0.115	0.209	0.222	-0.376		-0.215	-0.445	
Grouping of Texts	-0.335	-0.062	-0.368	0.208	-0.051	-0.657	0.056	-0.044	-0.004	-0.289	-0.201	0.376
Grouping of Links	-0.235	0.548	-0.015	0.009	0.087	-0.114	-0.231	-0.292	-0.066	-0.393	0.131	-0.556
Grouping of Graphics	-0.320	-0.195	0.372	-0.097	0.284	-0.241	-0.407	-0.378	-0.178	0.429	0.162	0.149

Table A10. PCA – loadings for users' ratings of the twelve complexity items (First Experiment) (continued)

DE												
Loadings	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7	Comp. 8	Comp. 9	Comp. 10	Comp. 11	Comp. 12
Alignment of Texts	-0.387	-0.037	0.056	0.192	-0.279	0.399	-0.029	0.097	-0.718	-0.114	-0.158	0.093
Alignment of Links	-0.328	0.245	0.022	0.152	-0.522	0.219	-0.193	-0.172	0.533	0.280	-0.213	0.114
Alignment of Graphics	-0.359	-0.040	0.317	0.127	-0.067	-0.220	0.712	0.077	0.039	0.238	0.085	-0.348

Table A11. PCA - importance of components for users' ratings of twelve complexity items (First Experiment)

Importance of components	EE											
	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7	Comp. 8	Comp. 9	Comp. 10	Comp. 11	Comp. 12
Standard deviation	2.31	1.25	1.08	1.01	0.85	0.82	0.63	0.59	0.48	0.46	0.41	0.34
Proportion of Variance	0.444	0.130	0.097	0.084	0.060	0.056	0.033	0.029	0.019	0.017	0.014	0.009
Cumulative Proportion	0.444	0.575	0.673	0.758	0.818	0.875	0.908	0.938	0.957	0.975	0.990	1.00

Table A12. PCA - loadings for users' ratings of the twelve complexity items (First Experiment)

Loadings	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7	Comp. 8	Comp. 9	Comp. 10	Comp. 11	Comp. 12
Size of Texts		-0.324	0.538	0.507	-0.353	-0.030	0.298	0.108	-0.150	-0.285	-0.041	0.036
Size of Links	-0.257	0.045	0.415	-0.332	-0.193	0.604	0.095	-0.147	0.051	0.442	0.096	0.100
Size of Graphics	-0.331	0.067	0.069	0.175	0.307	0.407	-0.579	-0.044	-0.280	-0.363	-0.206	-0.016
Density of Texts	-0.286	-0.481	-0.190	0.042	-0.131	-0.173	-0.212	-0.238	-0.311	0.319	0.282	-0.471
Density of Links	-0.286	-0.345	-0.345	-0.346	-0.243	-0.054	0.027	0.112	-0.221	-0.168	-0.087	0.633
Density of Graphics	-0.297	-0.385	-0.227	0.130	0.233	0.249	0.152	0.052	0.709	-0.167	0.123	-0.081
Grouping of Texts	-0.285	0.019	0.440	-0.021	0.321	-0.527	-0.275	-0.068	0.204	0.137	0.273	0.358
Grouping of Links	-0.328	0.154	0.127	-0.370	-0.152	-0.156	-0.059	0.692	0.053	-0.119	-0.043	-0.408
Grouping of Graphics	-0.343	0.019	0.034	-0.124	0.455	-0.163	0.541	-0.214	-0.210	0.066	-0.475	-0.139

Table A12. PCA - loadings for users' ratings of the twelve complexity items (First Experiment) (continued)

Loadings	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7	Comp. 8	Comp. 9	Comp. 10	Comp. 11	Comp. 12
Alignment of Texts	-0.307	0.258	-0.164	0.293	-0.449	-0.166	-0.206	-0.161	0.320	0.259	-0.508	0.035
Alignment of Links	-0.294	0.445	-0.098	-0.099	-0.238	-0.062	0.192	-0.435	0.000	-0.474	0.419	-0.094
Alignment of Graphics	-0.265	0.318	-0.282	0.462	0.146	0.119	0.216	0.385	-0.238	0.323	0.329	0.189

Table A13. Users' ratings of Development Experiment using the modified model screens (Second Experiment)

Homepages	Size Complexity			Density Complexity			Grouping Complexity			Alignment Complexity		
	(TSC)	(GSC)	(LSC)	(TDC)	(GDC)	(LDC)	(TGC)	(GGC)	(LGC)	(TAC)	(GAC)	(LAC)
forbes.com	0.59	0.45	0.59	0.55	0.28	0.48	0.15	0.32	0.17	0.22	0.25	0.31
msn.com	0.57	0.59	0.58	0.69	0.64	0.64	0.36	0.37	0.31	0.16	0.3	0.15
graceLand.com	0.38	0.63	0.61	0.35	0.75	0.61	0.29	0.31	0.25	0.24	0.33	0.28
dujour.com	0.48	0.24	0.47	0.32	0.2	0.37	0.12	0.13	0.16	0.2	0.25	0.25
psu.edu	0.37	0.3	0.45	0.32	0.27	0.38	0.2	0.15	0.16	0.18	0.16	0.28
Average	0.47	0.44	0.54	0.44	0.42	0.49	0.22	0.25	0.21	0.2	0.25	0.25

Table A14. Users' ratings of evaluation experiment using the modified model screens (Second Experiment)

Rating Values of Evaluation Experiment												
Homepages	Size Complexity			Density Complexity			Grouping Complexity			Alignment Complexity		
	(TSC)	(GSC)	(LSC)	(TDC)	(GDC)	(LDC)	(TGC)	(GGC)	(LGC)	(TAC)	(GAC)	(LAC)
thougthteconomics.com	0.81	0.74	0.14	0.93	0.02	0.03	0.09	0.21	0.19	0.01	0.11	0.16
theguardian.com	0.72	0.80	0.42	0.06	0.09	0.83	0.25	0.81	0.42	0.06	0.06	0.13
www.guggenheim.org	0.63	0.24	0.57	0.61	0.19	0.27	0.5	0.24	0.75	0.03	0.08	0.16
www.hampshire.edu	0.56	0.22	0.62	0.44	0.20	0.34	0.65	0.21	0.69	0.08	0.07	0.09
www.oracle.com	0.6	0.24	0	0.25	0.13	0.47	0.54	0.33	0.63	0.07	0.18	0.13
Average	0.6	0.45	0.35	0.46	0.13	0.39	0.40	0.36	0.53	0.05	0.10	0.13

Table A15. Metric values of development experiment (Second Experiment)

Homepages	Size Complexity			Density Complexity			Grouping Complexity			Alignment Complexity		
	(TSC)	(GSC)	(LSC)	(TDC)	(GDC)	(LDC)	(TGC)	(GGC)	(LGC)	(TAC)	(GAC)	(LAC)
forbes.com	0.75	0.50	0.13	0.47	0.11	0.42	0.03	0.24	0.23	0.28	0.16	0.16
msn.com	1.00	0.89	0.33	0.01	0.49	0.49	0.33	0.13	0.27	0.33	0.51	0.54
graceLand.com	0.50	0.72	0.24	0.68	0.17	0.28	0.29	0.68	0.45	0.16	0.18	0.26
dujour.com	0.77	0.80	0.11	0.66	0.11	0.23	0.04	0.14	0.20	0.24	0.16	0.33
psu.edu	0.57	0.71	0.07	0.43	0.33	0.31	0.42	0.29	0.26	0.73	0.25	0.16
Average	0.72	0.72	0.18	0.45	0.24	0.35	0.22	0.29	0.28	0.35	0.25	0.29

Table A16. Metric values of evaluation experiment (Second Experiment)

Homepages	Size Complexity			Density Complexity			Grouping Complexity			Alignment Complexity		
	(TSC)	(GSC)	(LSC)	(TDC)	(GDC)	(LDC)	(TGC)	(GGC)	(LGC)	(TAC)	(GAC)	(LAC)
thoughteconomics.com	0.25	0.25	0.33	0.51	0.30	0.18	0.1	0.15	0.20	0.05	0.28	0.28
theguardian.com	0.32	0.36	0.5	0.39	0.72	0.55	0.18	0.27	0.18	0.13	0.37	0.42
www.guggenheim.org	0.41	0.29	0.47	0.38	0.28	0.34	0.18	0.12	0.21	0.12	0.26	0.44
www.hampshire.edu	0.23	0.16	0.34	0.21	0.26	0.13	0.02	0.16	0.13	0.05	0.21	0.18
www.oracle.com	0.13	0.07	0.21	0.15	0.15	0.20	0.07	0.05	0.05	0.05	0.15	0.22
Average	0.27	0.23	0.37	0.33	0.34	0.28	0.11	0.15	0.15	0.08	0.25	0.31

Table A17. PCA - importance of components for users' ratings of twelve complexity items using the modified model screens of the development experiment (Second Experiment)

Importance of components	Comp .1	Comp .2	Comp .3	Comp .4	Comp .5	Comp .6	Comp .7	Comp .8	Comp .9	Comp .10	Comp .11	Comp .12	Comp .13
Standard deviation	2.01	1.15	1.13	1.00	0.92	0.85	0.83	0.79	0.77	0.75	0.70	0.67	0.60
Proportion of Variance	0.31	0.10	0.10	0.08	0.07	0.06	0.05	0.05	0.05	0.04	0.04	0.03	0.03
Cumulative Proportion	0.31	0.41	0.51	0.59	0.65	0.71	0.76	0.81	0.86	0.90	0.94	0.97	1.00

Table A18. PCA - importance of components for users' ratings of twelve complexity items using the modified model screens of the evaluation experiment (Second Experiment)

Importance of components	Comp .1	Comp .2	Comp .3	Comp .4	Comp .5	Comp .6	Comp .7	Comp .8	Comp .9	Comp .10	Comp .11	Comp .12	Comp .13
Standard deviation	2.14	1.24	1.12	1.04	0.92	0.85	0.79	0.73	0.70	0.66	0.58	0.56	0.50
Proportion of Variance	0.35	0.12	0.10	0.08	0.07	0.06	0.05	0.04	0.04	0.03	0.03	0.02	0.02
Cumulative Proportion	0.35	0.47	0.57	0.65	0.71	0.77	0.82	0.86	0.90	0.93	0.96	0.98	1.00

Table A19. PCA - loadings of users' ratings of twelve complexity items using the modified model screens of the development experiment (Second Experiment)

Loadings	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7	Comp. 8	Comp. 9	Comp. 10	Comp. 11	Comp. 12	Comp. 13
Textual-Size	-0.26	0.28	-0.35	0.31	0.39	0.13	0.22	-0.02	-0.04	-0.29	0.25	0.51	-0.12
Textual-Density	-0.30	0.17	-0.36	0.22	0.15	-0.04	-0.20	0.08	0.61	0.15	-0.02	-0.49	0.02
Textual-Grouping	-0.28	0.34	0.03	-0.02	0.25	-0.63	0.02	0.01	-0.49	0.20	-0.12	-0.18	0.12
Textual-Vertical-Alignment	-0.24	0.45	0.28	-0.03	-0.34	0.28	-0.10	0.36	-0.18	-0.09	0.42	-0.23	-0.25
Textual-Horizontal-Alignment	-0.26	0.44	0.28	0.09	-0.26	0.21	0.06	-0.30	0.18	0.02	-0.55	0.22	0.25
Graphical-Size	-0.31	-0.10	-0.27	-0.45	0.11	0.24	-0.22	0.04	-0.19	0.02	-0.44	0.03	-0.51
Graphical-Density	-0.31	-0.11	-0.36	-0.36	-0.18	0.21	-0.08	-0.05	-0.16	0.17	0.28	0.08	0.63
Graphical-Grouping	-0.29	-0.15	0.13	-0.30	-0.01	-0.11	0.78	0.30	0.27	0.02	-0.03	-0.03	-0.04

Table A19. PCA - loadings of users' ratings of twelve complexity items using the modified model screens of the development experiment (Second Experiment) (continued)

Loadings	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7	Comp. 8	Comp. 9	Comp. 10	Comp. 11	Comp. 12	Comp. 13
Graphical-Alignment	-0.26	-0.18	0.43	0.01	0.39	0.23	-0.01	-0.44	0.03	0.46	0.29	-0.03	-0.15
Link-Size	-0.23	-0.31	-0.12	0.53	-0.41	-0.09	0.00	0.22	-0.13	0.46	-0.08	0.25	-0.17
Link-Density	-0.31	-0.28	-0.08	0.20	-0.27	-0.06	0.17	-0.50	-0.19	-0.48	0.04	-0.37	-0.10
Link-Grouping	-0.30	-0.13	0.22	-0.22	-0.13	-0.49	-0.41	-0.02	0.35	-0.25	0.20	0.38	-0.08
Link-Alignment	-0.25	-0.34	0.33	0.24	0.34	0.20	-0.20	0.43	-0.13	-0.31	-0.22	-0.08	0.34

Table A20. PCA - loadings of users' ratings of twelve complexity items using the modified model screens of the evaluation experiment (Second Experiment)

Loadings	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7	Comp. 8	Comp. 9	Comp. 10	Comp. 11	Comp. 12	Comp. 13
Textual-Size	-0.29	-0.10	-0.08	0.42	0.07	-0.24	0.58	-0.19	0.35	-0.02	0.25	-0.30	0.09
Textual-Density	-0.26	0.06	0.14	0.51	0.50	-0.05	-0.24	-0.29	-0.22	0.05	-0.43	0.14	-0.09
Textual-Grouping	-0.22	-0.11	-0.38	-0.23	0.67	0.06	-0.09	0.41	-0.03	-0.19	0.22	-0.16	-0.08
Textual-Vertical-Alignment	-0.05	-0.66	0.15	0.09	-0.16	-0.34	-0.03	0.14	-0.51	0.10	0.17	-0.05	-0.24
Textual-Horizontal-Alignment	-0.06	-0.67	0.17	-0.10	0.02	0.29	-0.18	-0.07	0.49	-0.12	-0.31	-0.07	0.18
Graphical-Size	-0.30	0.06	-0.37	0.09	-0.38	-0.34	-0.02	0.37	0.07	-0.32	-0.50	0.07	-0.02

Table A20. PCA - loadings of users' ratings of twelve complexity items using the modified model screens of the evaluation experiment (Second Experiment) (continued)

Loadings	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7	Comp. 8	Comp. 9	Comp. 10	Comp. 11	Comp. 12	Comp. 13
Graphical-Density	-0.34	-0.02	-0.19	0.28	-0.21	0.07	-0.51	-0.03	0.12	0.08	0.51	0.34	0.26
Graphical-Grouping	-0.33	-0.02	-0.33	-0.24	-0.10	0.10	0.00	-0.14	0.09	0.74	-0.17	-0.13	-0.29
Graphical-Alignment	-0.33	0.12	0.34	-0.32	0.08	-0.31	-0.02	0.10	-0.15	0.21	-0.08	-0.16	0.67
Link-Size	-0.31	0.00	0.32	0.06	0.00	0.40	0.44	0.42	-0.04	0.10	-0.01	0.49	-0.08
Link-Density	-0.32	0.18	0.22	0.13	-0.26	0.44	-0.16	0.06	-0.20	-0.22	0.04	-0.63	-0.14
Link-Grouping	-0.30	-0.11	-0.22	-0.36	-0.07	0.15	0.23	-0.57	-0.34	-0.38	0.01	0.22	0.08
Link-Alignment	-0.28	0.15	0.43	-0.31	0.03	-0.35	-0.17	-0.12	0.35	-0.18	0.16	0.12	-0.50

APPENDIX B. REAL AND MODEL SCREENS OF THE HOME WEBPAGES

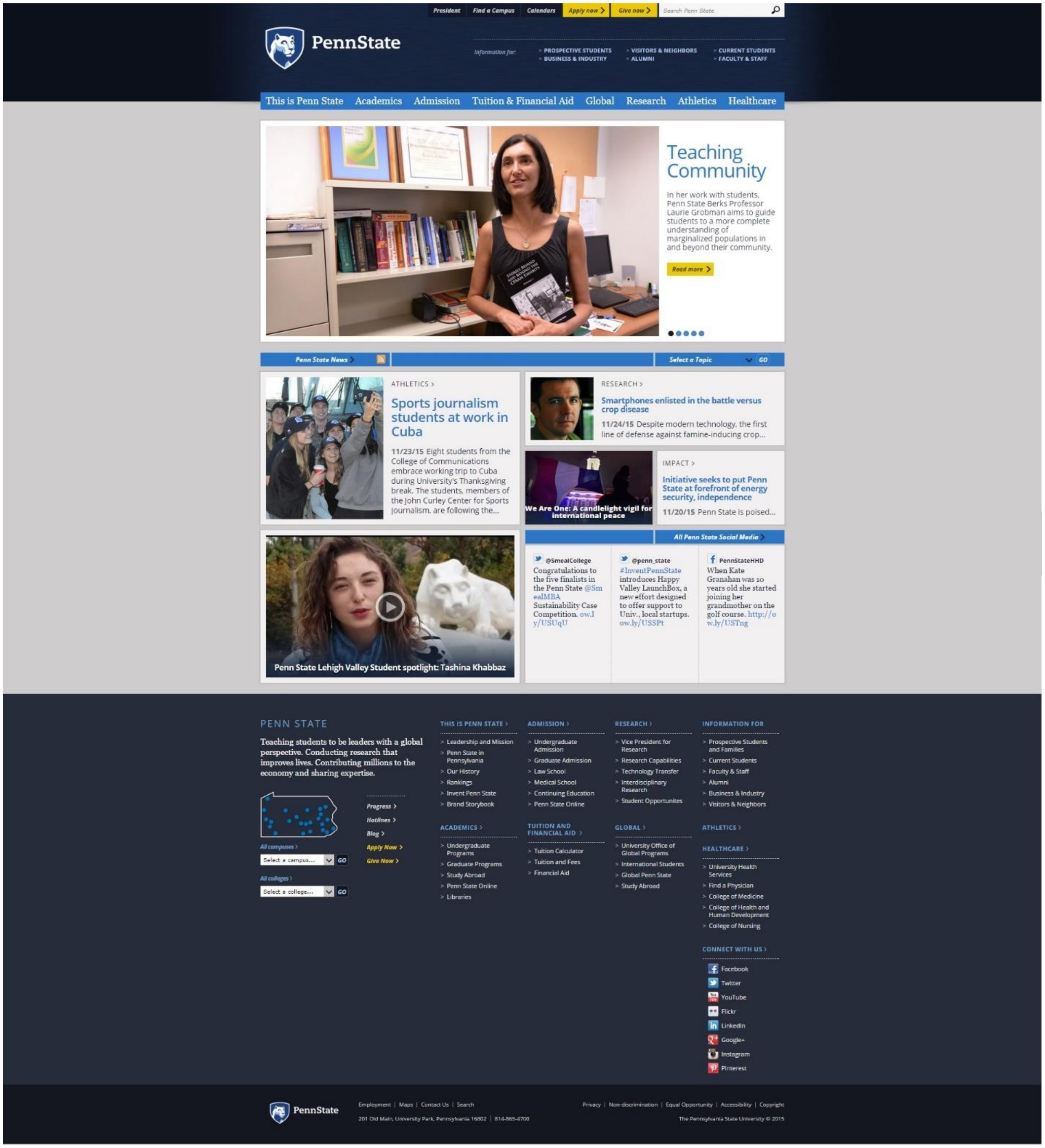


Figure B1. Homepage of www.psu.edu website

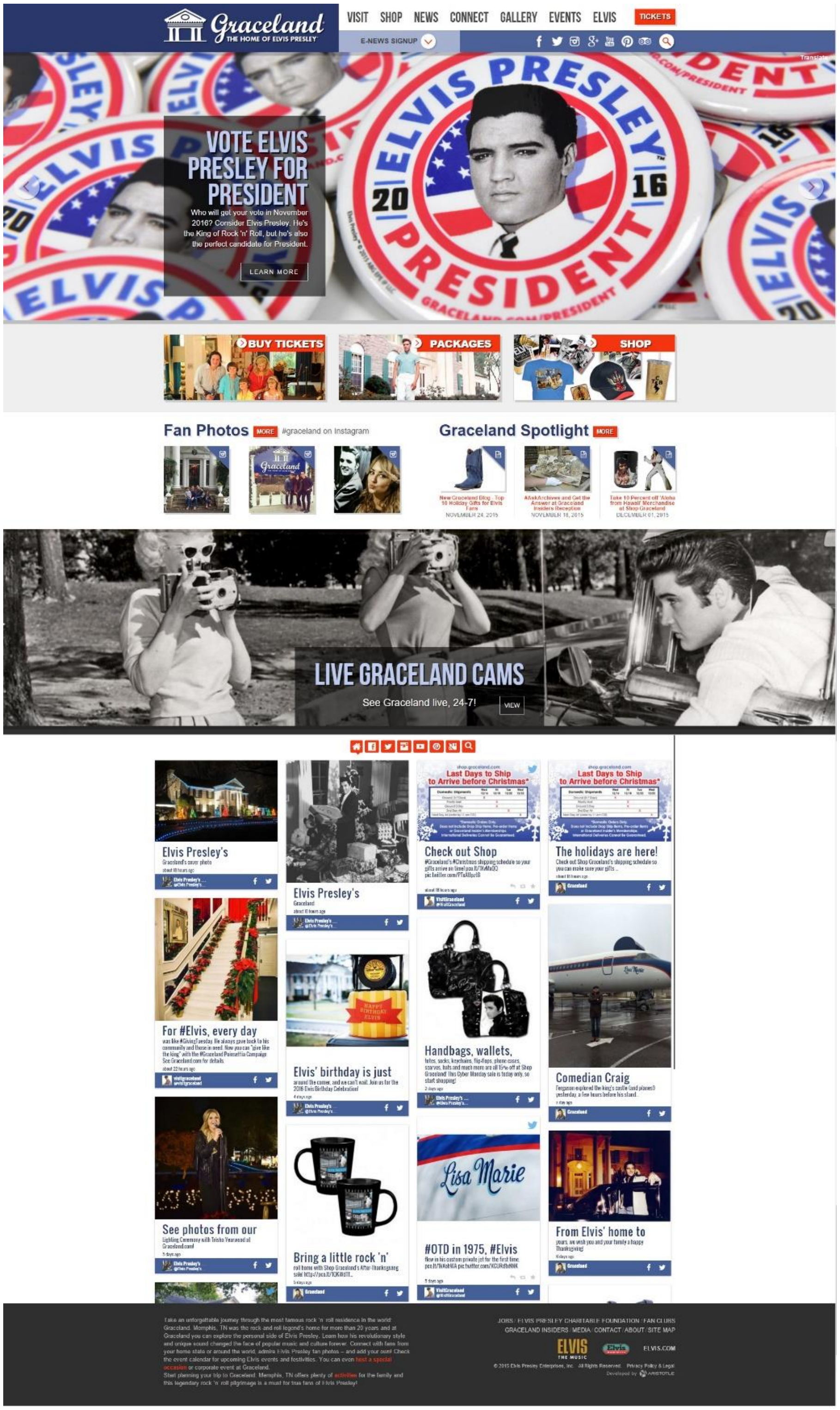


Figure B2. Homepage of www.graceland.com website

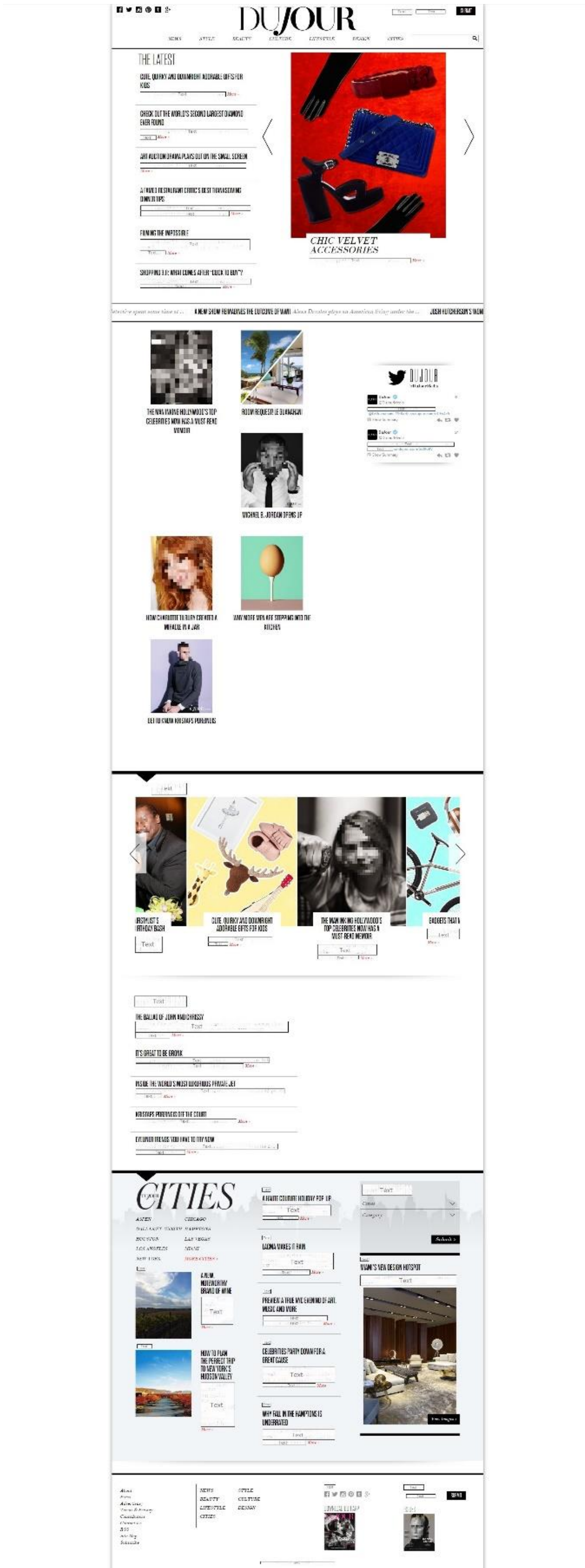


Figure B3. Homepage of www.dujour.com website

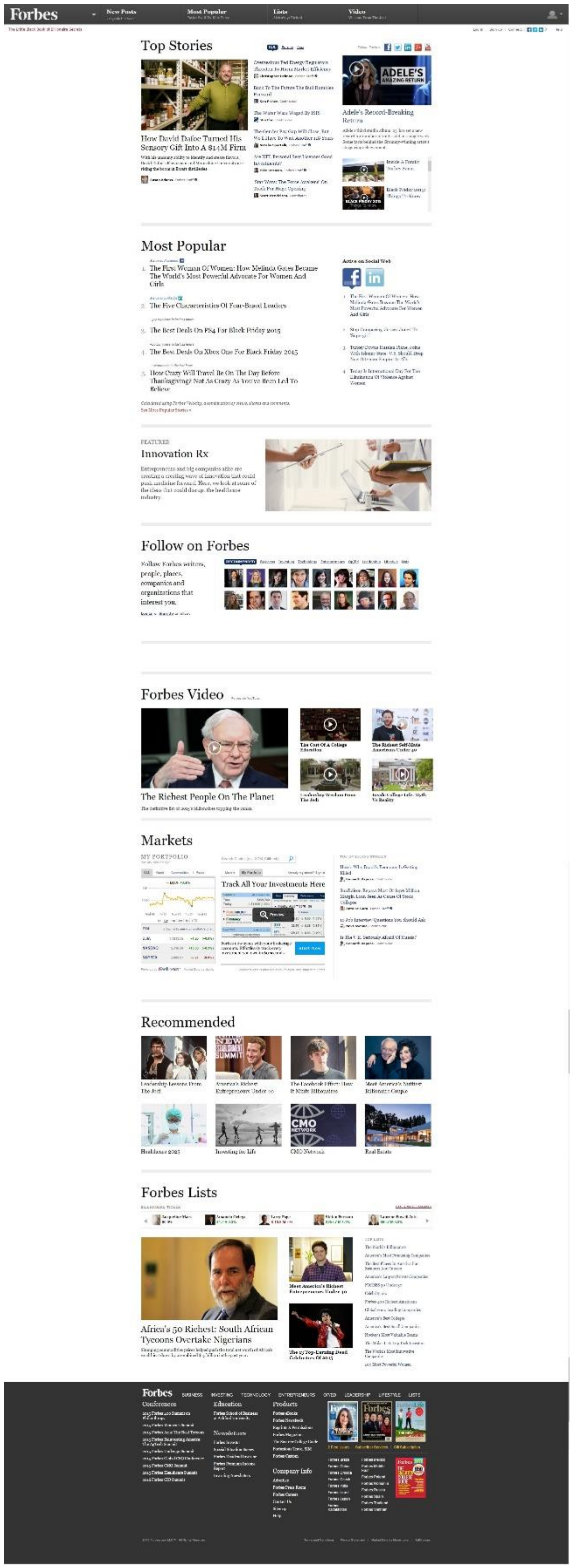


Figure B4. Homepage www.forbes.comwebsite

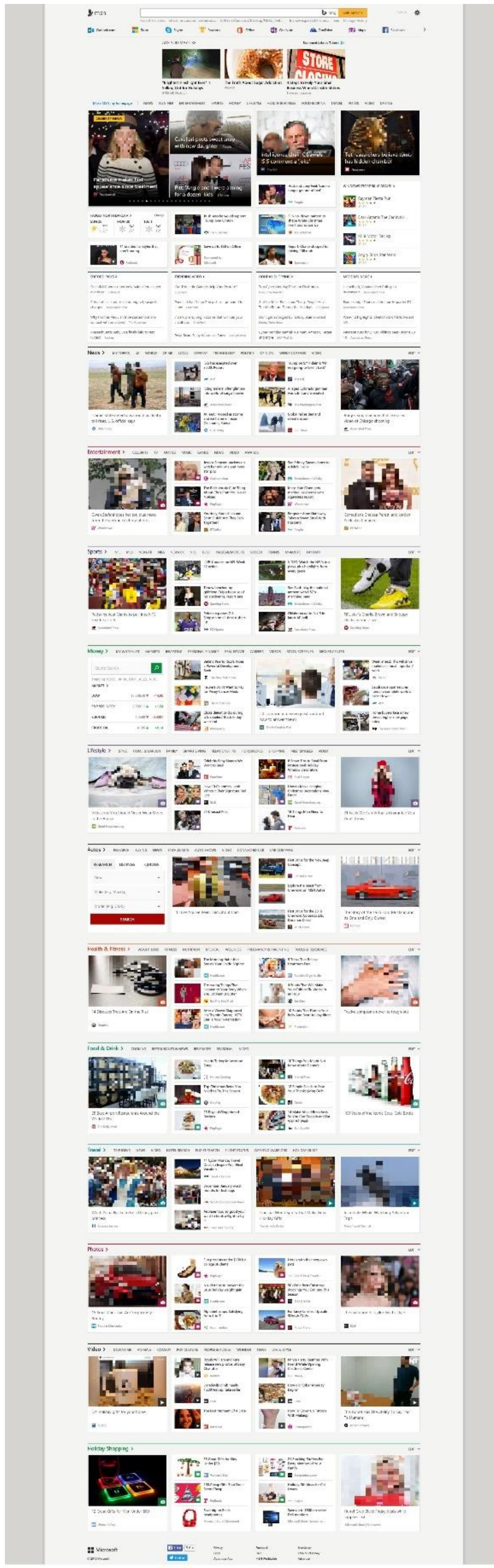


Figure B5. Homepage of www.msn.com website

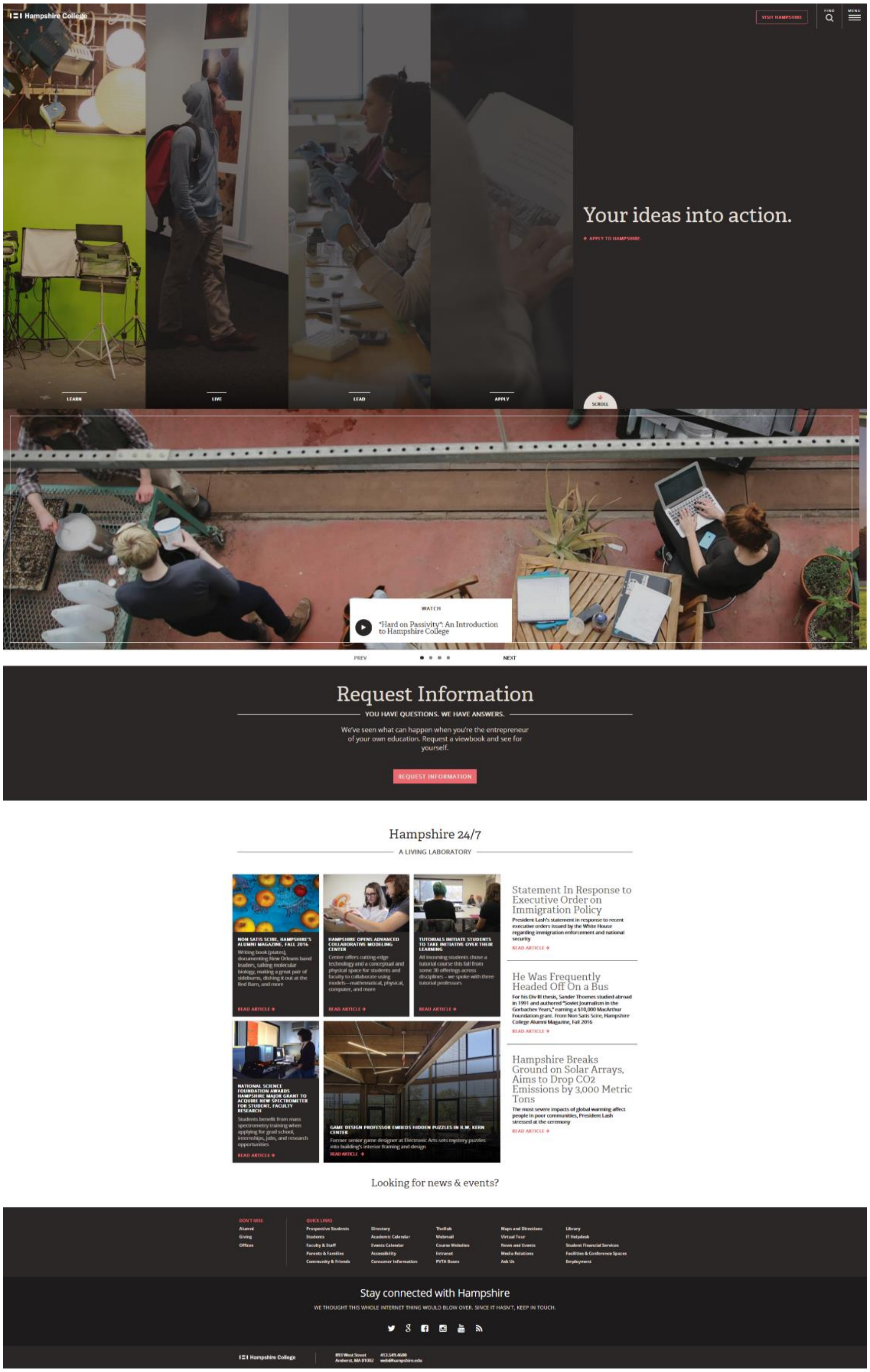


Figure B6. Homepage of www.hampshire.edu website

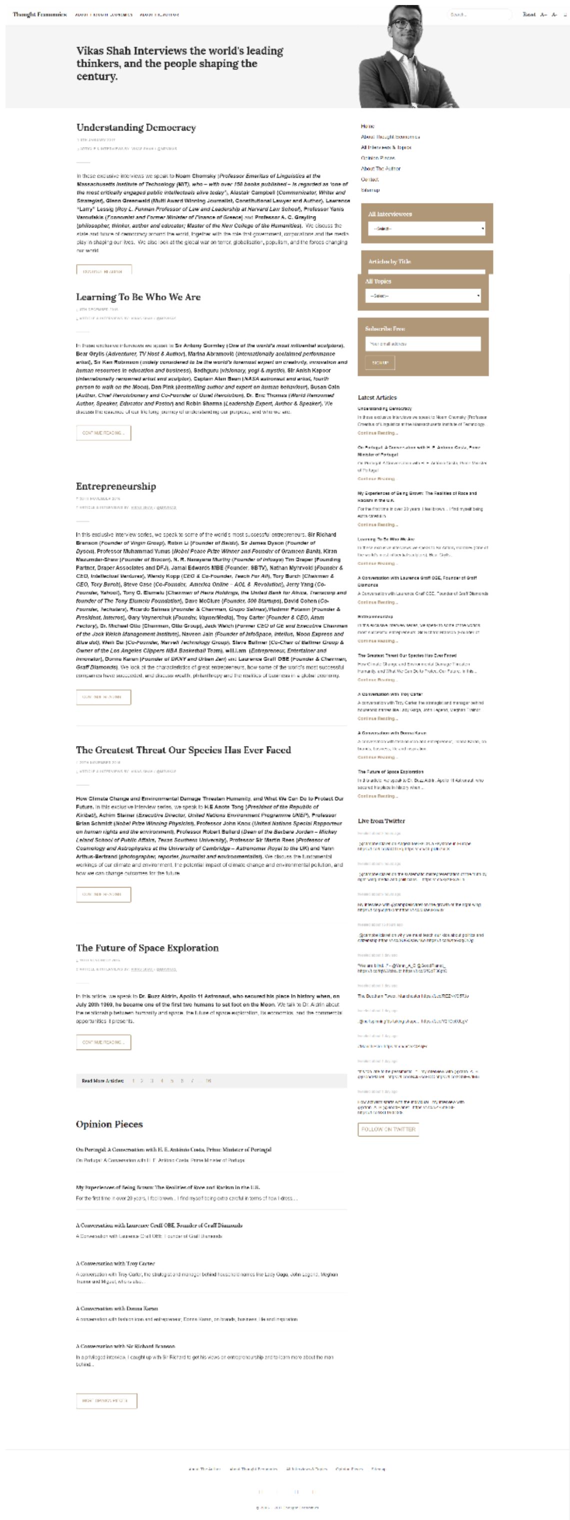


Figure B7. Homepage of www.thoughteconomics.com website

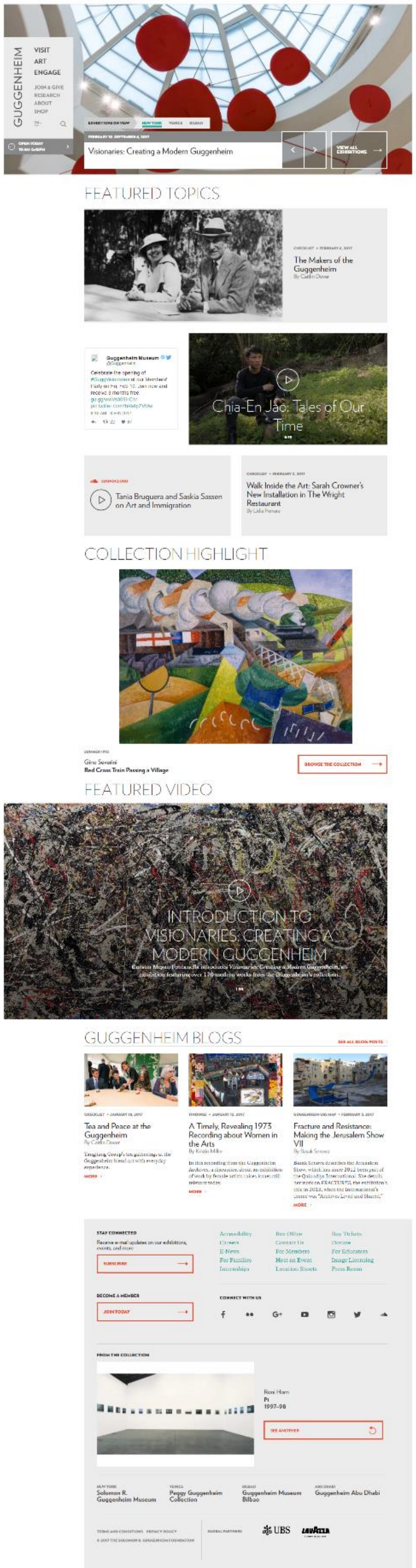


Figure B8. Homepage of www.guggenheim.org website

Figure B9. Homepage of www.oracl.com website

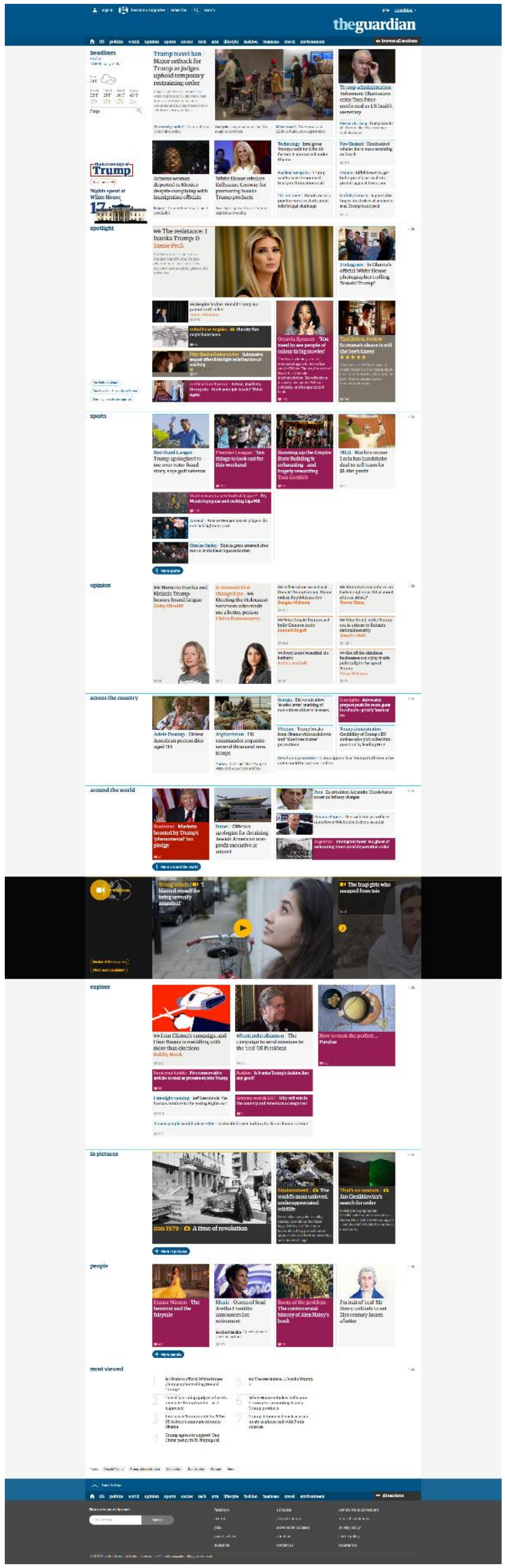


Figure B10. Homepage of www.theguardian.com website

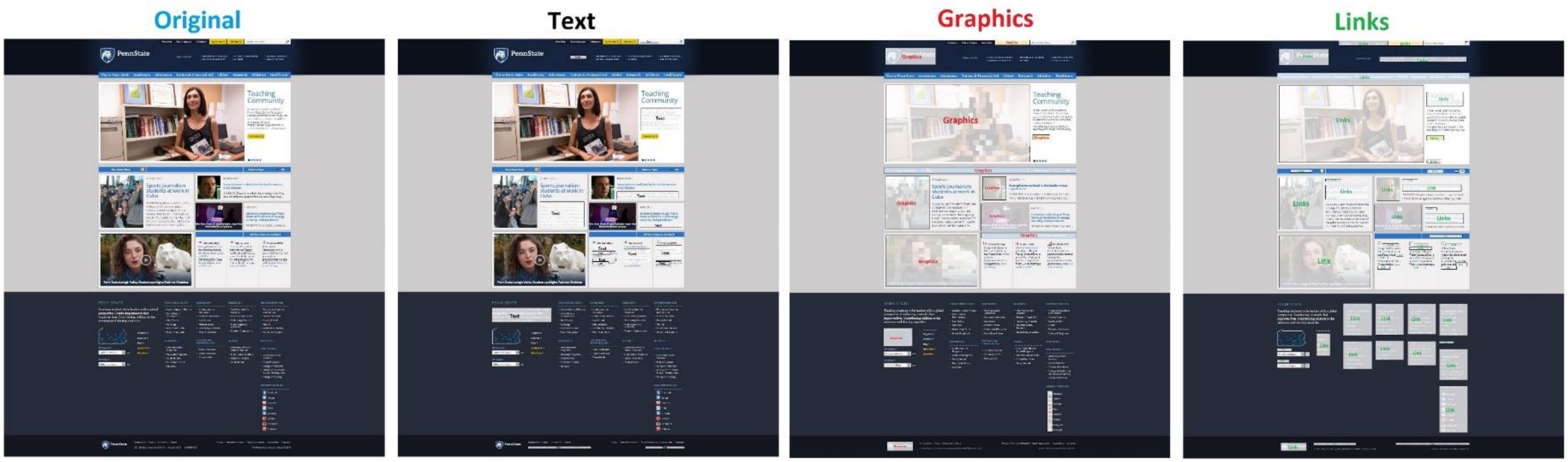


Figure B11. Three model screens of www.psu.edu website

Figure B12. Three model screens www.graceland.com website

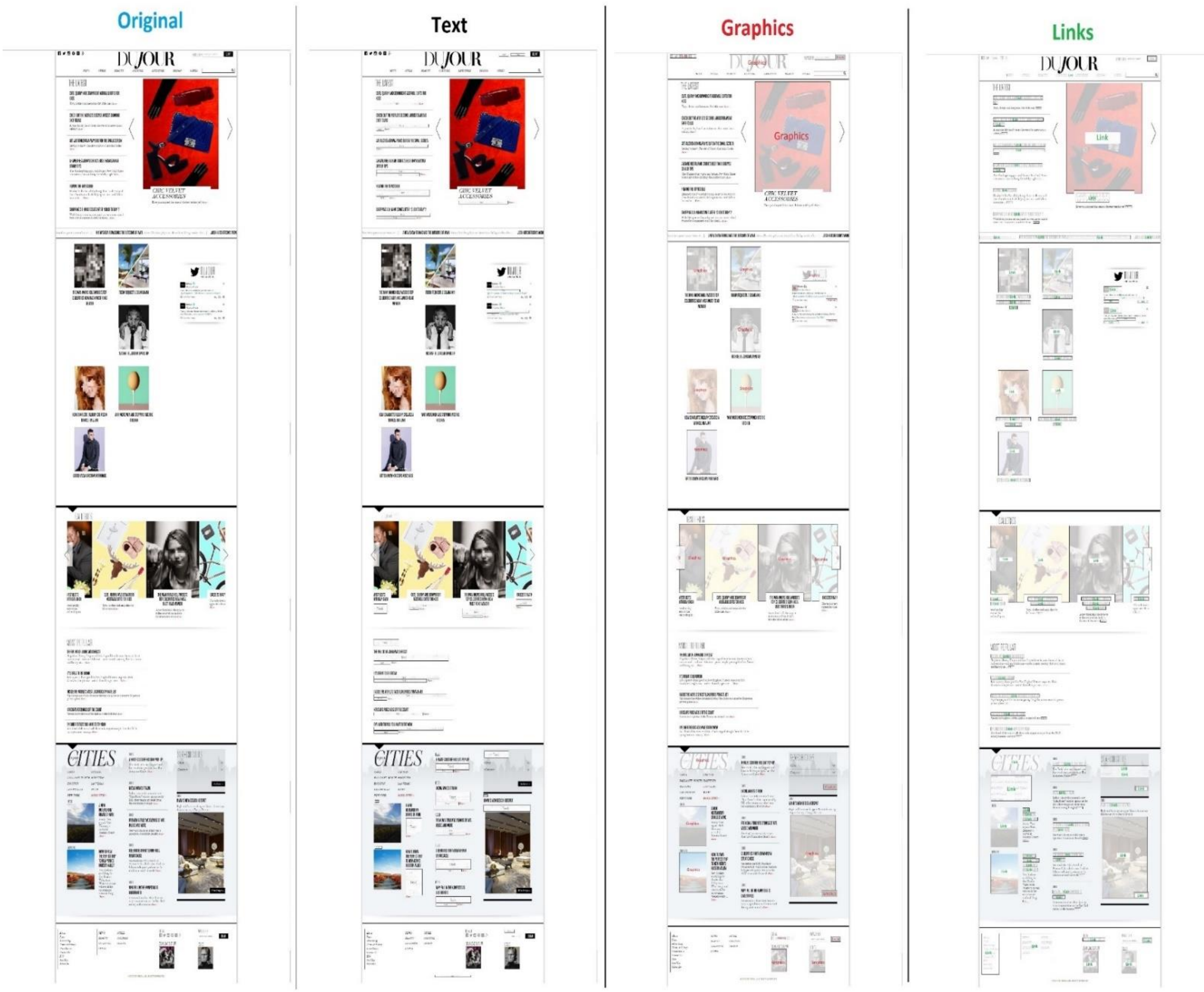


Figure B13. Three model screens of www.dujour.com website

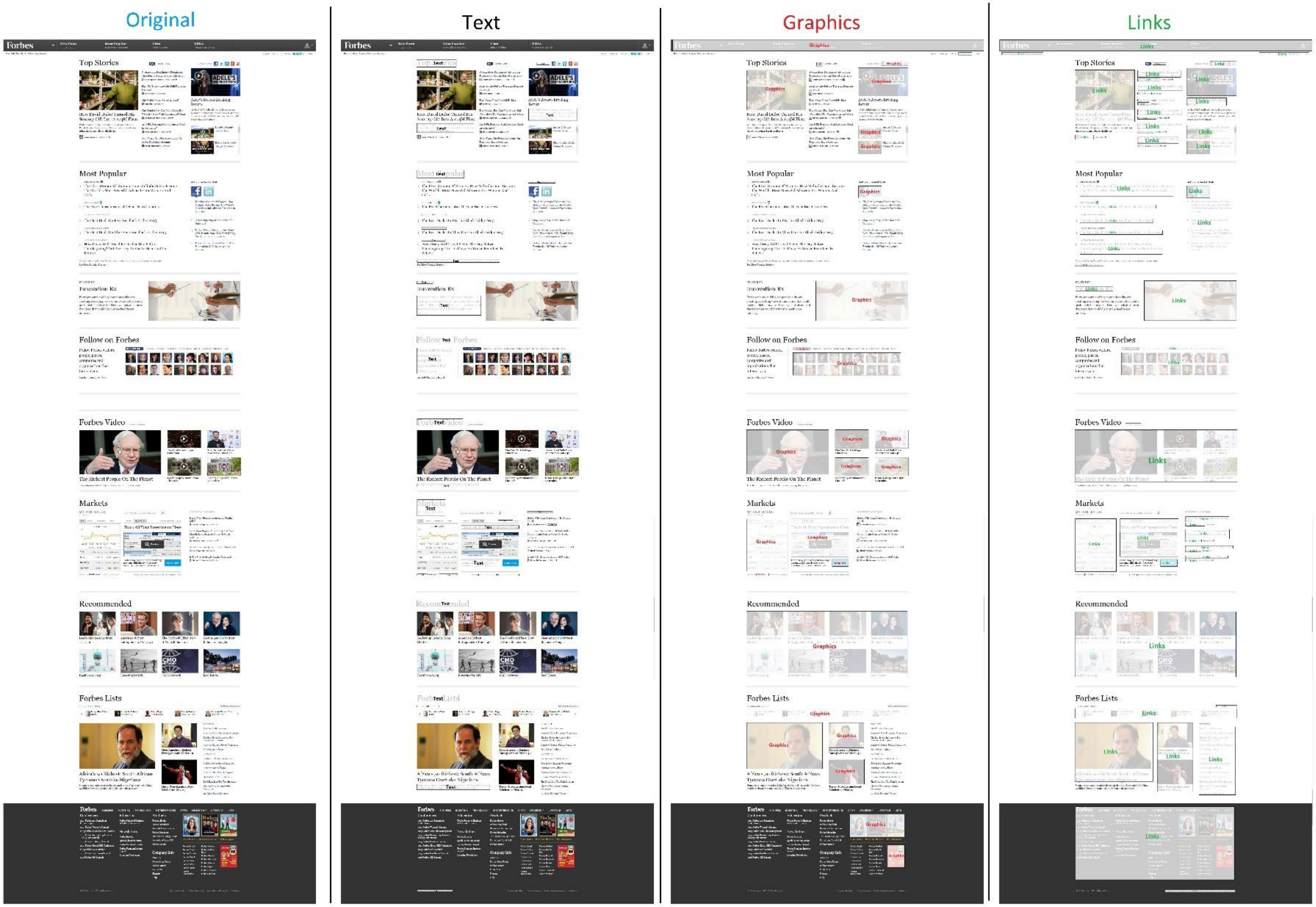


Figure B14. Three model screens of www.forbes.com website

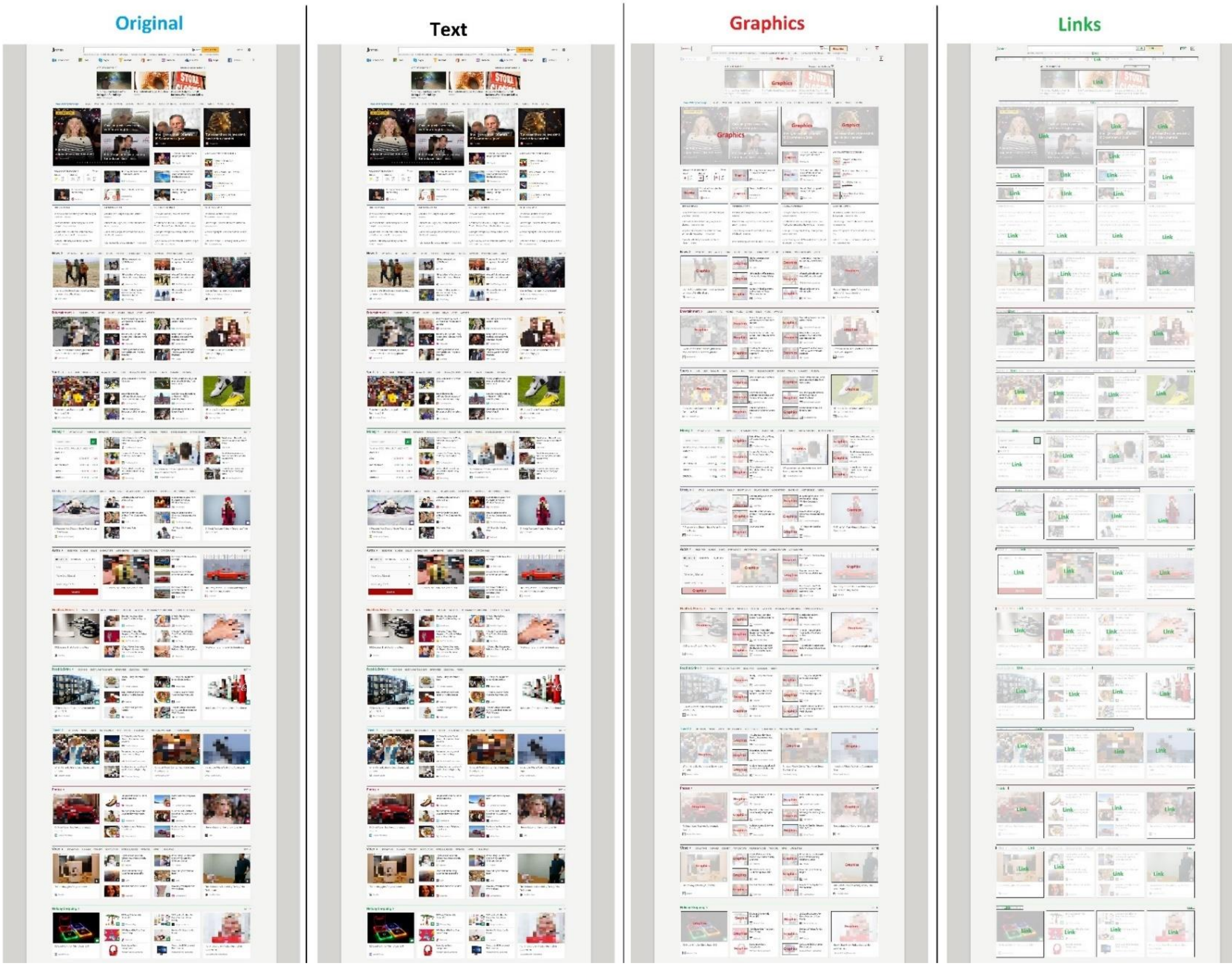


Figure B15. Three model screens of www.msn.com website

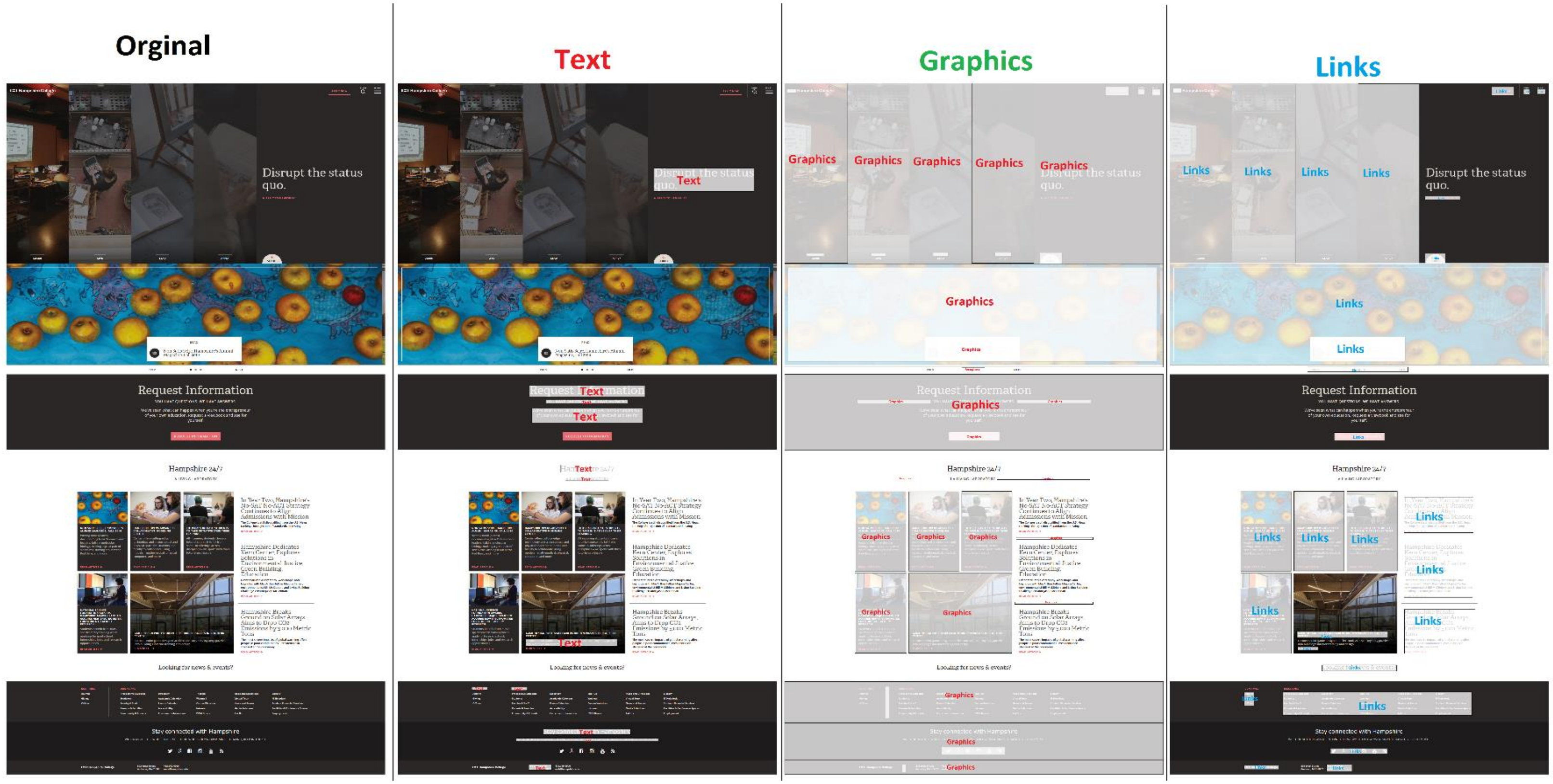


Figure B16. Three model screens of www.hampshire.edu website

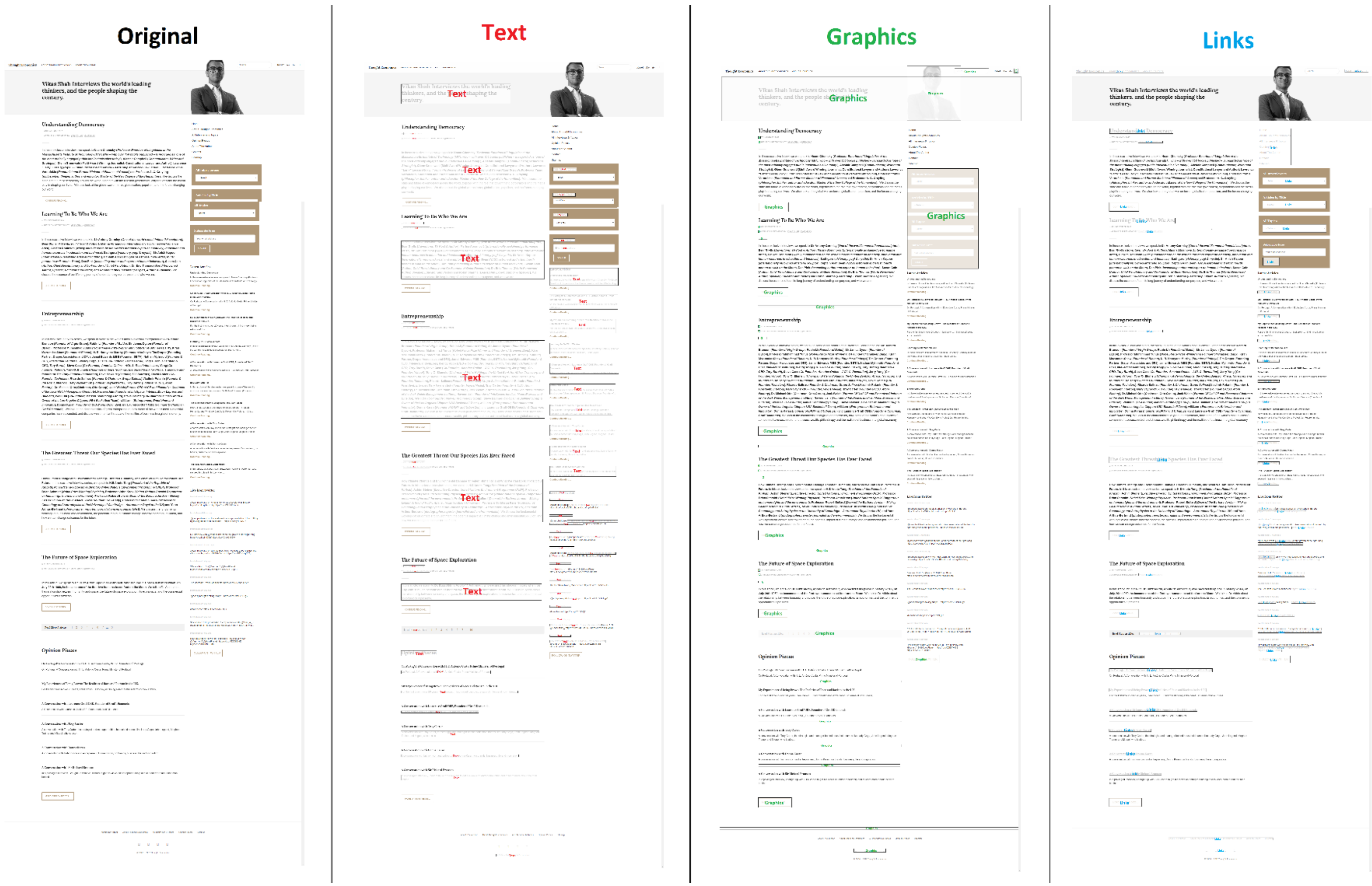


Figure B17. Three model screens of www.thoughteconomics.com website

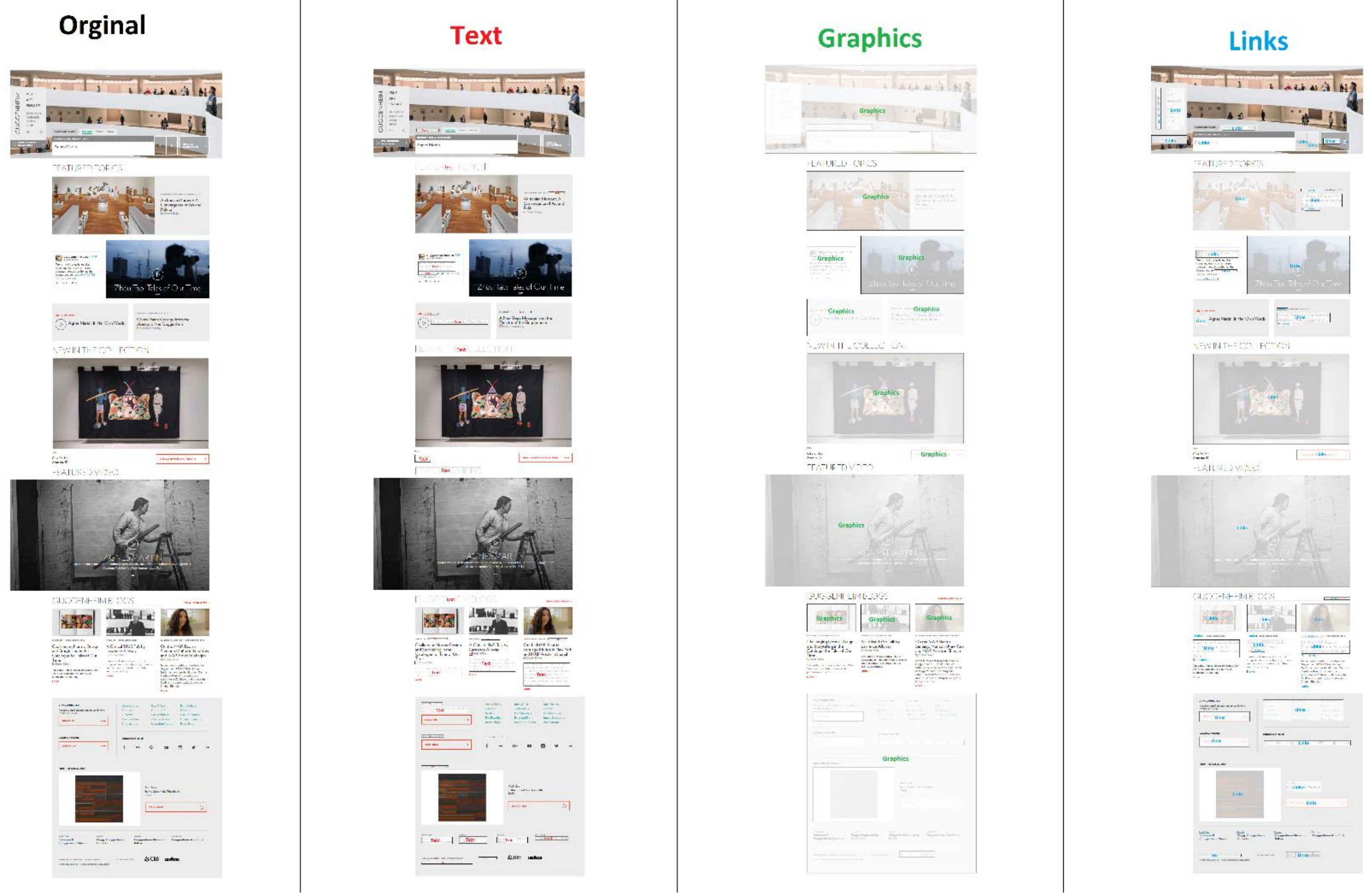


Figure B18. Three model screens of www.guggenheim.org website

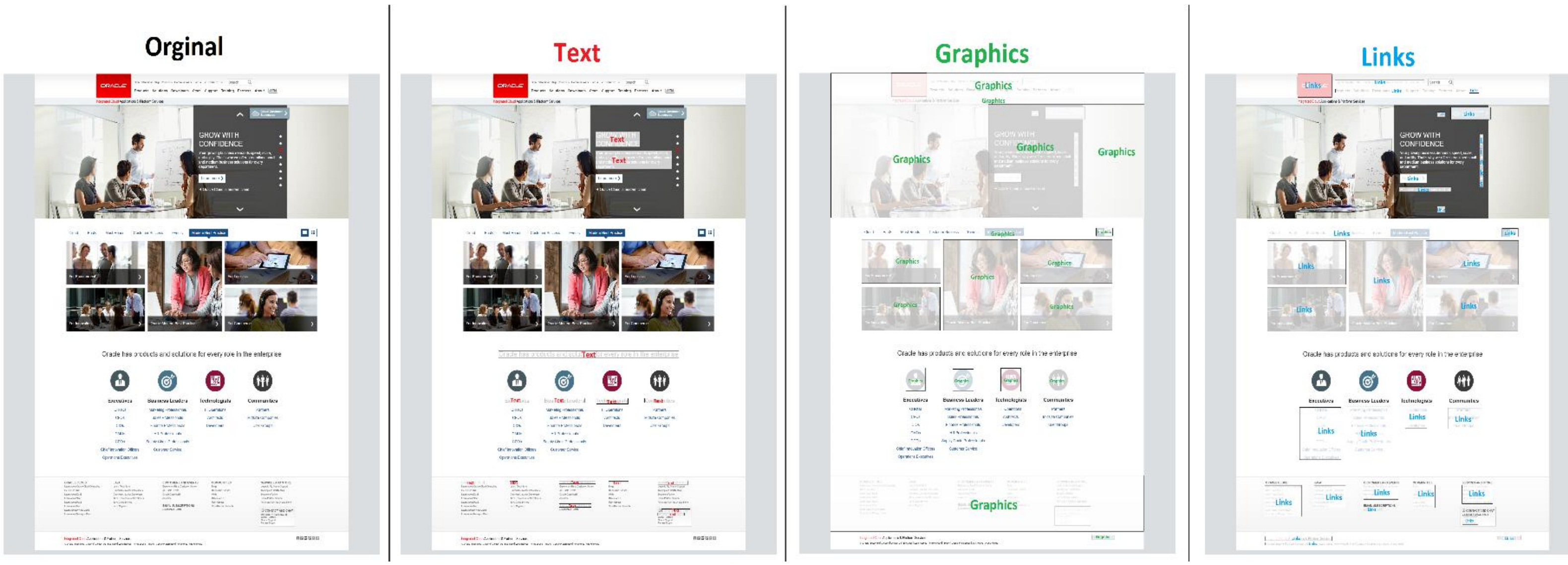


Figure B19. Three model screens of www.oracle.com website

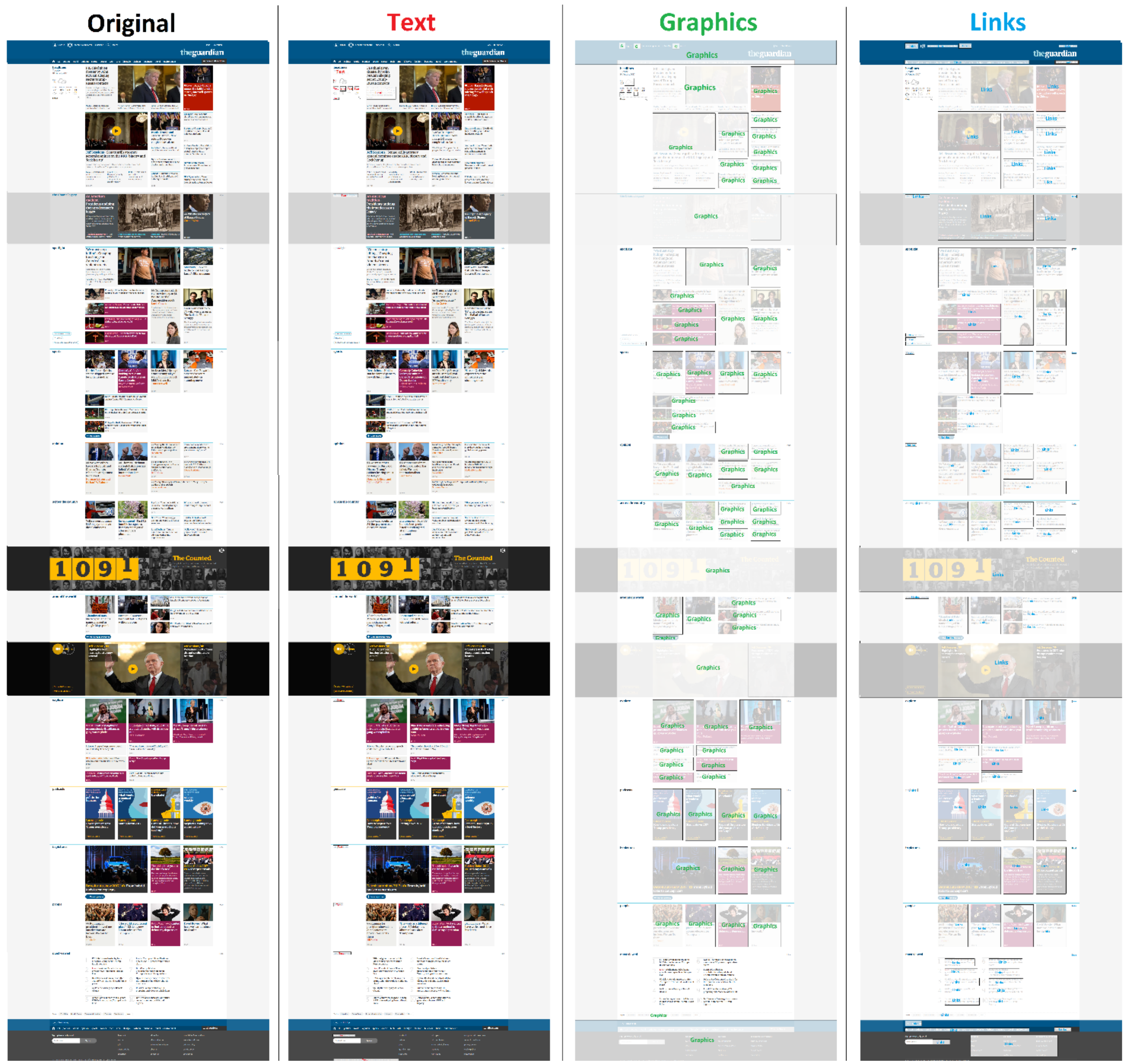


Figure B20. Three model screens of www.theguardian.com website

APPENDIX C. PRE-EXPERIMENTS QUESTIONNAIRE

Survey 1

* Required

Introduction & Consent

NDSU

North Dakota State University
Department of Computer Science
Campus Address
NDSU Dept. #2740
PO Box 6050
Fargo, ND 58108-6050
701.231.8562

Title of Research Study: A Structural Complexity Model for Web Interfaces

Dear NDSU Participant,

My name is Abdulaziz Attaallah. I am a graduate student in Computer Science Department at North Dakota State University, and I am conducting a research project to measure the complexity level of websites visual layouts. It is our hope, that with this research, we will learn more about what aspects of websites structures resulting in this sense of complexity to users.

Because you are a websites user and have some background in dealing with World Wide Web (WWW), you are invited to take part in this research project. Your participation is entirely your choice, and you may change your mind or quit participating at any time, with no penalty to you.

It is not possible to identify all potential risks in research procedures, but we have taken reasonable safeguards to minimize any known risks such leaking your personal information

By taking part in this research, you and we will be compensating participants for their time with $10.00 cash reward, or you can chose not to participate and as a result you will not get in return any rewards.

However, you may not get any benefit from being in this research study other than what has been mentioned above. Benefits to others such software engineering community are likely to include the quantitative results of the experiment in future for advancement in the same research area.

It should take about 45 minutes to complete the experiment's tasks and the questions.You will be doing the following:
You will be filling four online survey forms, the first one is about your computer and internet background and the others are about your evaluation of some websites' layouts.

We will keep private all research records that identify you. Your information will be combined with information from other people taking part in the study, we will write about the combined information that we have gathered. You will not be identified in these written materials. We may publish the results of the study; however, we will keep your name and other identifying information private.

If you have any questions about this project, please contact me at Email: abdulaziz.attaallah@ndsu.edu and Phone: (407)-818-6933, or contact my advisor at Dr. Keneth Magel at Email: Kenneth.Magel@ndsu.edu and Phone: (701) 231-8189.

This study has been accepted by the NDSU Institutional Review Board (IRB), approval number #SM16077
If you have any questions about the rights of human research participants, or if you would like to report a problem, please contact the NDSU IRB Office at (701) 231-8995 or email at NDSU.IRB@ndsu.edu. In addition, if you have any questions regarding this study or would like additional information about the study, please contact me at Abdulaziz.Attaallah@Ndsu.edu or (407) 818-6933 or you can also contact Dr. Kenneth Magel at Kenneth.Magel@ndsu.edu or (701) 231-8189

Thank you for your taking part in this research. If you wish to receive a copy of the results, please contact me using my contact information mentioned above.

Physical Address:
Quentin Burdick Building Room 112

Fargo, ND 58102

Abdulaziz Attaallah, Ph.D. Candidate
Software Engineering
Abdulaziz.attaallah@ndsu.edu
(701) 799-4139

Kenneth Magel, Ph.D.
Software Engineering
Associate Head
Kenneth.Magel@ndsu.edu
(701) 231-8189

If you pressed the continue button at the bottom, basically you are agreeing to participate in this study.

We appreciate your participation.

1. **First Name** *

2. **Last Name** *

Questions

3. **Assigned User ID:**
   * Filled by the researcher

4. **1 - What is your Gender?**
   Mark only one oval.
   - Male
   - Female

5. **2 - How old are you?**
   Mark only one oval.
   - 18 - 25
   - 26 - 39
   - 40 - 59
   - > 60

6. **3 - What kind of student are you?**
   Mark only one oval.
   - Graduate Student
   - Undergraduate Student

7. **4 - Specify your school major?**

Mark only one oval.

- ⬭ Yes, I do.
- ⬭ No, I don't.
- ⬭ Other: ____________

If your answered "No" on question 5, please answer questions 6 and 7. Otherwise, skip them.

9. **6 - Have you taken any computer courses?**
Mark only one oval.

- ⬭ Yes, I have.
- ⬭ No, I have not.

Computer Experience & Web Familiarity

10. **7- What kind(s) of programs have you worked with? Check all that apply**
Check all that apply.

- ☐ Word Processing
- ☐ Spreadsheets
- ☐ Graphics
- ☐ Other: ____________

11. **8- How many hours do you surf the internet daily?**
Mark only one oval.

- ⬭ Less than 1 hour
- ⬭ 1 - 3 Hours/Day
- ⬭ 4 - 6 Hours/Day
- ⬭ More than 7 Hours/Day

12. **9- How would you rate your experience and proficiency using the Internet?**
Mark only one oval.

	1	2	3	4	5	
Very Poor	⬭	⬭	⬭	⬭	⬭	Excellent

13. **10- How long have you been using personal computers?**
Mark only one oval.

- ⬭ 0 - 3 months
- ⬭ 4 - 6 months
- ⬭ 7 - 8 months
- ⬭ 1 - 2 years
- ⬭ 3 - 4 years
- ⬭ 5 - 6 years
- ⬭ Other: ____________

Check all that apply.

- [ ] Chrome
- [ ] Internet Explorer
- [ ] Safari
- [ ] Firefox
- [ ] Microsoft Edge
- [ ] Other: ______________________

15. **12- Where do you use email? Check all that apply.**
Check all that apply.

- [ ] From a personal home account
- [ ] From an account at work
- [ ] From a public access terminal (library, Internet cafe, other)
- [ ] From school

16. **13- What do you use email for? Check all that apply.**
Check all that apply.

- [ ] Work
- [ ] Personal
- [ ] Other: ______________________

17. **14- How often do you visit your university website?**
Mark only one oval.

	1	2	3	4	5	
Rarely	◯	◯	◯	◯	◯	Regularly

Domain Knowledge

18. **15- For what purposes do you visit your university website? Check all that apply.**
Check all that apply.

- [ ] Grading & Assessment
- [ ] Assignments & Collaboration
- [ ] Course Registration
- [ ] Studying
- [ ] Examination
- [ ] Researching
- [ ] Reading
- [ ] Looking up people contacts

APPENDIX D. FIRST EXPERIMENT SURVEY

Survey 2

* Required

1. **Participant ID**
   Assigned by researcher

2. **First Name** *

3. **Last Name** *

Instructions: - SCROLL up and down, and switch between tabs of the 6 websites. - DO NOT use the mouse lift or right click - Based on your observation answer the following questions

Participants View On Complexity

Type FIVE reasons or factors that make webpages complex (confusing or less understandable).

4. **Reason or Factor 1**
   Factor Definition: An influence that contributes to a result or outcome

5. **Reason or Factor 2**
   Factor Definition: An influence that contributes to a result or outcome

6. **Reason or Factor 3**
   Factor Definition: An influence that contributes to a result or outcome

7. **Reason or Factor 4**
   Factor Definition: An influence that contributes to a result or outcome

Factor Definition: An influence that contributes to a result or outcome

Name five OBJECTS or ELEMENTS that make webpages complex (confusing or less understandable).

Example

Objects: Menus

9. **Object or Element 1**
   Object Definition: A material thing that can be seen and touched

10. **Object or Element 2**
    Object Definition: A material thing that can be seen and touched

11. **Object or Element 3**
    Object Definition: A material thing that can be seen and touched

12. **Object or Element 4**
    Object Definition: A material thing that can be seen and touched

13. **Object or Element 5**
    Object Definition: A material thing that can be seen and touched

Webpage Factors

Look at the following Illustrations and evaluate each factor in the followed question

Factors' Illustrations

Symmetry of objects

Good
Bad
Sequence of objects
Good
Bad
Unity of objects
Shapes
Good
Bad
Grouping of objects
Good
Bad
Size of objects
Good
Bad
Spacing of objects
Good
Bad
Density of objects
Good
Bad
Alignment of objects
Good
Bad
Balance of objects

Fonts of objects

Colors of objects

14. **Rate the factors based on the negative impact on a page?**
Less Negative Impact--- More Negative Impact
Mark only one oval per row.

	1	2	3	4	5
Unity of objects					
Density of objects					
Size of objects					
Balance of objects					
Fonts of objects					
Alignment of objects					
Regularity of objects					
Grouping of objects					
Sequence of objects					
Symmetry of objects					
Colors of objects					
Spacing of objects					

Webpage Objects

15. **Rate the objects or elements based on the negative impact on the page?**
Less Negative Impact--- More Negative Impact
Mark only one oval per row.

	1	2	3	4	5
Lists					
Graphics					
Videos					
Links					
Audios					
Text					
Buttons					
Menues					
Search boxes					

Study of Webpages

Less Negative Impact--- More Negative Impact
Mark only one oval per row.

	1	2	3	4	5
Links	◯	◯	◯	◯	◯
Text	◯	◯	◯	◯	◯
Graphics	◯	◯	◯	◯	◯

17. **Rate the complexity factors based on the negative impact on the page?**
Less Negative Impact--- More Negative Impact
Mark only one oval per row.

	1	2	3	4	5
Density of Objects	◯	◯	◯	◯	◯
Alignment of Objects	◯	◯	◯	◯	◯
Grouping of Objects	◯	◯	◯	◯	◯
Size of Objects	◯	◯	◯	◯	◯

18. **Rate the the following webpage elements based on the SIZE which leads to more complex or confusing pages**
Less Negative Impact--- More Negative Impact
Mark only one oval per row.

	1	2	3	4	5
Size of Links	◯	◯	◯	◯	◯
Size of Text	◯	◯	◯	◯	◯
Size of Graphics	◯	◯	◯	◯	◯

19. **Rate the the following webpage elements based on the DENSITY which leads to more complex pages**
Less Negative Impact--- More Negative Impact
Mark only one oval per row.

	1	2	3	4	5
Density of Text	◯	◯	◯	◯	◯
Density of Links	◯	◯	◯	◯	◯
Density of Graphics	◯	◯	◯	◯	◯

20. **Rate the the following webpage elements based on the GROUPING which leads to more complex pages**
Less Negative Impact--- More Negative Impact
Mark only one oval per row.

	1	2	3	4	5
Grouping of Links	◯	◯	◯	◯	◯
Grouping of Text	◯	◯	◯	◯	◯
Grouping of Graphics	◯	◯	◯	◯	◯

21. **Rate the the following webpage elements based on the ALIGNMENT which leads to more complex pages**
Less Negative Impact--- More Negative Impact
Mark only one oval per row.

	1	2	3	4	5
Alignment of Text	◯	◯	◯	◯	◯
Alignment of Links	◯	◯	◯	◯	◯
Alignment of Graphics	◯	◯	◯	◯	◯

APPENDIX D. SECOND EXPERIMENT SURVEYS: MODEL SCREENS VERSUS REAL SCREEN SURVEY

Survey 3

* Required

1. **Participant ID**
   Assigned by researcher

2. **First Name** *

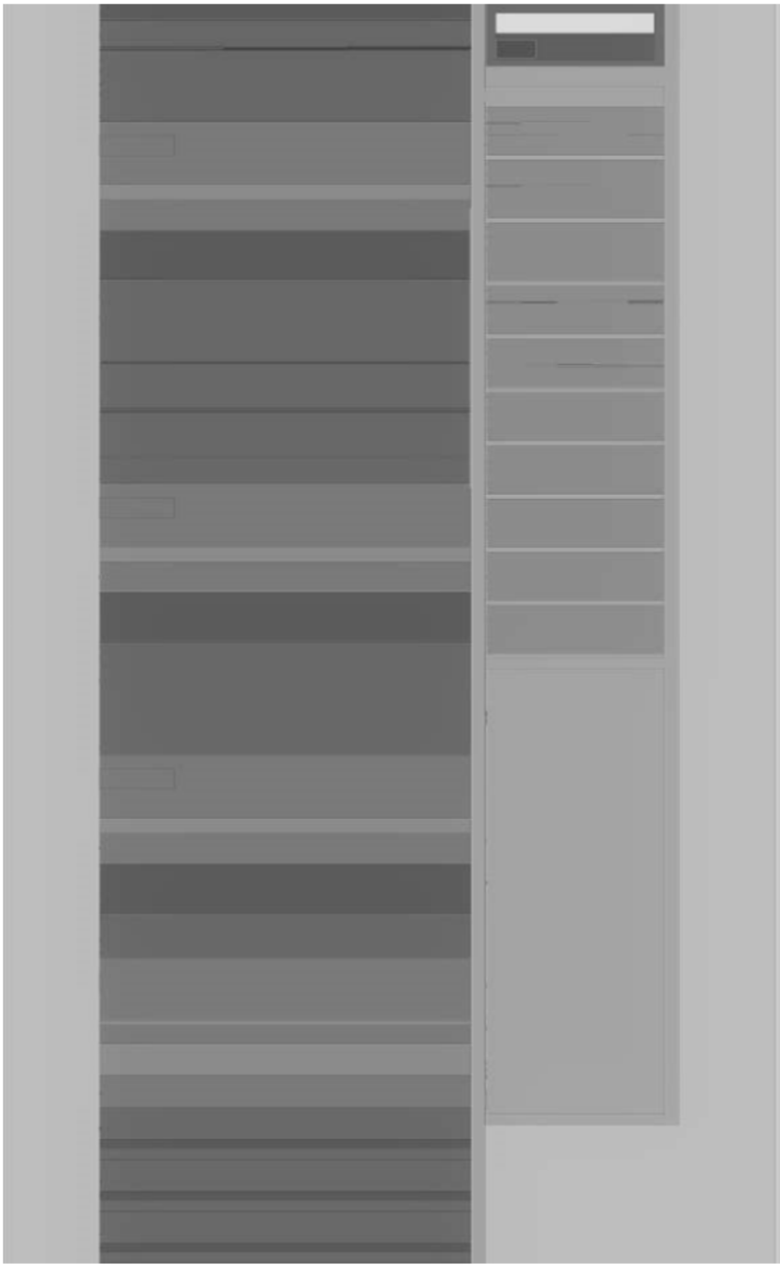
3. **Last Name** *

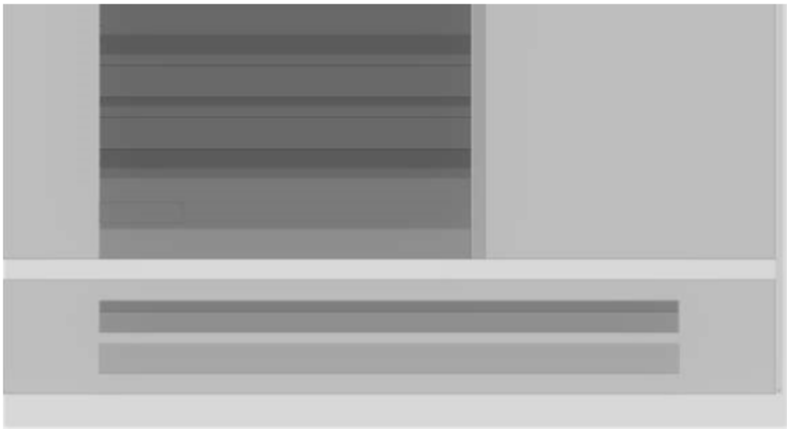
Instruction: Look at each webpage structure image, and evaluate each factor

Webpage Complexity

The following image is the structure layout of the website: www.thoughteconomics.com/. Evaluate the layout based on the factors highlighted below.

www.thoughteconomics.com/





4. **The SIZE of objects on the layout may result in:**
   Less Understandable Webpage--- More Understandable Webpage
   Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

5. **The DENSITY of objects on the webpage makes it**
   Less Understandable Webpage--- More Understandable Webpage
   Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

6. **The GROUPING of objects on the webpage makes it**
   Less Understandable Webpage--- More Understandable Webpage
   Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

7. **The ALIGNMENT of objects on the webpage makes it**
   Less Understandable Webpage--- More Understandable Webpage
   Mark only one oval per row.

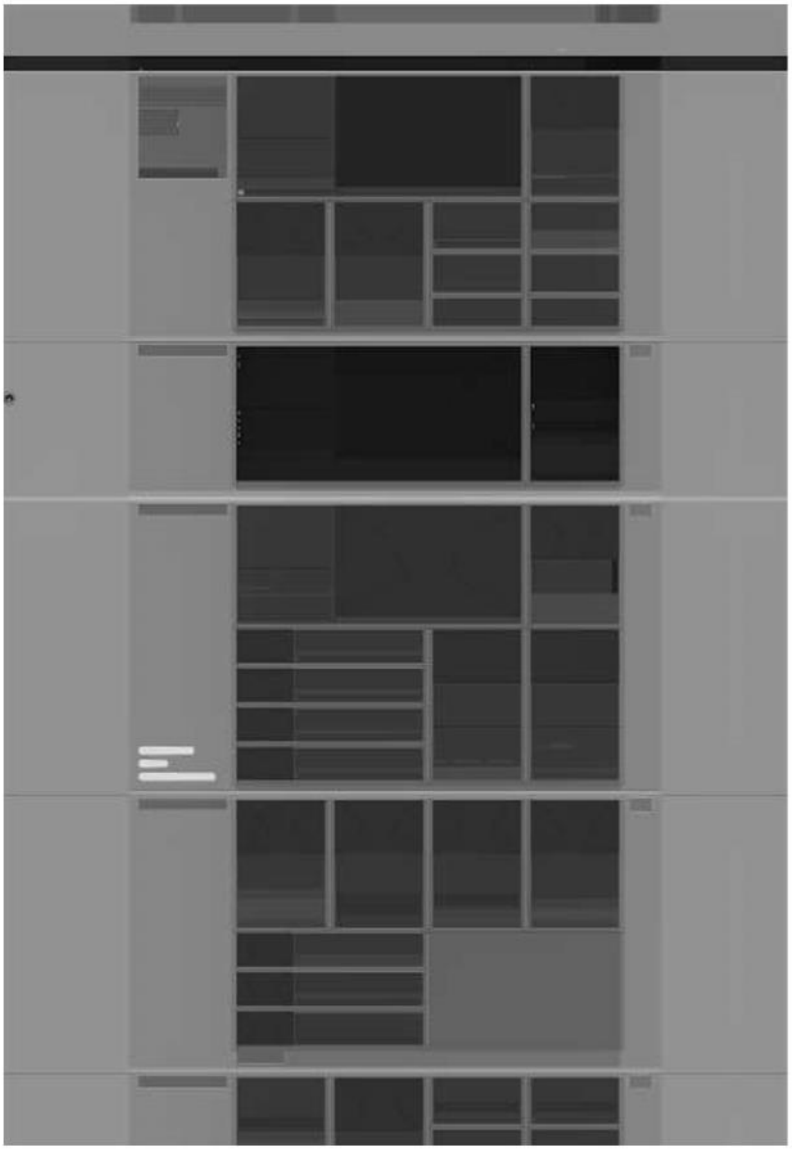
	1	2	3	4	5
	◯	◯	◯	◯	◯

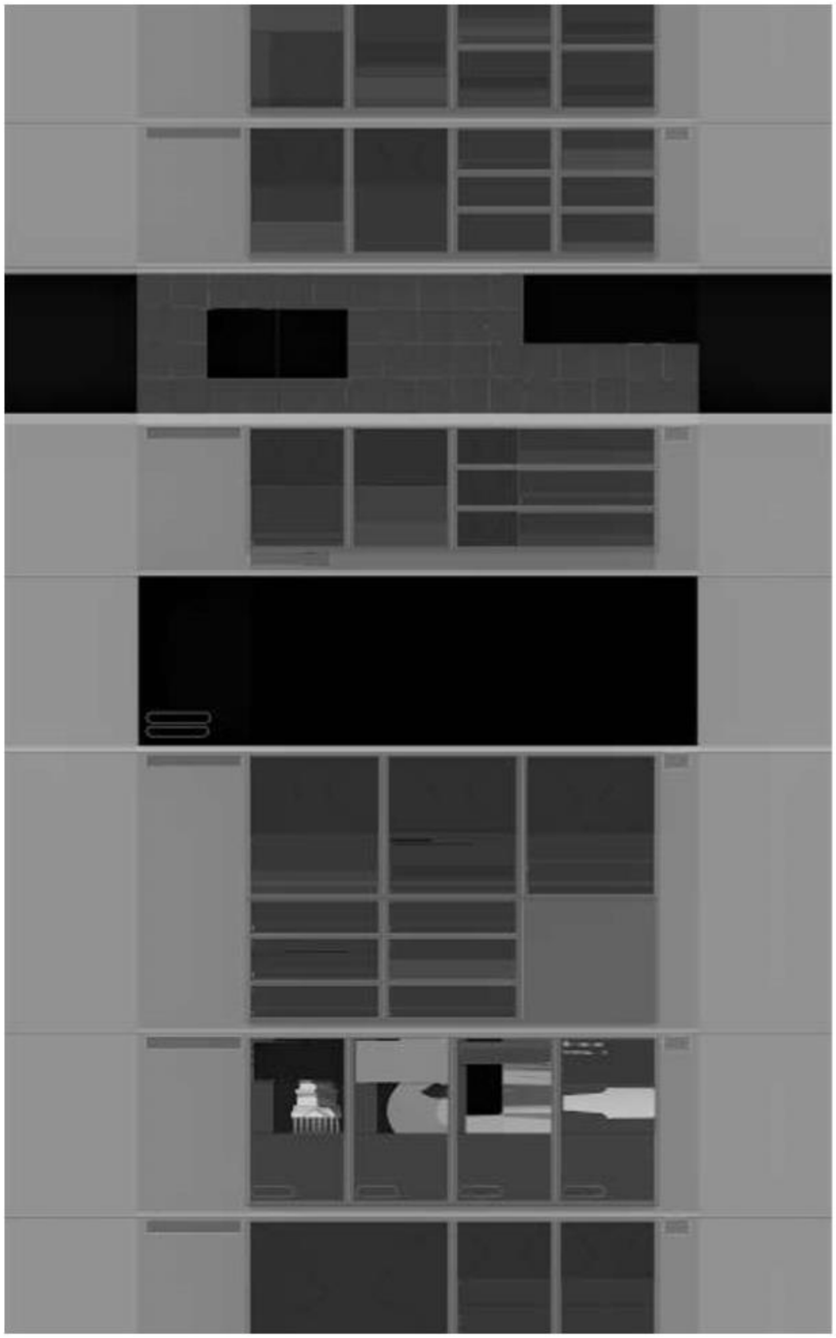
Webpage Complexity

The following image is the structure layout of the website: www.theguardian.com/. Evaluate the layout based on the factors

highlighted below.

www.theguardian.com/

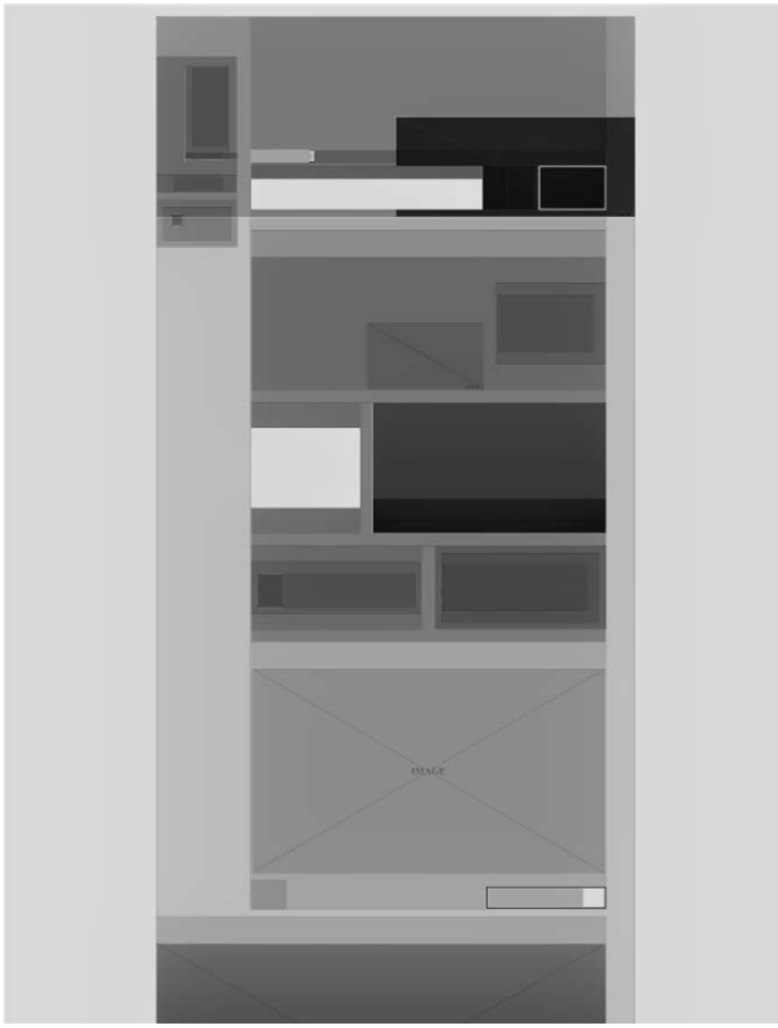




Webpage Complexity

The following image is the structure layout of the website: www.guggenheim.org. Evaluate the layout based on the factors highlighted below.

www.guggenheim.org/



14. **The GROUPING of objects on the webpage makes it**

Less Understandable Webpage-- More Understandable Webpage

Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

15. **The ALIGNMENT of objects on the webpage makes it**

Less Understandable Webpage-- More Understandable Webpage

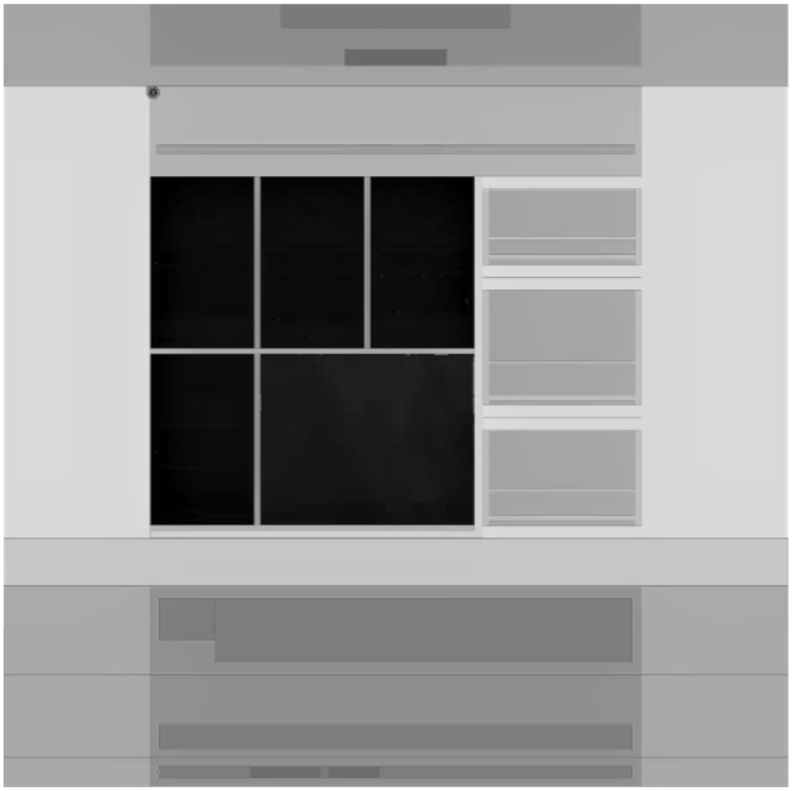
Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

Webpage Complexity

The following image is the structure layout of the website:www.hampshire.edu/. Evaluate the layout based on the factors highlighted below.

www.hampshire.edu/



16. **The SIZE of objects on the layout may result in:**

Less Understandable Webpage-- More Understandable Webpage

Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

17. **The DENSITY of objects on the webpage makes it**

Less Understandable Webpage-- More Understandable Webpage

Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

18. **The GROUPING of objects on the webpage makes it**

Less Understandable Webpage-- More Understandable Webpage

Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

19. **The ALIGNMENT of objects on the webpage makes it**

Less Understandable Webpage-- More Understandable Webpage

Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

Webpage Complexity

The following image is the structure layout of the website: www.oracle.com. Evaluate the layout based on the factors highlighted below.

www.oracle.com

20. **The SIZE of objects on the layout may result in:**
Less Understandable Webpage-- More Understandable Webpage
Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

21. **The DENSITY of objects on the webpage makes it**
Less Understandable Webpage-- More Understandable Webpage
Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

22. **The GROUPING of objects on the webpage makes it**

Less Understandable Webpage-- More Understandable Webpage

Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

23. **The ALIGNMENT of objects on the webpage makes it**

Less Understandable Webpage-- More Understandable Webpage

Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

APPENDIX E. SECOND EXPERIMENT SURVEYS: ELEMENTS' ATTRIBUTES SURVEY

Survey 4

* Required

1. **Participant ID**
   Assigned by researcher

2. **First Name** *

3. **Last Name** *

4. **Webpage Title**
   Entered by researcher

Instructions: - Use the Windows Photo Viewer application to observe 6 webpages. - Each webpage has 3 copies, and each one has an element HIGHLIGHTED with borders. - Answer each section of survey's questions by observing associated titled image.

Text (Size)

Rate the following statements:

Example

The current font size is: 6pt
The current font size is: 8pt
The current font size is: 9pt
The current font size is: 10pt
The current font size is: 10.95pt
The current font size is: 12pt
The current font size is: 14.4pt
The current font size is: 17.28pt
The current font size is: 20.74pt
The current font size is: 24.88pt

5. **Using a variety of font sizes make the page**
   Less Complex -- More Complex
   Mark only one oval per row.

1	2	3	4	5
◯	◯	◯	◯	◯

Less Complex -- More Complex
Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

Text (Density)

Rate the following statements:

Example

North Dakota State University is distinctive as a student-focused, land-grant, research university, that provides affordable access to an excellent education at a top-ranked institution that combines teaching and research in a rich learning environment, educating future leaders who will create solutions to national and global challenges that will shape a better world. NDSU is listed in the National Science Foundation's top 100 in several areas, including agricultural sciences, social sciences, physical sciences, chemistry, psychology and computer sciences. NDSU is fully accredited as an institution by the Higher Learning Commission.

North Dakota State University is distinctive a research university, that provides affordable access to an excellent education , educating future leaders who will create solutions to national and global challenges that will shape a better world. NDSU is listed in the National Science Foundation's top 100 in several areas, including agricultural sciences, social sciences, physical sciences, chemistry. NDSU is fully accredited as an institution by the Higher Learning Commission.

7. **The number of words covering each text area on the page makes it**
Less Complex -- More Complex
Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

8. **The number of different font sizes for each text area makes the page**
Less Complex -- More Complex
Mark only one oval per row.

	Neutral	1	2	3	4	5
	◯	◯	◯	◯	◯	◯

9. **Using complex font styles or families on the page makes it**
Less Complex -- More Complex
Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

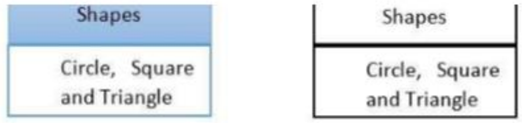
Text (Grouping)

Rate the following statements:

Example

Shapes:

- Circle
- Square
- Triangle



10. **The number of colors used for text (NOT LINKS) makes the page**
Less Complex -- More Complex
Mark only one oval per row.

1	2	3	4	5
◯	◯	◯	◯	◯

11. **Using highlighted-bordered regions for text areas make the page**
Less Complex -- More Complex
Mark only one oval per row.

1	2	3	4	5
◯	◯	◯	◯	◯

12. **Using highlighted-colored regions for text areas make the page**
Less Complex -- More Complex
Mark only one oval per row.

1	2	3	4	5
◯	◯	◯	◯	◯

13. **Using highlighted lists for text areas make the page**
Less Complex -- More Complex
Mark only one oval per row.

1	2	3	4	5
◯	◯	◯	◯	◯

Text (Alignment)

Rate the following statements:

Example

Example

20. **Using different graphic sizes on the page makes it**

Less Complex -- More Complex

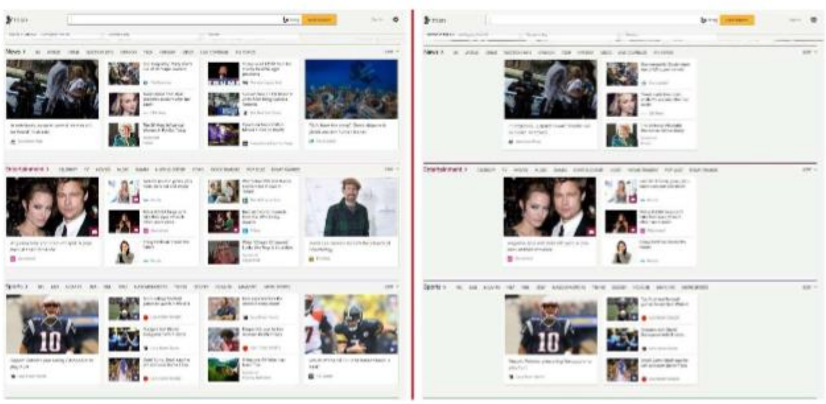
Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

Graphics (Density)

Rate the following statements:

Example



21. **The number of graphics covering each graphic area makes the page**

Less Complex -- More Complex

Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

22. **The number of different graphic sizes covering each graphic area makes the page**

Less Complex -- More Complex

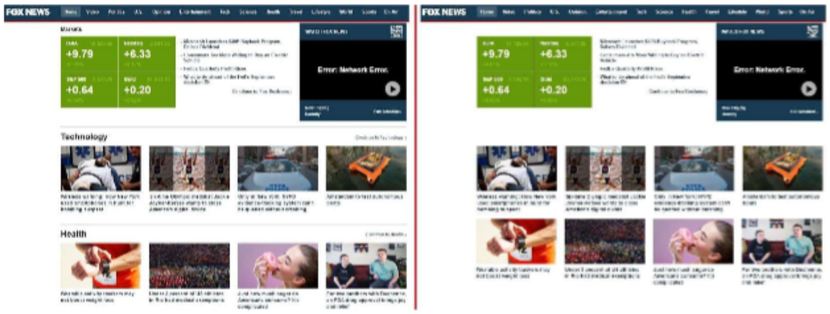
Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

Graphics (Grouping)

Rate the following statements:

Example



23. **Using highlighted-bordered regions for graphic areas make the page**
    Less Complex -- More Complex
    Mark only one oval per row.

    1 2 3 4 5

24. **Using highlighted-colored regions for graphic areas make the page**
    Less Complex -- More Complex
    Mark only one oval per row.

    1 2 3 4 5

25. **Using highlighted lists for graphic areas make the page**
    Less Complex -- More Complex
    Mark only one oval per row.

    1 2 3 4 5

Graphics (Alignment)

Rate the following statements:

Example

26. **The vertical alignment of graphic areas make the page**

Less Complex -- More Complex

Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

27. **The horizontal alignment of graphic areas make the page**

Less Complex -- More Complex

Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

28. **The number of graphic areas that aligned to LEFT make the page**

Less Complex -- More Complex

Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

29. **The number of of graphic areas that aligned to CENTER make the page**

Less Complex -- More Complex

Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

30. **The number of graphics areas that aligned to RIGHT make the page**

Less Complex -- More Complex

Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

31. **The number of graphic areas that are JUSTIFIED (Stretched) make the page more**

Less Complex -- More Complex

Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

Rate the following statements:

Example

32. **Using different font sizes for links on the page make it**
    Less Complex -- More Complex
    Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

33. **Using variety of font styles or families for links which makes it**
    Less Complex -- More Complex
    Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

34. **The number of links on the page that have less than 2 words or more than 4 words make the page**
    Less Complex -- More Complex
    Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

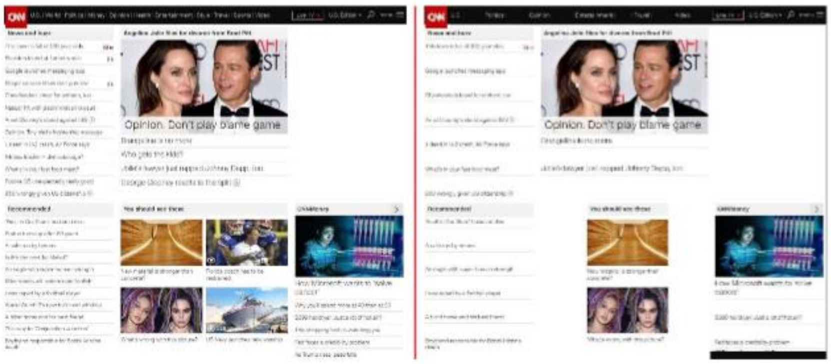
Less Complex -- More Complex
Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

Links (Density)

Rate the following statements:

Example



36. **The number of links for each link area makes the page**
Less Complex -- More Complex
Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

37. **The number of Graphic-Links for each link area makes the page**
Less Complex -- More Complex
Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

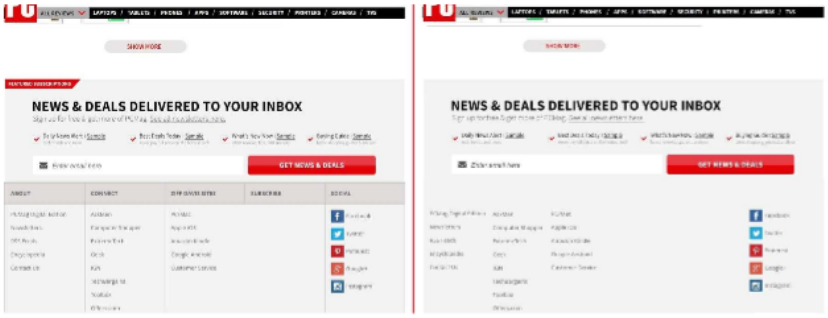
38. **The number of Non-Underlined Links for each link area makes the page**
Less Complex -- More Complex
Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

Links (Grouping)

Rate the following statements:

Example



39. **The number of colors used for links (NOT TEXT) makes the page**

Less Complex -- More Complex

Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

40. **Using highlighted-bordered regions for link areas make the page**

Less Complex -- More Complex

Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

41. **Using highlighted-colored regions for link areas make the page**

Less Complex -- More Complex

Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

42. **Using highlighted lists regions for link areas make the page**

Less Complex -- More Complex

Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

Links (Alignment)

Rate the following statements:

Example

43. **The vertical alignment of link areas make the page**
Less Complex -- More Complex
Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

44. **The horizontal alignment of link areas make the page**
Less Complex -- More Complex
Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

45. **The number of link areas that aligned to LEFT make the page**
Less Complex -- More Complex
Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

46. **The number of link areas that aligned to RIGHT make the page**
Less Complex -- More Complex
Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

Less Complex -- More Complex
Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

48. **The number of link areas that are JUSTIFIED make the page more**

Less Complex -- More Complex
Mark only one oval per row.

	1	2	3	4	5
	◯	◯	◯	◯	◯

APPENDIX F. STUDY OFFICIAL FORMS

NDSU NORTH DAKOTA STATE UNIVERSITY

October 12, 2015

Dr. Kenneth Magel
Computer Science

Re: IRB Certification of Exempt Human Subjects Research:
Protocol #SM16077 , "A Metric Model to Predict Navigation Disorientation on Screen Layout Structure and Web Page Complexity"

Co-investigator(s) and research team: Abdulaziz Attaallah

Certification Date: 10/12/15 Expiration Date: 10/11/18
Study site(s): NDSU
Sponsor: n/a

The above referenced human subjects research project has been certified as exempt (category # 2) in accordance with federal regulations (Code of Federal Regulations, Title 45, Part 46, Protection of Human Subjects). This determination is based on the original protocol submission with revised consent/recruitment materials (received 10/8/2015).

Please also note the following:

- If you wish to continue the research after the expiration, submit a request for recertification several weeks prior to the expiration.
- The study must be conducted as described in the approved protocol. Changes to this protocol must be approved prior to initiating, unless the changes are necessary to eliminate an immediate hazard to subjects.
- Notify the IRB promptly of any adverse events, complaints, or unanticipated problems involving risks to subjects or others related to this project.
- Report any significant new findings that may affect the risks and benefits to the participants and the IRB.

Research records may be subject to a random or directed audit at any time to verify compliance with IRB standard operating procedures.

Thank you for your cooperation with NDSU IRB procedures. Best wishes for a successful study.

Sincerely,

Kristy Shirley, CIP, Research Compliance Administrator

For more information regarding IRB Office submissions and guidelines, please consult http://www.ndsu.edu/research/integrity_compliance/irb/. This Institution has an approved FederalWide Assurance with the Department of Health and Human Services: FWA00002439.

INSTITUTIONAL REVIEW BOARD
NDSU Dept 4000 | PO Box 6050 | Fargo ND 58108-6050 | 701.231.8995 | Fax 701.231.8098 | ndsu.edu/irb

Shipping address: Research 1, 1735 NDSU Research Park Drive, Fargo ND 58102

Date Received

IRB Protocol #:

NDSU NORTH DAKOTA STATE UNIVERSITY
INSTITUTIONAL REVIEW BOARD
office: Research 1, 1735 NDSU Research Park Drive, Fargo, ND 58102
mail: NDSU Dept. #4000, PO Box 6050, Fargo, ND 58108-6050
p: 701.231.8995 **f:** 701.231.8098 **e:** ndsu.irb@ndsu.edu **w:** www.ndsu.edu/irb

Protocol Amendment Request Form

Changes to approved research may not be initiated without prior IRB review and approval, except where necessary to eliminate apparent immediate hazards to participants. Reference: *SOP 7.5 Protocol Amendments.*

Examples of changes requiring IRB review include, but are not limited to changes in: investigators or research team members, purpose/scope of research, recruitment procedures, compensation strategy, participant population, research setting, interventions involving participants, data collection procedures, or surveys, measures or other data forms.

Protocol Information:

Protocol #: **SM16077** Title: **A Structural Complexity Model for Web Interfaces**

Review category: ☒ Exempt ☐ Expedited ☐ Full board

Principal investigator: **Kenneth Magel** Email address: **kenneth.magel@ndsu.edu**
Dept: **Computer Science**

Co-investigator: **Abdulaziz Attaallah** Email address: **abdulaziz.attaallah@ndsu.edu**
Dept: **Computer Science**

Principal investigator signature, Date: Kenneth Magel (email) 10/12/2016

In lieu of a written signature, submission via the Principal Investigator's NDSU email constitutes an acceptable electronic signature.

Description of proposed changes:

1. Date of proposed implementation of change(s)*: **10/6/2016**
   ** Cannot be implemented prior to IRB approval unless the IRB Chair has determined that the change is necessary to eliminate apparent immediate hazards to participants.*

2. Describe proposed change(s), including justification:
   Recruitment methodology: Dr. Kenneth Magel will deliver verbally an invitation to the study in the classes that he lectures
   Compensation: Dr. Kenneth Magel will compensate each participant with ten points in current courses which are taken with him
   Consent: The consent will be provided on the first page of the surveys as an electronic version, and participants by pressing a "continue button" they are informed that they are consenting to the study

(attached a copy of the consent)
The Surveys' amendments are:

- **Reconfiguring the structure of the surveys**
- **Rephrasing some questions and terminologies**
- **Including some graphical examples and descriptions**

to fulfill the research hypothesis.
(attached a copies of the surveys)
Non-research alternative: Graphical User Interface (GUI) evaluation assignment, which will be given to students who are not willing to participate in the study

3. Will the change(s) increase any risks, or present new risks *(physical, economic, psychological, or sociological)* to participants?
   ☒ No
   ☐ Yes: *In the appropriate section of the protocol form, describe new or altered risks and how they will be minimized.*

4. Does the proposed change involve the addition of a vulnerable group of participants?
   Children: ☒ no ☐ yes – include the *Children in Research* attachment form
   Prisoners: ☒ no ☐ yes – include the *Prisoners in Research* attachment form
   Cognitively impaired individuals: ☒ no ☐ yes*
   Economically or educationally disadvantaged individuals: ☒ no ☐ yes*

   **Provide additional information where applicable in the revised protocol form.*

5. Does the proposed change involve a request to waive some or all the elements of informed consent or documentation of consent?
   ☒ no
   ☐ yes – Attach the *Informed Consent Waiver or Alteration Request.*

6. Does the proposed change involve a new research site?
   ☒ no
   ☐ yes

If information in your previously approved protocol has changed, or additional information is being added, incorporate the changes into relevant section(s) of the protocol. Draw attention to changes by using all caps, asterisks, etc. to the revised section(s) and attach a copy of the revised protocol with your submission. *(If the changes are limited to addition/change in research team members, research sites, etc. a revised protocol form is not needed.)*

Impact for Participants (future, current, or prior):

1. Will the change(s) alter information on previously approved versions of the recruitment materials, informed consent, or other documents, or require new documents?

☐ No

☒ Yes - [attach] attach revised/new document(s)

2. Could the change(s) affect the willingness of *currently* enrolled participants to continue in the research?

   ☒ No

   ☐ Yes - describe procedures that will be used to inform current participants, and re-consent, if necessary:

3. Will the change(s) have any impact to *previously* enrolled participants?

   ☒ No

   ☐ Yes - describe impact, and any procedures that will be taken to protect the rights and welfare of participants:

----------FOR IRB OFFICE USE ONLY--------

Request is: ☒ **Approved** ☐ **Not Approved**		
Review: ☒ **Exempt, category#:** 2b	☐ **Expedited method, category #** ____ ☐ **Expedited review of minor change**	☐ **Convened meeting, date:** ____
IRB Signature: Kristy Shirley		**Date:** 10/18/2016
Comments:		