EVALUATING USABILITY EVALUATIONS

A Thesis

presented to

the Faculty of California Polytechnic State University

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

Allen Dunlea

June 2013

COMMITTEE MEMBERSHIP

TITLE:	Evaluating Usability Evaluations
AUTHOR:	Allen Dunlea
DATE SUBMITTED:	June 2013
COMMITTEE CHAIR:	Franz Kurfess, Ph.D. Professor, Computer Science
COMMITTEE MEMBER:	David Janzen, Ph.D. Associate Professor, Computer Science
COMMITTEE MEMBER:	Gene Fisher, Ph.D. Professor, Computer Science

Abstract

Evaluating Usability Evaluations

Allen Dunlea

We live in an age when consumers can now shop and browse the web using hand-held devices. This means that competitive companies need to have a website to represent their brand and to conduct business [20, 22]. E-commerce sites need to pay special attention to the usability of their sites, since it has such an impact on how potential costumers view their brand [4].

Jakob Nielsen defines usability as a "quality attribute that assesses how easy user interfaces are to use" [25]; he separates usability into five quality components: **learnability, efficiency, memorability, errors and satisfaction**. The current standard for testing usability involves having a number of users physically use a site in order to determine where they have trouble [10]. This kind of usability testing can be time consuming and costly [24].

In order to mitigate some of these costs, many tools are being developed to help automate the process [5]. However, many automated tools evaluate only one of the five components, or simply look for errors. In an attempt to increase the reliability and scope of such testing, this paper investigates the effectiveness of automated usability evaluators and proposes methods for future researchers to test them. Specifically, this paper details an experiment performed to test the some freely available usability evaluators against more traditional usability evaluations. The experiment attempts to determine whether automatic usability evaluations might be used as a cheaper alternative to more traditional usability evaluations.

Acknowledgements

Thank you Abigail Dunlea, for your patience; Dr. Franz Kurfess, for your guidance; Dr. David Janzen and Dr. Gene Fisher for serving as committee memembers for my defense; Tommy Merrill and Ryan Palo, for your support.

Also a special thank you to my parents and my parents away from home, Linda Garcia-Inchausti and the Marist Priests, Fr. Kevin Duggan and Fr. John Ulrich.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Rosenbaum [28] defines usability as the ease with which a user can learn and use a product. Many software developers agree that usability is one of the most important aspects of any software project [6]. This is even truer for websites, especially those where business is conducted [4]. If a user cannot navigate a website then he or she may not be able to buy a product. The very nature of a website makes usability essential; although software is generally concerned with "doing" things, websites are more concerned with "communicating" things [32]. Since communication is often the major concern of a website, it makes sense that usability would be a more important factor. The focus of this thesis is therefore on the usability of websites.

For websites, usability is defined as the ease with which users (visitors to a web page) can learn and use a website. There are number of methods for determining the usability of a website while it is being developed, but most of them are time consuming and can have confusing results. This paper investigates the work being done to automate the usability evaluation process. Additionally, I propose a generalized testing method that can be used to determine the effectiveness of any usability evaluators. This work is done to answer the underlying question: "Can computers measure usability as well as humans?"

The rest of this section is organized as follows: Section 1.1 describes the current standard for determining usability and usability evaluations. Section 1.2 describes some of the theoretical benefits that should come from switching to an automated approach. Finally, Section 1.3 gives a more in depth definition of usability and explains some of the hurdles that automated usability evaluations have to overcome.

1.1 Usability Evaluations

The current standard for testing usability involves having a number of users investigate a site in order to determine where they have trouble [10]. Often, this method involves actually watching users interact with a website and then determining what kinds of usability errors exist from their problems and comments. Performing usability tests this way can be time consuming and costly [24].

Alternatively, usability can be inspected by consultants using heuristics. Usability inspection methods of this form are generally cheaper and faster to do, but can introduce unreliable results [9], as these methods are usually performed by only one or a few evaluators who are using a small set of heuristics (10 in some cases). Jakob Nielsen reports that "it takes 39 hours to usability test a website the first time you try. This time estimate includes planning the test, defining test tasks, recruiting test users, conducting a test with five users, analyzing results, and writing the report" [26].

Usability testing and heuristics can be used to evaluate sites for qualitative usability or for a quantitative measure of usability [19]. Qualitative measuring is generally used to find specific errors or problems to correct. Quantitative measurements of usability are more useful for comparisons. For example, if a web designer wanted to compare the usability of their site to competitors. In this thesis, quantitative usability testing is used to compare a large number of websites against each other. These tests are performed by non-experts using a rubric (similar to using a heuristic). The entire process is described in Section 3.

1.2 Benefits of Automation

In order to mitigate some of the costs of usability evaluations, many tools are being developed to help automate the process [5]. "Automated usability tools can help save time and money in design and user testing, improve consistency and quality of site design, and improve the systematic application standards of usability" [7].

There are number of benefits that can be realized by using an automated usability system over more traditional usability evaluations [11]:

- **Reduced Costs:** Usability evaluations can be expensive to perform. Hours can be spent designing tests and if expert evaluators are used they will cost money as well. Doing tests early in the development process can reduce costly changes later in development. Automated processes can be used early and often to detect errors when they are the easiest to correct.

- **Competitive Edge:** Human investigations can take a long time. If a web designer is trying to get a product to market as quickly as possible, they will want to do testing as quickly as possible, as well. Automated tests can be performed faster than human investigations and can allow web designers to detect problems in their webpages much faster.

- **Regression Testing:** When developers create new iterations of a product it is useful to make sure that new errors or problems are not introduced into the system. It is becoming increasingly important for developers to do usability tests in this way. Having an automated test that will give consistent results can go a long way in making sure that a website is being improved over time.
- **Comparative Testing:** At times, a designer might have multiple websites that they are considering for deployment. Automated testing makes it easier for developers to test these sites side by side and make useful comparisons between the two.
- **Meet Demands:** An automated tool would be better able to meet the large demand from the multitudes of web designers and developers. Expert usability evaluators cannot meet these demands, especially for smaller projects. A well-developed automated tool would likely be very popular among web developers.

Most of the benefits of automation stem from the efficiency of those processes. Automated tests can run very quickly and very often. Once they are developed or bought they cost virtually nothing to run. The major concern for automated tests is accuracy. The focus of this thesis is on how to evaluate the correctness of automated usability tests.

1.3 Nielsen's Quality Components

Jakob Nielsen, a usability consultant, is largely considered to be a usability guru. He defines usability as a "quality attribute that assesses how easy user interfaces are to use" [25]. He feels that usability cannot be expressed as just one idea. Usability encompasses the entire user experience which is comprised of many different parts. Nielsen separates usability into five quality components:

- **Learnability:** How fast can a user learn to use a website?
- **Efficiency:** How fast can a user perform tasks on a website?
- **Memorability:** How easily can users reestablish proficiency when using a site after not using it for an extended period of time?
- **Errors:** How severe and regularly are errors encountered for users visiting a website?
- **Satisfaction:** How pleasant is the website?

These factors can be difficult to measure since they are all inherently subjective in nature [6, 31]. People often make decisions based only on subjective measurements and this is especially true for E-commerce. Since usability is subjective in nature, user testing or usability evaluations are commonly used for testing these metrics.

This thesis uses the following roadmap to show how online evaluators can be used to calculate usability: I begin by assuming that usability is expressed by Nielsen's Quality Components that are described in this section. Later, in Section 2.4 I argue that these Quality Components are embedded in the Rubric that is used in the experimentation process. During the experiment, users grade websites according to the rubric (which I already likened to the Quality Components). Finally, the user evaluations are compared to scores from online evaluators. By showing that online evaluators can approximate the scores from the user evaluations, I assert that they can approximate the Quality Components. Furthermore, if they can approximate the Quality Components, then they should also be able to approximate usability.

The rest of this paper is organized as follows: Section 2 describes the proposed process for testing the effectiveness of online usability evaluators. An example of a performed test using the proposed methods is described in Section 3. Section 4 is a record of the results of that experiment and a discussion of the specific usability evaluators that were tested. Section 5 discusses the effectiveness of the experiment design from Section 2 and comments on how well it was carried out in this thesis. Finally, Section 6 discusses implications for further research.

Chapter 2

Design

There are two types of usability evaluations: qualitative and quantitative. Qualitative usability evaluations identify errors and other problems on websites. This is usually done to help website developers improve the quality of a website. Quantitative usability evaluations give a website a numerical score for usability. Rather than providing a list of problems to correct. Quantitative evaluations are useful for making comparative statements about websites. For example, a web developer might want to know how usable their website was compared to a competitor's website.

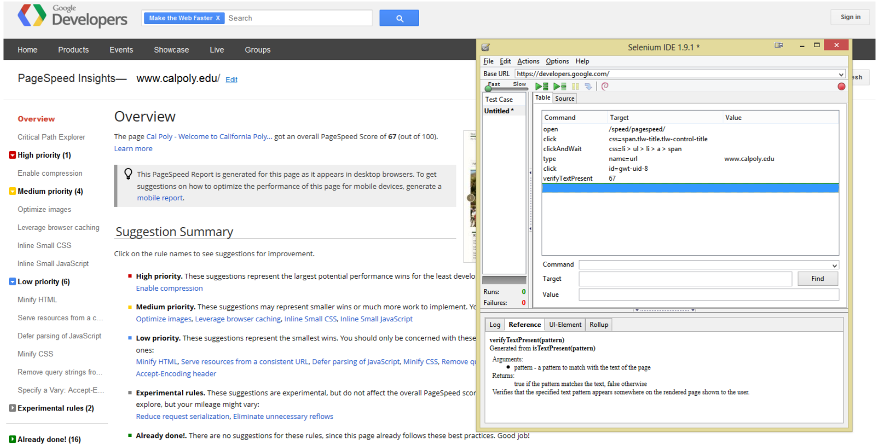
Trying to compare qualitative measurements from users and automated systems raises a number of difficulties. One problem relates to how the qualitative assessments are compared. Does a good automated evaluator find as many problems as a human does? What if it finds problems that are different than the problems the human evaluators find? Does that make the automated evaluator a stronger or weaker candidate? Comparing qualitative scores is too difficult to be informative.

For these reasons, quantitative usability evaluations are used to compare scores from users to scores from automated usability evaluations. The scores are normalized and correlations are investigated. Since there is only one usability score the numbers can be quickly compared. If an automated evaluator reports website scores similar to the scores from human evaluators, then I can trust the automated evaluator to be a replacement for determining website quality.

2.1 Automated Evaluators

To help web designers, numerous free and for-purchase automated evaluators are available online. The automated evaluators that are tested in this thesis are described in greater detail in Section 3.1, but they all generally work the same way. An automated evaluator receives a URL as input (see Figure **??**) and then loads the site to test. Once it has tested a site, it provides a page with a score for the site and usually provides additional information explaining the given score and what can be done to improve it. For the purposes of this thesis, only evaluators that provided a single overall score were used.

Online usability evaluators are powerful because they can run their tests very quickly. Unfortunately, it is difficult to determine whether they are accurate. With enough data, human evaluator scores can be compared to scores from the online usability evaluators to verify their accuracy. The following section describes how an automated web browser (Selenium) can be used to speed the process of testing with the online usability evaluators.



PageSpeed Insights— www.calpoly.edu/ Edit

Overview
Critical Path Explorer
High priority (1)
Enable compression
Medium priority (4)
Optimize images
Leverage browser caching
Inline Small CSS
Inline Small JavaScript
Low priority (6)
Minify HTML
Serve resources from a c…
Defer parsing of JavaScript
Minify CSS
Remove query strings fro…
Specify a Vary: Accept-E…
Experimental rules (2)
Already done! (16)

Overview

The page Cal Poly - Welcome to California Poly... got an overall PageSpeed Score of **67** (out of 100).
Learn more

This PageSpeed Report is generated for this page as it appears in desktop browsers. To get suggestions on how to optimize the performance of this page for mobile devices, generate a mobile report.

Suggestion Summary

Click on the rule names to see suggestions for improvement.

- **High priority.** These suggestions represent the largest potential performance wins for the least develo
  Enable compression
- **Medium priority.** These suggestions may represent smaller wins or much more work to implement. Yo
  Optimize images, Leverage browser caching, Inline Small CSS, Inline Small JavaScript
- **Low priority.** These suggestions represent the smallest wins. You should only be concerned with these ones:
  Minify HTML, Serve resources from a consistent URL, Defer parsing of JavaScript, Minify CSS, Remove qu
  Accept-Encoding header
- **Experimental rules.** These suggestions are experimental, but do not affect the overall PageSpeed scor explore, but your mileage might vary:
  Reduce request serialization, Eliminate unnecessary reflows
- **Already done!.** There are no suggestions for these rules, since this page already follows these best practices. Good job!

Base URL https://developers.google.com/

Command	Target	Value
open	/speed/pagespeed/	
click	css=span.tlw-title.tlw-control-title	
clickAndWait	css=li > ul > li > a > span	
type	name=url	www.calpoly.edu
click	id=gwt-uid-8	
verifyTextPresent	67	

Command
Target
Value

Runs: 0
Failures: 0

verifyTextPresent(pattern)
Generated from **isTextPresent(pattern)**
Arguments:
- pattern - a pattern to match with the text of the page

Returns:
true if the pattern matches the text, false otherwise
Verifies that the specified text pattern appears somewhere on the rendered page shown to the user.

Figure 2.1: Recording Tests with Selenium

2.2 Selenium

Selenium is an open source project that provides a framework for building regression tests for online applications. “Selenium is a suite of tools to automate web app testing across many platforms. Selenium runs in many browsers and operating systems and can be controlled by many programming languages and testing frameworks” [1]. Selenium provides an interface for users to record and subsequently run tests in Firefox, Chrome, Internet Explorer or Safari. Users can use the code generated by Selenium to perform their own tests or they can give the code to one of the many distributed systems that Selenium uses to run tests.

WebDriver has been added to the Selenium open source project. WebDriver allows usability testers to create tests to be run on many different browsers. “WebDriver aims to interact with a given Web browser as a person would; for example, keystrokes and mouse clicks are generated at the operating-system level rather than being synthesized in the Web browser.” [15] By mimicking human interactions with the browser, WebDriver is able to quickly perform simple usability tests, similar to ones that I are trying to create.

Selenium was used to create Java code that uses WebDriver to visit the online evaluators and then uses the evaluators to test websites. Figure 2.2 shows sample code that was generated by Selenium. In the method score, the first line of code tells the WebDriver to load the online evaluator front page. In the sample code Google's Pagespeed is being loaded. After the page is loaded WebDriver clears the "url" element (in this case it is a text field). It then enters in the target URL (the website being tested) in the text field. Once the URL is entered the "analyze" button is clicked. The "analyze" button has an ID of "gwt-uid-8" which is used to find the element on the web page. The code just described was all generated by Selenium. The code for getting the score was more difficult to determine. This involved searching the score page for HTML that could be used to locate the overall score. Each online evaluator was different in terms of identifying the overall score. Google's Pagespeed was the simplest since it had a unique class identifier that surrounded the score ("GCT5URKM3" in the example below).

All the code in quotations is individual to Google's Pagespeed. Similar code was generated for each online evaluator with the largest difference being the way in which the score was found, since each website placed their score on a different part of the webpage.

Unfortunately, if the placement of the score changes on the webpage (for example, if the web designer updated the layout of the scoring page), then the code for collecting the score could behave in unexpected ways or stop working altogether. Ideally, online evaluators would change infrequently enough that this scenario would not cause any major problems. There were no problems during the experimentation process, but it is possible that a more dynamic online evaluator would be unusable for this tool.

```
private WebDriver driver;

public String score(String url) {

//Go to the Pagespeed web site
driver.get("https://developers.google.com/pagespeed/");

//Enter the URL
driver.findElement(By.name("url")).clear();
driver.findElement(By.name("url")).sendKeys(url);

//Click the analyze button
driver.findElement(By.id("gwt-uid-8")).click();

//Find the score
WebElement ele = driver.findElement(By.className("GCT5URKBM3"));

return ele.getText();
}
```

Figure 2.2: Example Selenium Code

2.3 Rubric

Unfortunately, usability consultants are too expensive to give usability scores to the hundreds of websites required to make significant comparisons. Volunteers are often the most affordable option, but due to their lack of expertise, the volunteers generate lower quality results. To mitigate this problem, a rubric was developed that could allow volunteers to act as usability experts.

RubiStar, a free tool funded by a grant from the U.S. Department of Education, was used to make a rubric to grade the sites [2]. RubiStar was developed to assist educators with their project-based learning activities, and it was selected for use in this project because it contained a template rubric for web-design. The template was modified to fit the needs of the user evaluation. A summary of the rubric used can be seen below:

1. FONT: Is the font attractive and easy to read?

2. CONTENT: Is the purpose of the site clear?

3. INTEREST: Has the author made an effort to make the content interesting?

4. LAYOUT: Does the site have an attractive and usable layout?

5. NAVIGATION: Is it easy to navigate the site or do you often get lost?

6. BACKGROUND: Is the background appealing or does it distract?

7. COLOR CHOICES: Are the colors pleasant or do they distract?

8. LOAD TIME: Does the site load quickly or is it noticeably slow?

Using a rubric was very important since only non-expert evaluators participated. More popular websites were selected for this experiment to better control the nature of the content, but familiarity with websites could have led to unexpected bias from the volunteers. Imposing a rubric on the volunteers was an attempt to reduce possible bias. The complete rubric that the volunteers used is available in Table A of the Appendix.

2.4 Quality Components

While descriptive, Nielsen's Quality Components (**Learnability, Efficiency, Memorability, Errors, and Satisfaction**) are too abstract to be useful for providing quantitative scores for websites. The aforementioned rubric is used to provide human testers with a more concrete framework with which to evaluate websites. In the following sections I argue that there is an overlap between the Five Quality Components and the eight traits in the rubric.

2.4.1 Learnability

Learnability is a measure of how fast a user can learn to use a website [25]. A user should not have to spend a long time getting oriented before he or she can start interacting with a website in meaningful ways. If it takes too long for a user to learn to use a website, he or she may become discouraged, and even give up. To be effective, it is important for a website to be intuitive and easy to use.

When asked to grade the Content of a website, volunteers were asked if the purpose of the website was clear. If the purpose of the website was unclear, they gave the website a lower grade. Content is a strong indicator of Learnability because if the purpose of a website was clear, then volunteers were able to quickly understand the website and how to use it. If the purpose of a website was vague or unclear, then it was far more difficult for volunteers to learn to use the website.

Navigation and Layout also correspond to the Learnability of a website. Both Navigation and Layout are concerned primarily with where objects are located on a webpage. If objects were not where expected, volunteers gave the website a lower score. Navigation and Layout correspond to Learnability because if objects are in strange or non-obvious places, then a website will be harder to learn to use.

2.4.2 Efficiency

Efficiency is a measure of how quickly a user can perform tasks on a website after he or she has established proficiency (users have established proficiency with a website when they can use a website in its intended function with few or no errors) [25]. Just because a website is easy to learn to use does not mean that it can be used quickly. Suppose a news website located all of its articles on a single webpage. It would probably be very easy to learn to use this hypothetical website (probably a just a lot of scrolling), but it would be extremely inefficient to use. Users would have to depend on the find feature of the browser to use the website. Additionally, the website would require a long time to load, especially if the articles contained pictures or video. Users should be able to navigate to the information they are interested in as quickly as possible. For websites, this involves mainly the load time of the pages and the number of clicks it takes to perform a task. Generally, efficiency is measured by giving a user a task and then timing how long it takes to complete that task. While a survey seems to be a nonobvious way to measure efficiency, this Quality Component has been represented in the survey traits.

To measure Efficiency in the survey, Load Time was added as a trait. Unfortunately, Load Time turned out to be a fairly ineffective measurement since most users were not able to notice significant differences in Load Time between webpages. I suspect that this is largely due to advances in modern technology that make discrepancies in Load Time less noticeable.

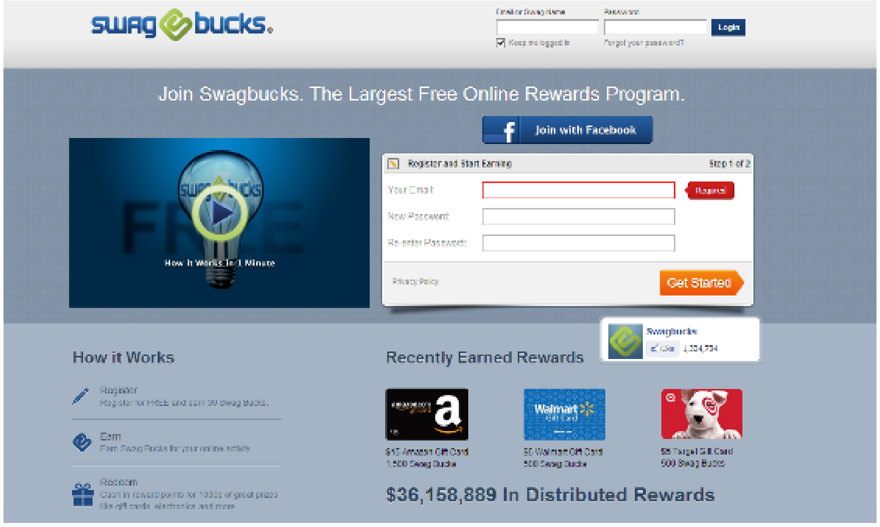
Navigation captures the spirit of Efficiency. In order to use a website efficiently, a user must be able to move between pages quickly. While taking the survey, if users felt lost when navigating a website, they gave that website a lower Navigation score. If users felt that it was easy to move between pages of a website, then they gave that website a higher score. To some degree, the Efficiency of a website was measured by the survey scores relating to Load Time and Navigation.

2.4.3 Memorability

Memorability expresses how difficult it is for a user to reestablish proficiency with a website after not using it for an extended period time [25]. Memorability is similar to Learnability in that both are concerned with how quickly a user is able to establish proficiency with a website. Ideally, a website that is easy to learn should also be a memorable site; however, a user should not have to feel a novice each time he or she uses that website. This is where Learnabiliy and Memorability differ. During the survey, users were asked to grade websites after only a single visit. Although, it might not be clear how Memorability was measured with one viewing, I argue that this is not an issue. I believe there are elements of a site which make it easier or harder to relearn and that these elements can be evaluated with only one viewing.

Similar to Learnability, Layout and Navigation should both capture some of the spirit of Memorability. If it is not obvious to the user how to operate a website the first time, it will probably not be obvious the second time, either. It should be easy for a user to locate elements of interest on a website. The authors of [30] recommend that website developers "try to maximize the memorability by creating logical steps and consistent design throughout the site." Furthermore, the authors of [23] noted that imageability (the "shape, color, or arrangement which facilitates the making of vividly identified, powerfully structured, highly useful mental images of the environment" [23]) led to designs that were more memorable.

The Background and Color Choices of a website also can be used as a measure for Memorability (as noted by [23]. A well-crafted background should leave an impression on the user that helps differentiate one website from another. Consider



Join Swagbucks. The Largest Free Online Rewards Program.

How it Works

Register
Register for FREE and earn 30 Swag Bucks.

Earn
Earn Swag Bucks for your online activity

Redeem
Cash in reward points for 1000s of great prizes like gift cards, electronics and more

Recently Earned Rewards

$15 Amazon Gift Card
1,500 Swag Bucks

$5 Walmart Gift Card
500 Swag Bucks

$5 Target Gift Card
500 Swag Bucks

$36,158,889 In Distributed Rewards

Figure 2.3: Swag Bucks Homepage

the home page shown in Figure 2.3. The different colors in the background help the user to separate the different parts of the page. A returning user should quickly recognize which "section" they want to be looking at just by the colors. While completing the survey, users graded the Background of a website; they gave higher scores to backgrounds that added to the theme or purpose of a website. Strong backgrounds and color choices contribute to the Memorability of a website. While weak backgrounds will not necessarily not detract from the Memorability of a website, it certainly will not make it better.

If the web designer made an exceptional attempt to ensure that that the content of his or her website was interesting, then it stands to reason that the website would also be memorable. If a website is interesting, users will remember it on the second encounter. This would allow them to reestablish proficiency more quickly than for a website they did not find interesting.

The Layout, Navigation, Background, Color Choices and Interest rubric scores were used to ensure that Memorability was measured by the survey. Due to the nature of this research, a longer term study of memorability is not possible, and these factors were used to approximate it.

2.4.4 Errors

Errors encompass any expected or unexpected errors that might cause difficulties when interacting with a site [25]. These errors include a page not loading, a broken link, or anything that does not work as intended. While errors can be caused by poor programming, they can also be caused by human mistakes. No matter how usable a developer makes a website, humans will always find a way to do something they should not have done. It is important that the designers make websites that can handle all of these types of errors. A user should not be lost or have to start over because he or she accidentally clicked the wrong link. The user should not be punished for clicking the wrong button (especially considering that it might be the developer's fault that the button's meaning was not obvious to the user). Nielsen's Errors not only measures how severe and often errors occur, but also how gracefully they can be recovered from.

During the survey, Errors was a difficult Quality Component for users to measure with only the rubric that was provided. It was not feasible to orchestrate errors for all the websites that were graded. I suspect that lower scores were given to sites when errors were encountered. Since users did not spend long periods of time with a website, if any errors were encountered, then it negatively influenced the volunteer's perception of the website. I believe that volunteers would score both Navigation and Load Time lower when errors were encountered. If a volunteer quickly got lost using a site by clicking incorrect links then he or she would probably reflect that in their Navigation score. Also, if a page failed to load because of errors, I suspect that the user would have given the website a lower Load Time score.

2.4.5 Satisfaction

Satisfaction is a measure of the visual quality of a website or how pleasant the site is to the users [25]. Users should find the visual aspects of a website appealing. If a user does not find a website visually appealing, he or she will not want to shop or spend time there. Having a visually appealing website also makes users more willing to forgive other failings that a website might have.

If Errors was hard for humans to grade but easy for computers to grade, then Satisfaction is the exact opposite. A website that is visually satisfying is very easy for a human to recognize. Computers have a much harder time determining what is visually appealing to humans, and even the best algorithms have a difficult time measuring the beauty of a webpage. The web designers are trying to impress the humans, not the computer.

That being said, most of the survey questions map better to Satisfaction than to any of the other Quality Components. Fonts, Background and Color Choices are the most obvious items that pertain to Satisfaction. All of these items specifically asked the graders to evaluate the visual appeal of a website. A good collection of readable fonts makes a webpage easier to digest. A good background can set the mood for a website and can make the purpose of the site much clearer, whereas a poor background can make text hard to read and can be even more distracting than a bad font. Color Choices encompasses many of the problems that might be apparent in Fonts or Background, but I felt that it was useful to consider it separately, as well, since Color Choices can refer to many

	Learnability	Efficiency	Memorability	Errors	Satisfaction
Background			X		X
Color Choices			X		X
Content	X				
Fonts					X
Interest			X		X
Layout	X		X		
Load Time		X		X	
Navigation	X	X	X	X	

Table 2.1: Cross Chart of Quality Components

other elements of a webpage.

What might be less obvious is how Satisfaction might correlate to Interest. During the survey, users were asked to give high Interest scores to websites that made the content of their site interesting for the intended users. One of the best ways to make content interesting is to make sure that it is presented in a visually appealing way. If a web designer has not put the effort into making the content interesting, then they probably have put little work into making the website visually appealing. Survey scores for Font, Background, Color Choices and Interest should all correspond to the Satisfaction Quality Component.

Chapter 3

Experiment

In order to affirm the effectiveness of the testing process described in Section 2, a small-scale experiment was performed.

First, six freely available online evaluators were selected to test. The six online evaluators were tested using 100 websites as input; these websites were selected from Alexa's Top 500 Sites [3]. Alexa is a Web Information Company that collects data about the most frequently searched terms and the most popular websites online. Alexa ranks these sites based on popularity. The websites were selected from Alexa's top 500 sites so that more familiar sites would be selected. It seemed less likely that content would be an issue with more familiar sites than with less familiar sites. A number of sites were removed because of inappropriate content, but otherwise the top 100 sites that could be tested were selected.

Only English-language websites were selected. I investigated using international sites to remove bias from the evaluators but decided that it was not a good idea. Popular Asian websites generally had a very different layout than what English users would be accustomed to. I suspect that this is partially due to cultural differences, but even the characters of the language forced the sites to have very different layouts. My fear was that these sites would receive much lower ratings than Western sites.

To test the effectiveness of the online evaluators, human evaluators were used. To get more reliable results, multiple users graded each website. Their scores were then averaged to give the most accurate score possible. The scoring procedure is described in greater detail in Section 4. The rest of this section describes the online evaluators and how the human testing was performed.

3.1 Online Evaluators

Six freely available online evaluators were identified to use for the experiment. These graders were chosen because they all gave an "overall" score for the site instead of a number of scores. To reduce costs, free evaluators that were available online were chosen. To simplify comparisons between user evaluations, only graders that had a maximum score were selected. (For instance, some graders reported the number of errors of a website or some other figure with no obvious maximum value.) These graders were not selected since the distribution of grades was harder to interpret and was largely dependent on the size of the web page.

Many other online evaluators could not be used, only these six graders were consistent enough to be used in the testing process. Some of the other evaluators would fail to grade a site and needed to be restarted to get a useful number. The automated system was not prepared to handle this situation, so some graders had to be thrown out.

In the following sections, the individual evaluators that were used are described in greater detail.

3.1.1 PageSpeed

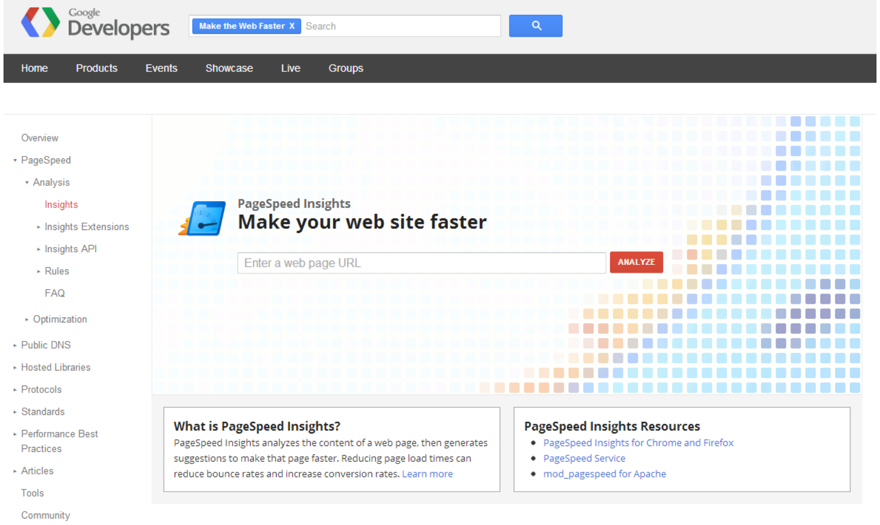
Google's PageSpeed provides a score (out of 100) that measures the **expected** efficiency of a given webpage. "[PageSpeed] runs a number of diagnostic tests against a web page, and analyzes the page's performance on a number of 'rules' that are known to speed up page load time. The rules are based on general principles of web page performance, including resource caching, data upload and download size, and client-server round-trip times. They examine factors such as web server configuration, JavaScript and CSS code, image file properties, and so on. For each rule, PageSpeed gives a general score, using a simple red-yellow-green grading scheme, and suggests specific techniques for correctly implementing each rule. It also provides some automatic optimization of external resources included on a page, such as minifying JavaScript code and compressing images" [13].

In this project, only the overall score was used to test PageSpeed against human user scores, but it would be simple to make the testing suite include the additional scores that are reported by PageSpeed.

3.1.2 Yottaa

Yottaa is a cloud service that aims to make a website faster, safer and more reliable [18]. Yottaa optimizes sites by controlling traffic to a website. It optimally routes packets over the internet to reduce load on a web server and deflects malicious traffic to a site.

In order to sell more subscriptions, Yottaa provides a free scoring service that



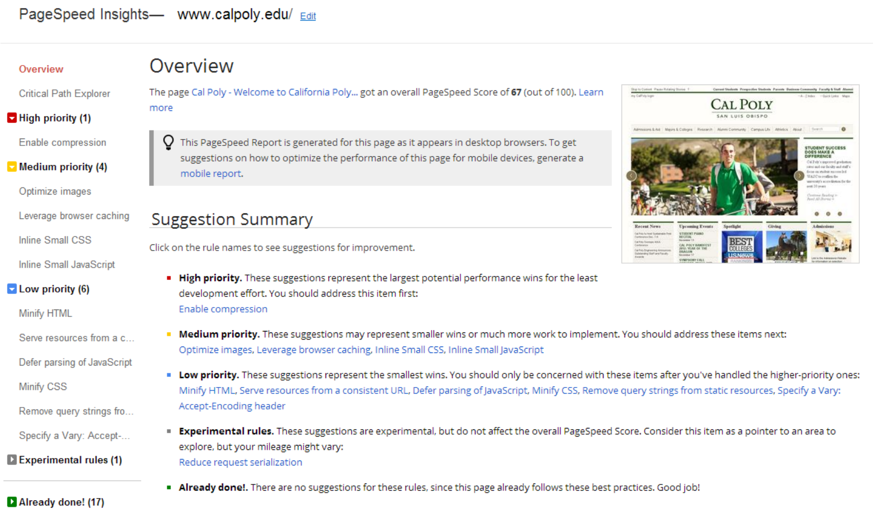
What is PageSpeed Insights?

PageSpeed Insights analyzes the content of a web page, then generates suggestions to make that page faster. Reducing page load times can reduce bounce rates and increase conversion rates. Learn more

PageSpeed Insights Resources

- PageSpeed Insights for Chrome and Firefox
- PageSpeed Service
- mod_pagespeed for Apache

Figure 3.1: Pagespeed Online Evaluator



Overview

The page Cal Poly - Welcome to California Poly... got an overall PageSpeed Score of **67** (out of 100). Learn more

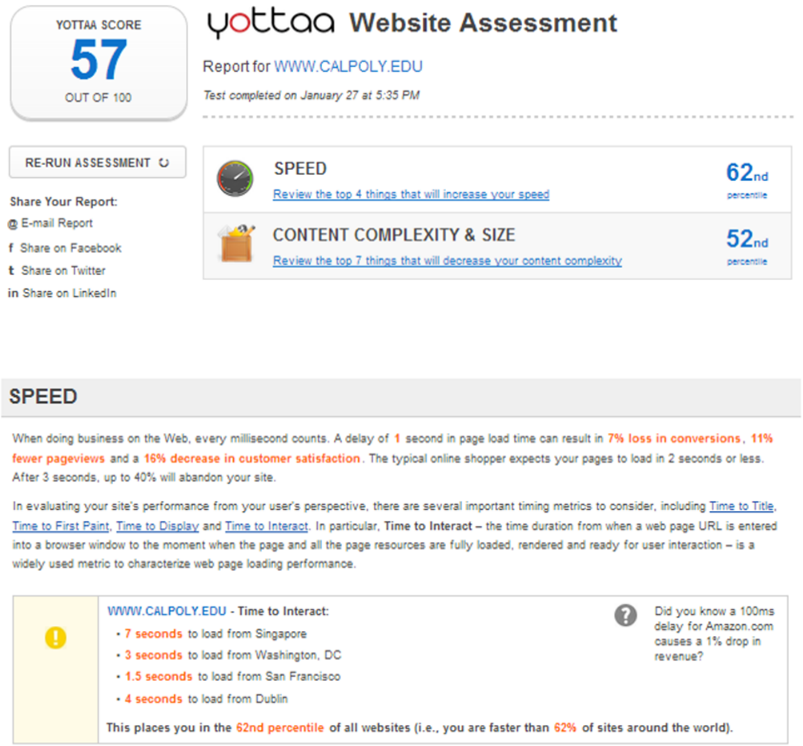
This PageSpeed Report is generated for this page as it appears in desktop browsers. To get suggestions on how to optimize the performance of this page for mobile devices, generate a mobile report.

Suggestion Summary

Click on the rule names to see suggestions for improvement.

- **High priority.** These suggestions represent the largest potential performance wins for the least development effort. You should address this item first:
  Enable compression
- **Medium priority.** These suggestions may represent smaller wins or much more work to implement. You should address these items next:
  Optimize images, Leverage browser caching, Inline Small CSS, Inline Small JavaScript
- **Low priority.** These suggestions represent the smallest wins. You should only be concerned with these items after you've handled the higher-priority ones:
  Minify HTML, Serve resources from a consistent URL, Defer parsing of JavaScript, Minify CSS, Remove query strings from static resources, Specify a Vary: Accept-Encoding header
- **Experimental rules.** These suggestions are experimental, but do not affect the overall PageSpeed Score. Consider this item as a pointer to an area to explore, but your mileage might vary:
  Reduce request serialization
- **Already done!.** There are no suggestions for these rules, since this page already follows these best practices. Good job!

Figure 3.2: Pagespeed Results Page



Website Assessment

Report for WWW.CALPOLY.EDU

Test completed on January 27 at 5:35 PM

SPEED — Review the top 4 things that will increase your speed — 62nd percentile

CONTENT COMPLEXITY & SIZE — Review the top 7 things that will decrease your content complexity — 52nd percentile

SPEED

When doing business on the Web, every millisecond counts. A delay of 1 second in page load time can result in **7% loss in conversions**, **11% fewer pageviews** and a **16% decrease in customer satisfaction**. The typical online shopper expects your pages to load in 2 seconds or less. After 3 seconds, up to 40% will abandon your site.

In evaluating your site's performance from your user's perspective, there are several important timing metrics to consider, including Time to Title, Time to First Paint, Time to Display and Time to Interact. In particular, **Time to Interact** – the time duration from when a web page URL is entered into a browser window to the moment when the page and all the page resources are fully loaded, rendered and ready for user interaction – is a widely used metric to characterize web page loading performance.

WWW.CALPOLY.EDU - **Time to Interact:**

- **7 seconds** to load from Singapore
- **3 seconds** to load from Washington, DC
- **1.5 seconds** to load from San Francisco
- **4 seconds** to load from Dublin

Did you know a 100ms delay for Amazon.com causes a 1% drop in revenue?

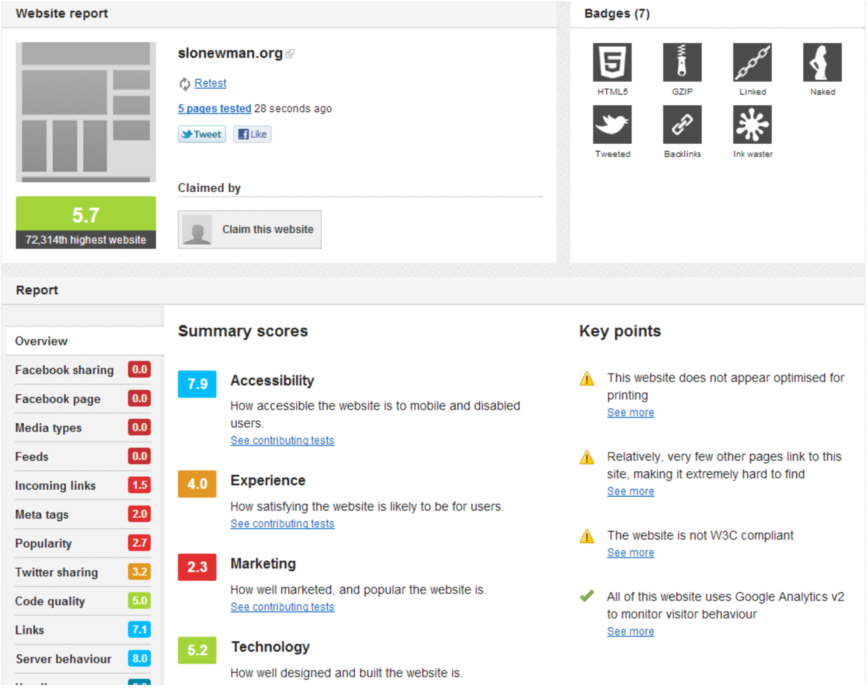
This places you in the 62nd percentile of all websites (i.e., you are faster than 62% of sites around the world).

Figure 3.3: Yottaa Website Assessment

measures your page load time and then shows how much faster your page load time could be using Yottaa. In addition to a page load time, Yottaa reports a Yottaa Score which is a score relative to all other websites that are tested with Yottaa. For example, a score of 57 would mean that the tested website was considered better than 56% of all websites [18].

3.1.3 Nibbler

Nibbler provides an overall score for a webpage, as well as four summary scores (sub-scores that combine to make the overall score) that include Accessibility, Experience, Marketing and Technology [29]. These scores all range from one to ten. Nibbler also reports a number of other scores, which are summarized by the



Website report

slonewman.org

5 pages tested 28 seconds ago

Claimed by

5.7

72,314th highest website

Badges (7)

Report

Overview	
Facebook sharing	0.0
Facebook page	0.0
Media types	0.0
Feeds	0.0
Incoming links	1.5
Meta tags	2.0
Popularity	2.7
Twitter sharing	3.2
Code quality	5.0
Links	7.1
Server behaviour	8.0

Summary scores

7.9 Accessibility

How accessible the website is to mobile and disabled users.

See contributing tests

4.0 Experience

How satisfying the website is likely to be for users.

See contributing tests

2.3 Marketing

How well marketed, and popular the website is.

See contributing tests

5.2 Technology

How well designed and built the website is.

Key points

⚠ This website does not appear optimised for printing

See more

⚠ Relatively, very few other pages link to this site, making it extremely hard to find

See more

⚠ The website is not W3C compliant

See more

✓ All of this website uses Google Analytics v2 to monitor visitor behaviour

See more

Figure 3.4: Nibbler Report Example

summary scores. Nibbler scores a number of items, including the existence of a Facebook page, Headings, and Meta tags.

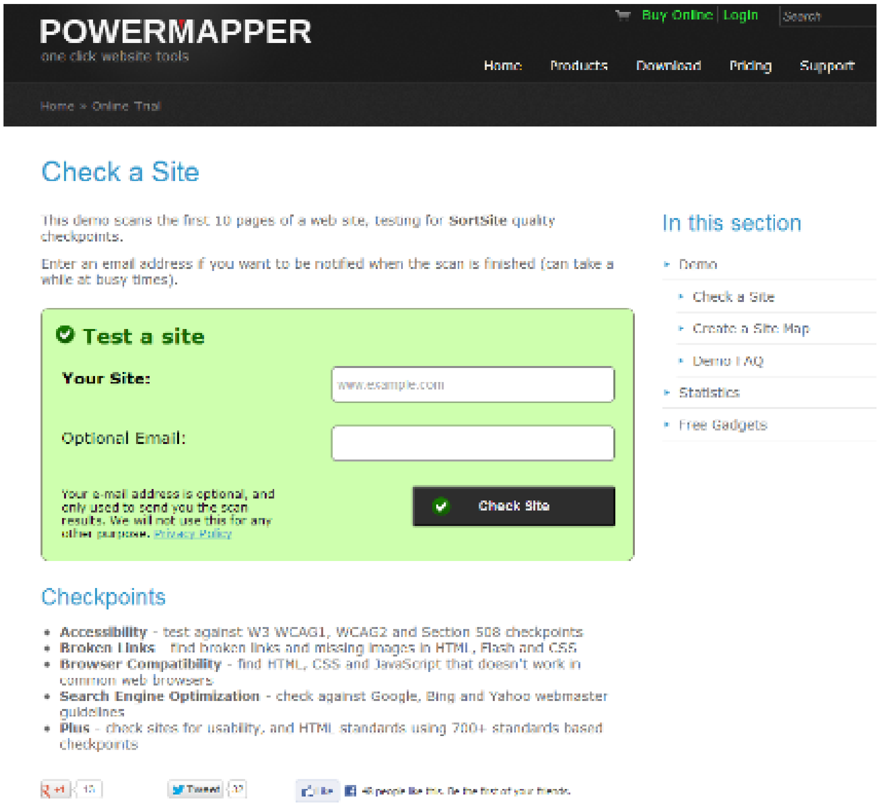
To make the grading a little more exciting and to encourage website developers to make better sites, Nibbler also includes badges that can be earned for a site. These badges can be tied to a user profile and then shared so that other users can see which badges other website developers have earned [29]. Figure 3.4 is an example of a website report that includes a summary of scores and badges. For this project, only the overall score was measured; it was unclear if there was any correlation between number of badges and results from the usability evaluation.

3.1.4 SortSite

Powermapper's SortSite creates graphical representations of site maps and helps users to identify problems with their website. "SortSite is a one-click web site testing tool used by federal agencies, Fortune 100 corporations and independent consultancies" [27]. SortSite tests a website for a number of different checkpoints, including:

- **Accessibility:** Test against W3 WCAG1, WCAG2 and Section 508 checkpoints
- **Broken Links:** Find broken links and missing images in HTML., Flash and CSS
- **Browser Compatibility:** Find HTML, CSS and JavaScript that doesn't work in common web browsers
- **Search Engine Optimization:** Check against Google, Bing and Yahoo webmaster guidelines
- **Plus:** Check sites for usability, and HTML standards using 450+ standards based checkpoints

Unlike the other online evaluators that were tested, SortSite tests the first ten pages of a website instead of only the first page. At the end of a scan, SortSite reports the percentage of webpages on which errors were found. To simplify testing, only that number was recorded.



Check a Site

This demo scans the first 10 pages of a web site, testing for **SortSite** quality checkpoints.

Enter an email address if you want to be notified when the scan is finished (can take a while at busy times).

Test a site

Your Site: www.example.com

Optional Email:

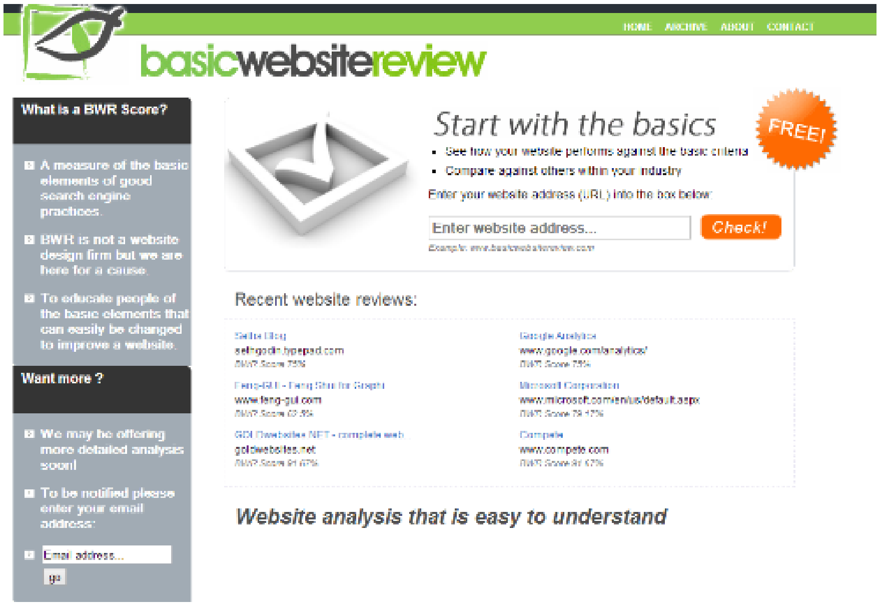
Your e-mail address is optional, and only used to send you the scan results. We will not use this for any other purpose. Privacy Policy

Checkpoints

- **Accessibility** - test against W3 WCAG1, WCAG2 and Section 508 checkpoints
- **Broken Links** - find broken links and missing images in HTML, Flash and CSS
- **Browser Compatibility** - find HTML, CSS and JavaScript that doesn't work in common web browsers
- **Search Engine Optimization** - check against Google, Bing and Yahoo webmaster guidelines
- **Plus** - check sites for usability, and HTML standards using 700+ standards based checkpoints

48 people like this. Be the first of your friends.

Figure 3.5: PowerMapper's SortSite



What is a BWR Score?

- A measure of the basic elements of good search engine practices.
- BWR is not a website design firm but we are here for a cause.
- To educate people of the basic elements that can easily be changed to improve a website.

Want more ?

- We may be offering more detailed analysis soon!
- To be notified please enter your email address:
- Email address...

go

Start with the basics

FREE!

- See how your website performs against the basic criteria
- Compare against others within your industry

Enter your website address (URL) into the box below:

Enter website address... Check!

Example: www.basicwebsitereview.com

Recent website reviews:

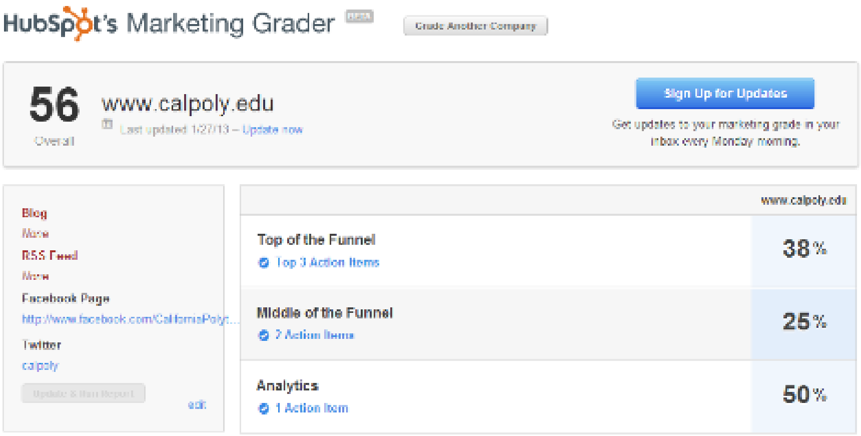
Seths Blog sethgodin.typepad.com BWR Score 75%	Google Analytics www.google.com/analytics/ BWR Score 75%
Feng-GUI - Feng Shui for Graphi www.feng-gui.com BWR Score 62.5%	Microsoft Corporation www.microsoft.com/en/us/default.aspx BWR Score 79.17%
GDI Dwebsites NET - complete web gdidwebsites.net BWR Score 91.67%	Compete www.compete.com BWR Score 91.67%

Website analysis that is easy to understand

Figure 3.6: Basic Website Review Front Page

3.1.5 Basic Website Review

Basic Website Review (BWR) claims that the score it reports is “a measure of the basic elements of good search engine practices” [12]. BWR evaluates a website with an emphasis on search engine optimization instead of focusing on something narrower, like website efficiency. “BWR was built as a statement, [sic] if your website cannot achieve a reasonable score with as few as ten factors or criteria then perhaps you need to reevaluate your online efforts” [12]. The score from BWR is based on ten parameters. Most of them are simple items, like the existence of a title and the count of H1 tags. As of the writing of this thesis, BWR was not operating correctly, so a complete list of items could not be provided.



56 www.calpoly.edu
Overall
Last updated 1/27/13 – Update now

Get updates to your marketing grade in your inbox every Monday morning.

Blog
None
RSS Feed
None
Facebook Page
http://www.facebook.com/CaliforniaPoly...
Twitter
calpoly

	www.calpoly.edu
Top of the Funnel Top 3 Action Items	38 %
Middle of the Funnel 2 Action Items	25 %
Analytics 1 Action Item	50 %

Figure 3.7: HubSpot's Marketing Grader

3.1.6 Marketing Grader

HubSpot [17] provides marketing strategies for online businesses. To demonstrate the effectiveness of their systems, they also developed a Marketing Grader that will grade websites based on how well the website is marketed. This might seem like a departure from usability, but evaluating Marketing Grader was a useful way to investigate the correlation between well-marketed websites and usable websites.

Marketing Grader [17] reports an overall score for each site (which was recorded for this experiment) and then reports actionable items that can help make a site more marketable. Everything from Tweets to the presence of analytic software is analyzed by Marketing Grader to score to each site.

3.1.7 Summary

The table below presents an overview of the information that was presented in the previous sections. Very few of the online evaluators claim to have anything

	What is Being Measured
Yotta	Predominately concerned with load time
Pagespeed	Concerned with load times Uses algorithms to look at potential problems in website code that could be adversely impacting rendering performance
Nibbler	An average of scores that include: Accessibility: How accessible the website is to mobile and disable users Experience: How satisfying the website is likely to be for users Marketing: How well marketed, and popular the website is Technology: How well designed and built the website is
Power Mapper	Measures the percent of pages with “issues” Broken links of other errors Accessibility problems Browser specific issues Compliance of legal issues Search engine issues W3C standards issues Usability issues
Basic Website Review	A measure of the basic elements of good search engine practices
Marketing Grader	A measure of well marketed a site is

Table 3.1: Summary of Online Evaluators

to do with measuring usability, but there is overlap. For instance, a website that takes a long time to load might be viewed as unusable by a user who wants to complete a task quickly.

3.2 The Human Evaluators

To gather potential evaluators, a Facebook event was set to happen on Saturday, June 30th, 2012. Volunteers were invited to this event so that they could help evaluate the websites. There was not actually any event that occurred, but it was useful to use Facebook's event mechanism to make friends and family aware of the need for help with the thesis evaluations. When potential evaluators viewed the description of the event they were greeted with the following text:

> Hi Everyone,

> I'm working on my Thesis for my Master's Degree in Computer Science on Website Usability. The Thesis looks at how accurately automated website graders can predict how visitors actually feel about a site.
>
> I've created a 30 minute survey to evaluate how real people feel about certain sites. I'm testing a lot of sites, so I broke them into ten groups. To take the survey, click the link with the last digit of the day of your birth (So if you were born on February 20, 1991, then you take test 10).
>
> Additional information will be provided on the first page of the survey.
>
> If you want to take more than one test you are welcome to, just please do not take the same test twice as that will sway my results.
>
> Thank you so much for your help.
>
> Allen Dunlea

These instructions were given to try and randomize which evaluators tested which sites and to make sure that each site was tested at least once. Even with these instructions, most users defaulted to evaluating the first survey that was provided. Having additional evaluators gave more precise scores, but did not necessarily give more accurate scores. Even websites with few evaluators were useful to compare online usability evaluators against.

Roughly 30 people (mainly consisting of friends and family) volunteered to participate in the survey research. Each website was graded a minimum of three (3) times, although several websites were graded 11 times. I invited volunteers to evaluate websites at their leisure, and the scores all came between mid-June to early July.

3.3 The Evaluation

Once they selected a survey, a user was taken to one of ten websites. Before he or she could take the survey, the user was presented with a consent form with information about the experiment and warnings about the potential nature of the websites (whose content could not be controlled). If they consented to the terms of the evaluation they were taken to a second page and were provided with the following instructions:

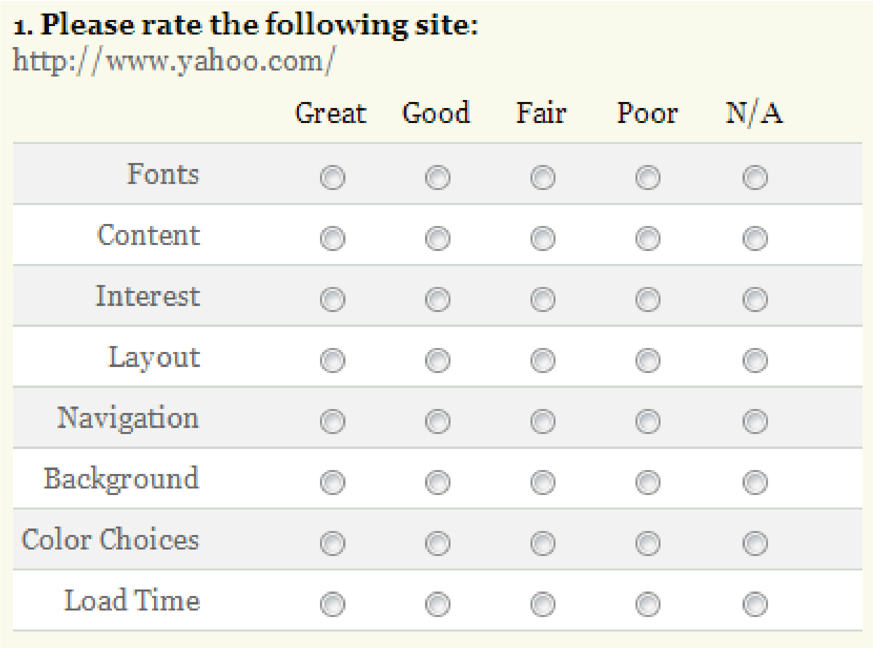
> You are being asked to evaluate each site on the following criteria:
>
> 1. FONT: Is the font attractive and easy to read?
> 2. CONTENT: Is the purpose of the site clear?
> 3. INTEREST: Has the author made an effort to make the content interesting?
> 4. LAYOUT: Does the site have an attractive and usable layout?
> 5. NAVIGATION: Is it easy to navigate the site or do you often get lost?
> 6. BACKGROUND: Is the background appealing or does it distract?
> 7. COLOR CHOICES: Are the colors pleasant or do they distract?
> 8. LOAD TIME: Does the site load quickly or is it noticeably slow?
>
> If you have questions about grading, please refer to the following rubric:
>
> https://docs.google.com/open?id=0B0SI2k7H6Ac7QWFVN0E2WDFEaHc.

Each survey had ten websites to be graded. Figure 3.8 shows an example of what an individual site's evaluation looked like for the user.

When an evaluator finished grading all ten websites, the grades were recorded in a Google Form, which was later processed to find average user scores.

To verify the effectiveness of the automated evaluators, the overall scores from the evaluators were compared to the **Total User Score** of each website. To calculate the Total User Score, the trait scores were averaged across users. For example: http://imgur.com/ (an image sharing site) received **Color Choices** scores of 3, 4, 4, and 3. So the average **Color Choices** score for http://imgur.

com/ is 3.50. Once all the average trait scores were determined, they were summed to create a Total User Score. Each website was graded on eight traits and each trait had a maximum score four (Great) so the maximum Total User Score was 32. Since the lowest score a trait could receive was a one (Poor) the minimum Total User Score a site could receive was 8.



1. Please rate the following site:
http://www.yahoo.com/

	Great	Good	Fair	Poor	N/A
Fonts	○	○	○	○	○
Content	○	○	○	○	○
Interest	○	○	○	○	○
Layout	○	○	○	○	○
Navigation	○	○	○	○	○
Background	○	○	○	○	○
Color Choices	○	○	○	○	○
Load Time	○	○	○	○	○

Figure 3.8: A Sample Evaluation

Chapter 4

Results

To test the quality of the evaluations the Cronbach's α was calculated for the rubric. Cronbach's α is a popular measure for determining internal consistency [16]. It is especially useful for determining the reliability of questionnaires. Generally, if the questions of a questionnaire (or in the case of this thesis, a survey) have a high Cronbach's α, then the questions are probably measuring the same underlying metric. Opinions of what a good Cronbach's α is varies between applications, but generally a score of .70 is considered minimally acceptable and a score of .90 or above is generally considered very high [14]. Unfortunately, results from a questionnaire are needed to compute Cronbach's α so it cannot be used to measure the effectiveness of a survey before it is used.

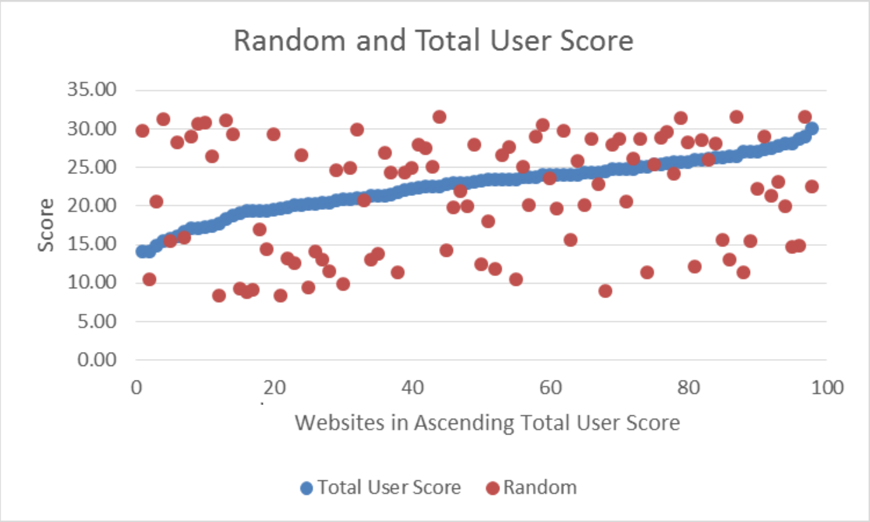
The Cronbach's α for the survey used in this thesis was **0.91**. This means that the survey performed well at measuring a single underlying trait. There is no way to verify that that the underlying trait was actually usability but every attempt was made to make sure that usability was actually measured. Having a Cronbach's α so high means that I can be confident in the survey results and that the individual questions were effective at measuring usability.

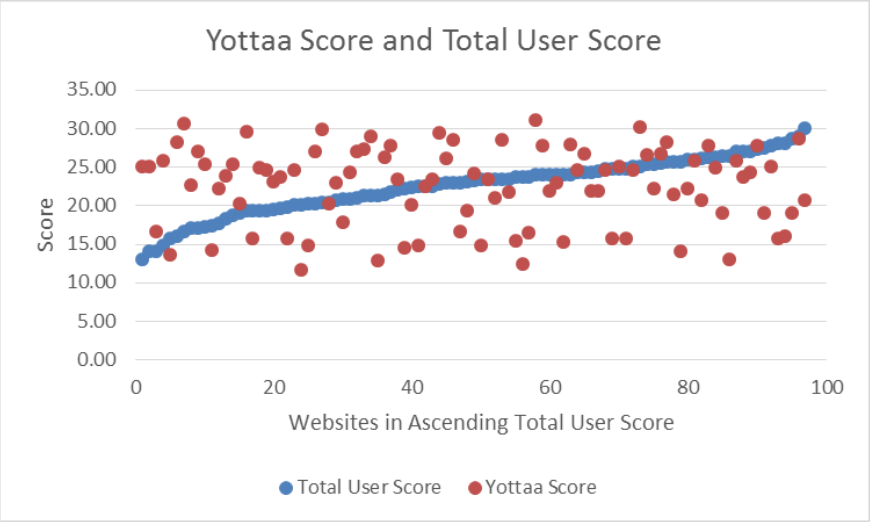
4.1 Online Evaluators

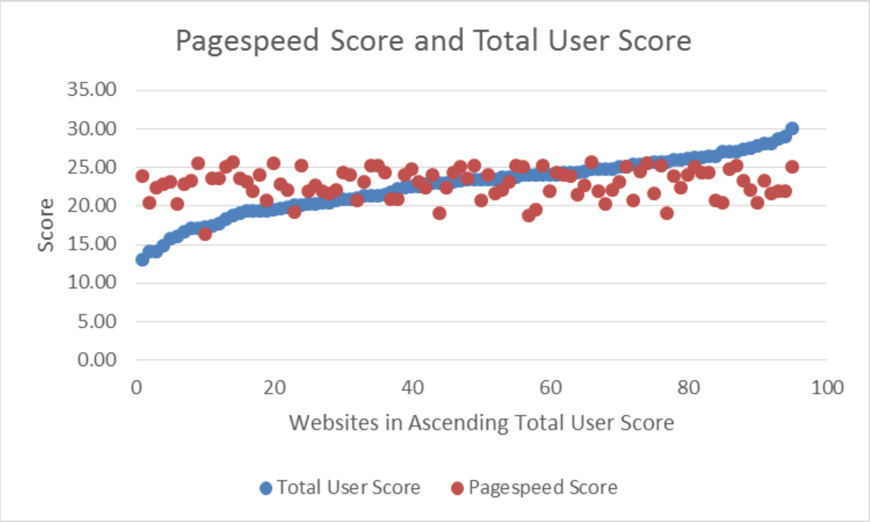
The online evaluators were graded by investigating how similar their scores were to the Total User Score. To accurately compare the scores of the online evaluators to the Total User Scores a multiplier and offset was applied to the scores from the online evaluators. This was done so that their scores would be normalized to the scores given by the users. For example, Yottaa reports a score from 0 to 100. The Total User Score ranges from 8 to 32 (24 possible values) so the multiplier applied to the Yottaa scores was .24. Doing so returns scores ranging from 0 to 24 so an offset still needed to be applied. It would have been easy enough to set the offset to 8 (making the scores range from 8 to 32) but that would have overlooked the possibility that Yottaa scores might be naturally higher than scores from users.

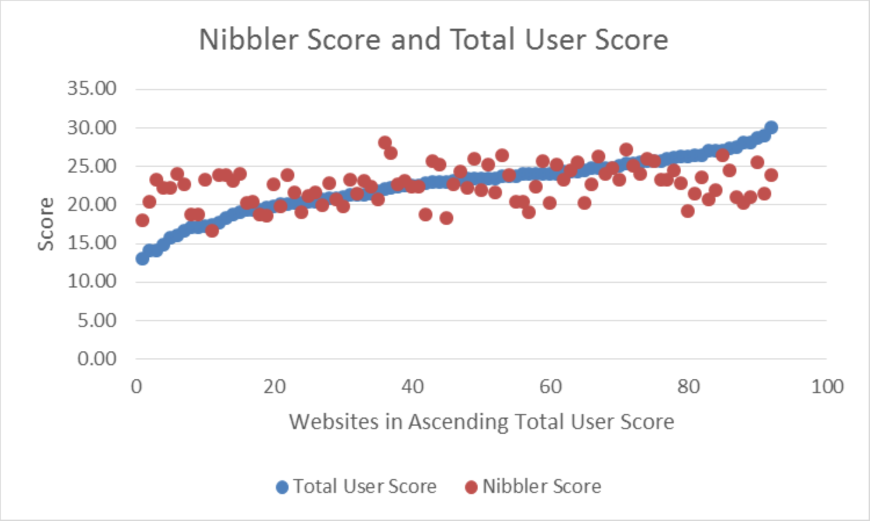
Since there was no way to determine what offset would be the “best” the offset was adjusted until the average error was as low as possible. Manipulating the error to be as low as possible was not an attempt to skew the results. Instead, it was an attempt to give the online evaluators the benefit of the doubt. That way, if the online evaluators failed to produce promising results I could be sure that they were not effective enough to be considered useful. If they did produce promising results then additional research would certainly be pursued.

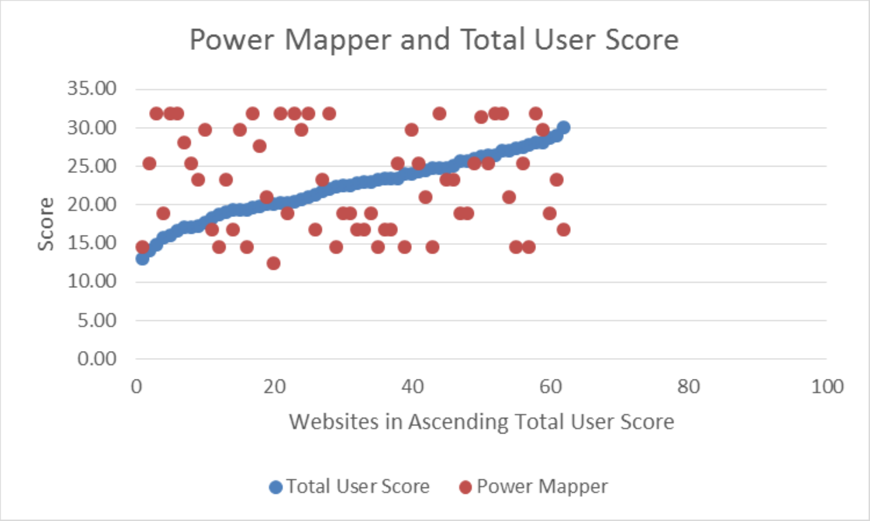
The results are presented graphically (below) and in Table 4.1.

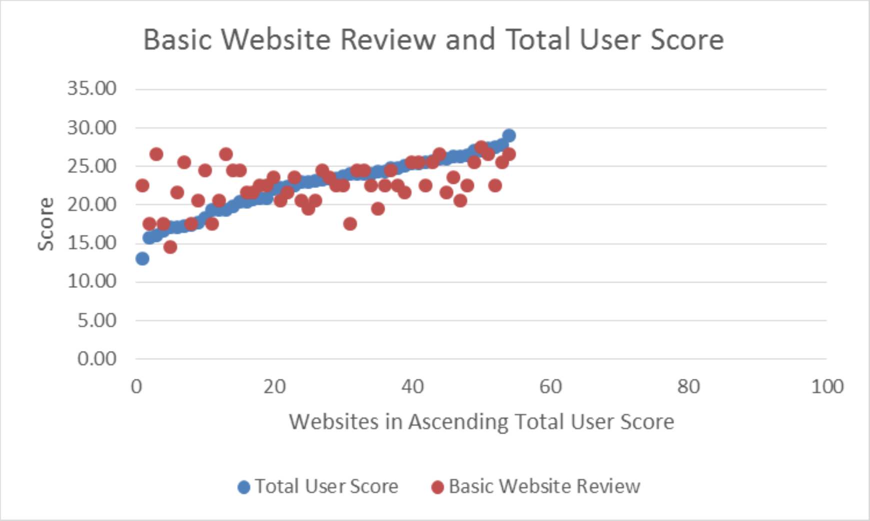

Random and Total User Score
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
Score
0
20
40
60
80
100
Websites in Ascending Total User Score
● Total User Score
● Random


Yottaa Score and Total User Score
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
Score
0
20
40
60
80
100
Websites in Ascending Total User Score
● Total User Score
● Yottaa Score


Pagespeed Score and Total User Score
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
Score
0
20
40
60
80
100
Websites in Ascending Total User Score
● Total User Score
● Pagespeed Score


Nibbler Score and Total User Score
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
Score
0
20
40
60
80
100
Websites in Ascending Total User Score
● Total User Score
● Nibbler Score


Power Mapper and Total User Score
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
Score
0
20
40
60
80
100
Websites in Ascending Total User Score
● Total User Score
● Power Mapper


Basic Website Review and Total User Score
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
Score
0
20
40
60
80
100
Websites in Ascending Total User Score
● Total User Score
● Basic Website Review

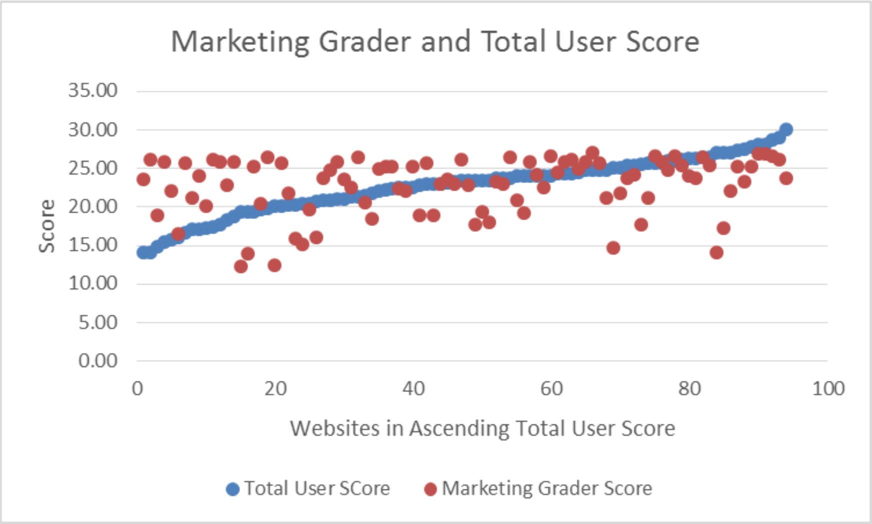


Figure 4.1: Evaluator Results

Note that the number of websites is not consistent across the six graphs. Some online evaluators did not retrieve valid results for all the websites. Four of the online evaluators were able to score about 90 of the websites, but Power Mapper scored only 62 of the 100 websites. Basic Website Review only scored 54 (a 46% failure rate), but seemed to be the best a predicting the Total User Score (with an average error of .083). The second most accurate was Nibbler (with an average error of .088). The rest of the graders all had average errors of .1 or above with Power Mapper being the worst at .19.

In addition to the average error, a Pearson correlation was calculated for each online evaluator [21]. A Pearson Product-Moment Correlation is useful for determining the correlation of two linear items. Because I selected graders that provide Quantitative, Linear Scores I know that a Pearson Correlation will produce useful results. Unfortunately, there is little precedence for determining how strong correlations between usability are. Generally an r value of .10 to .29 is considered a weak correlation, an r value of .30 to .49 is considered a medium correlation and an r value of .50 to 1 is considered a strong correlation. It is important to note however, that these values can vary significantly depending on the application . For example, in biology a Pearson correlation of .90 might be considered too small, but in Tourism and Hospitality, and Pearson correlation of .40 might be considered quite large.

A good way to think about r values is to consider their squared value. An r value of .316 corresponds to an r^2 value of about .10. This means that the item being tested accounts for about 10% of the variability of the original item. For example, from our tests I discovered that Nibbler had a Pearson Correlation value of .309. This means that the r^2 value is .095. Which in turn means that Nibbler was able to account for 9.5% of the variance in the usability scores from the human testers.

The best performing evaluators were Nibbler and Basic Website Review with Pearson Correlations of .308 and .407 respectively. These graders were able to account for 9% and 16% of the variability respectively. For some applications this might be very useful. Being able to predict about 16% of the perceived usability of a website might go a long way in increasing business. Especially considering that these online evaluators are freely available and only take about a minute to run. When they were combined they performed even better, which makes it hopeful that a tool could be created that would incorporate many different evaluators to make a stronger tool.

Both Yotta and Power Mapper had large standard deviations. Both of these evaluators performed very poorly when compared to the human evaluations. The variability of their scores makes them unpredictable and probably not useful for usability evaluations. Pagespeed, on the other hand, had a low standard deviation. Pagespeed was primarily concerned with speed and the optimization of speed, so it is not surprising that Alexa's Top 500 Websites performed similarly.

	Average Error	Pearson	Equation
Random Score	0.206893	0.014877	*Rand()* * 24 + 8
Yottaa	0.159858	-0.03519	*Score* * .24 + 7.3
Pagespeed	0.101095	0.054372	*Score* * .24 + 2.4
Nibbler	**0.087986**	**0.308722**	*Score* * 2.4 + 7.7
Power Mapper	0.190154	0.081516	*Score* * 24 + 8
Basic Website Review	**0.083056**	**0.407523**	*Score* * 24 + 8
Marketing Grader	0.111321	0.184852	*Score* * 24 + 8
Nib + BWR	**0.079776**	**0.425272**	Average
Nib + BWR + PS	0.103284	**0.487581**	Average

Table 4.1: Errors of Graders

	Mean	Stdev
Yottaa	22.44722	**5.005048**
Pagespeed	22.86063	**1.897436**
Nibbler	22.52	2.390754
Power Mapper	22.99097	**6.390127**
Basic Website Review	22.574	2.814335
Marketing Grader	22.72213	3.683332

Table 4.2: Statistics of Automated Graders

One of the things I was originally interested in when I started this thesis was whether or not online evaluators could be combined to give better coverage and provide users with fuller feedback than they could individually. To try and test this I tried averaging the scores of some of the better graders to see if they would correlate better with the user scores than they did individually. The most promising results came from combining Basic Website Review and Nibbler (the two evaluators with the lowest average errors). Combined they were able to reduce the average error to .080 (reducing it .003 from Basic Website Review and .008 from Nibbler) they also achieved a higher Pearson's value (.425). Additionally, Basic Website Review, Nibbler and Pagespeed were combined. They had higher average error than any of those three evaluators had individual but did have a higher Pearson's value (.489).

Ultimately, there is no easy answer for stating whether or not the online evaluators performed well enough to be considered viable options for replacing traditional usability evaluations. The best this paper can do is to point out the strengths and limitations of these graders.

These tools have a long way to go before they can be considered strong replacements for usability evaluations, but there does seem to be evidence that they are at least making progress towards estimating usability. In a business situation these predictions might even be useful alternatives to expensive and time consuming usability evaluations.

4.2 Usability Traits

To determine which traits were most important in the graders' minds they were put through the same calculations as the online evaluators. Presumably, traits that are better predictors of the Total User Score are more important in the graders' minds. This is because traits that are better predictors correlate better with the rest of the scored traits.

Each trait represents $\frac{1}{8}$th of the total user score. This means that if all the traits were scored randomly (with no relation to one another) you would expect to see most of the errors around 0.125 ($\frac{1}{8}$) and a Pearson Correlation of about .35. A trait receiving such a score would have effectively a random distribution compared to the seven other traits. A trait with an error below 0.125 would represent a trait that has at least some correlation with the other traits (I.E. generally high user scores would correspond to generally high scores in that trait).

According to the average errors, Navigation was the best predictor of the

	Average Error	Pearson	Equation
Background	0.069383	0.763229	*Score* * 8
Color Choices	0.068446	0.742232	*Score* * 8
Content	0.074205	0.774066	*Score* * 8
Fonts	0.061182	0.797638	*Score* * 8
Interest	0.077631	0.853067	*Score* * 8
Layout	0.078248	**0.900682**	*Score* * 8
Load Time	**0.125084**	**0.469802**	*Score* * 8
Navigation	**0.058002**	0.845938	*Score* * 8

Table 4.3: Errors of Usability Traits

Average User Score with an average error of only .058 (which is much better than the best online evaluator score of .083). This means that Navigation was generally good at predicting what users would think of the other usability traits as well as usability in general. However, Layout received the highest Pearson Correlation (.901) and had the second lowest average error (.078). When measuring Layout, users were asked to specifically consider the usability and attractiveness of the website so it seems logical that it would have the highest correlation with the Total User Score.

Equally interesting is how poorly Load Time performed. Load Time had the highest average error (.125) and the lowest Pearson Correlation (.470). Being able to tell the difference of load times between different websites was probably difficult for users since most pages load in seconds. More importantly, if pages are cached on the evaluators' computers then they might take less time to load that would otherwise be expected. It might be worth investigating the importance of Load Time by creating an experiment where the websites load time could be controlled and made artificially slower.

The standard deviation of the Interest was highest of all the rubric traits. The Interest of a website was the most subjective rubric trait that I asked volunteers

	Mean	Stdev
Background	2.6552	0.5791
Color Choices	2.7139	0.5687
Content	2.9094	0.5601
Fonts	2.7927	0.5354
Interest	2.7067	**0.7270**
Layout	2.6221	0.6988
Load Time	**3.2465**	*0.4301*
Navigation	2.8101	0.5835

Table 4.4: Statistics of Usability Traits

to grade, so it follows that volunteers would grade the websites very differently. Also noteworthy is that the Load Time had the highest mean and lowest standard deviation. I expect that the load times were not very noticeable when volunteers were grading websites, due to advances in technology. This led to most websites receiving similar, high scores of Load Time.

Chapter 5

Conclusion

Performing usability evaluations is important, but using human evaluators is often too expensive and time consuming for projects. Instead, automation has been proposed to reduce costs and improve the consistency of testing.

In this thesis, I investigated the accuracy of freely available online usability evaluators. The results show that the automated evaluators are not ready to be used as a replacement for real human evaluations, although they might be beneficial if used in tandem with human evaluations. There does seem to be hope for automated evaluators in the future, but the correlations are still too low to be taken seriously.

One of the contributions of this thesis is a quantitative testing method that can be performed by relative unskilled human evaluators. The testing method was validated with a high Cronbach's α and can be used to make quantitative comparisons between websites or to perform experiments similar to the one described in this thesis. The rubric can be used to test additional websites and increase the number of websites to test the automated evaluators with.

Developers of future online evaluators should consider focusing on the layout and navigation of webpages. The survey results showed that these two aspects were most important to users when considering usability. Online usability evaluators should encourage website developers to spend a significant portion of their design process considering the usability of their layout.

Although this thesis focused on website usability, many of these conclusions could also apply to software in general. Online usability evaluators only evaluate websites, but the experiment described in this thesis could be applied to any kind of usability evaluator. The experiment could even be performed to determine the effectiveness of software usability testing procedures. Additionally, anyone creating any kind of software with a user interface should consider the impact of their layout and navigation.

Chapter 6

Future Works

This thesis was only a brief investigation into some of the many things that might be done to improve website usability. There are many ways that automated website grading could be improved. This section discusses some additions that could be made to this thesis and some future areas that could also be investigated.

6.1 Use VIPS Algorithm

The results of the rubric showed that users considered Layout to be one of the most important attributes when considering usability. Unfortunately, trying to determine the layout of a website programmatically can be difficult. In my research I found an algorithm that could help programmers make calculations on the visual layout of a website. The Vision-based Page Segmentation Algorithm or **VIPS** is described in detail in [8] but a brief description is provided here.

The VIPS algorithm combines the DOM (Document Object Model) tree of a web page with a visual rendering of the web page. Since pages often utilize the

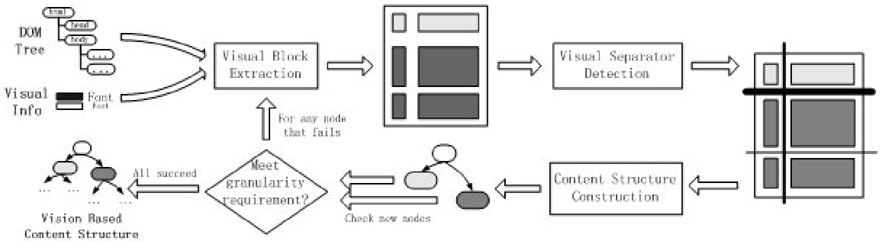

html
head
body
DOM Tree
Visual Info
Font
Visual Block Extraction
Visual Separator Detection
For any node that fails
Vision Based Content Structure
All succeed
Meet granularity requirement?
Check new nodes
Content Structure Construction

Figure 6.1: VIPS Algorithm Flowchart

flexibility of HTML syntax and do not obey W3C html specifications it is easy to create mistakes in the DOM tree structure that cause it to not reflect the real or intended layout of a web page.

Visual blocks are extracted from the DOM tree by comparing the nodes in the DOM tree to visual information from a rendering of the web page. Each iteration of the algorithm splits the DOM tree by identifying visual separators on a page. A series of 13 rules (outlined in the paper [8]) are applied to new blocks to determine if they should be removed from the tree or split into more blocks. The iterations complete once all the blocks have met or exceeded a minimum degree of coherency. The degree of coherency is a measure that is used to keep blocks with similar content together.

The algorithm outputs a tree that better represents the visual layout of the page. Theoretically, calculations performed on the tree and the resulting imdata could be mined to determine if there were any correlations between properties of the tree and the usability of website. One thought is that a tree with a large number of leaf nodes would probably look very complex when displayed in a web browser. It likely that there is a correlation between websites with a large number of leaf nodes and websites that users find too complex.

6.2 Record More Information

Most of the online evaluators that were investigated reported additional information beyond just an overall score for the site. These items could be extracted from pages and tested to see if more correlations existed. Realistically, only a small amount of the data has been reviewed and much can still be done to investigate automated evaluators.

One of the weakness of this study is that the websites tested in this study were considered some of the best and most popular web pages online. These high quality websites may have skewed the results; it would be beneficial to find websites that are less well known and not as high quality as the sites used in this study. A larger range in the popularity or quality of websites would make for a more diverse population to test from.

6.3 Pay Money for Evaluators

To conserve money only freely available online evaluators were selected for this experiment. These online evaluators were useful for providing an experimental framework for how testing and evaluation can be done but might be more limited in ability than commercial software. A number of evaluators were available for purchase that were overlooked. Now that a framework has been developed it would be fruitful to investigate the effectiveness of these evaluators.

Alternatively, the human testers could be replaced with usability consultants. Friends and family were asked to help by taking surveys to evaluate sites. This proved fruitful for the purposes of this experiment, but having professional usability evaluators do evaluations might be more informative.

6.4 Create a Real Tool

Rather than taking the time to develop an actual tool, I instead attempted to prove or disprove the viability of such a tool. I believe that the results of this paper prove to some degree that a tool for measuring usability is viable and could be beneficial to web designers. Perhaps the most obvious next step of this paper is to use these results to create a new usability evaluation tool.

It would be beneficial to see how effective the tool would actually be for website designers. Rather than just studying their accuracy, it would be informative to interview the website designers to see if a product like the one described in this paper would be useful for them. A tool could then be created and tested by the web designers. They could report the usefulness of the tool and whether or not they would consider using it regularly for the maintenance of their website.

Bibliography

[1] SelenuimHQ. http://seleniumhq.org/.

[2] 4Teachers.org. Web Site Design Rubric. http://rubistar.4teachers.org/.

[3] Alexa Internet, Inc. Alexa. http://www.alexa.com/.

[4] A. Anandhan, S. Dhandapani, H. Reza, and K. Namasivayam. Web usability testing CARE methodology. *Generations Journal Of The American Society On Aging*, 2006.

[5] R. Atterer. Model-based automatic usability validation: a tool concept for improving web-based UIs. In *Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges*, pages 13–22, 2008.

[6] S. Baker, F. Au, G. Dobbie, and I. Warren. Automated Usability Testing Using HUI Analyzer. In *19th Australian Conference on Software Engineering*, pages 579–588, 2008.

[7] T. Brinck and E. Hofer. Automatically evaluating the usability of web sites. *CHI '02 extended abstracts on Human factors in computer systems - CHI '02*, page 906, 2002.

[8] D. Cai, S. Yu, J. Wen, and W. Ma. VIPS: a visionbased page segmentation algorithm. Technical report, Citeseer, 2003.

[9] G. Chattratichart, Jarinee; Lindgaard. A comparative evaluation of heuristic-based usability inspection methods. In *Proceeding of the twenty-sixth annual CHI conference extended abstracts on Human factors in computing systems - CHI '08*, pages 2213–2220, 2008.

[10] E. de Kock, J. van Biljon, and M. Pretorius. Usability evaluation methods: Mind the gaps. In *Proceedings of the 2009 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists*, pages 122–131, 2009.

[11] A. Dingli and J. Mifsud. USEFul : A Framework to Mainstream Web Site Usability Through Automated Evaluation. *International Journal*, (2):10–30, 2011.

[12] M. Fok. About BWR. http://www.basicwebsitereview.com/about.html.

[13] Google Development. Frequently Asked Questions - PageSpeed Insights Google Developers. https://developers.google.com/speed/docs/insights/faq.

[14] J. Hair, R. Anderson, R. Tatham, and W. Black. *Multivariate Data Analysis with Readings*. Prentice Hall, 1988.

[15] J. Harty. Finding Usability Bugs with Automated Tests. pages 20:20–20:27, Jan 2011.

[16] K. Hornbaek and E. L.-C. Law. Meta-Analysis of Correlations Among Usability Measures. In *CHI 2007 Proceedings Emprical Models*, pages 617–626. ACM, 2007.

[17] HubSpot. Marketing Grader. http://marketing.grader.com/?s=wsg.

[18] Y. Inc. About Us: Yottaa Web Performance Optimization Solutions. https://www.yottaa.com/about, 2012.

[19] S. Krug. *Rocket Surgery Made Easy*. New Riders, 2010.

[20] J. Kuzic and Z. Rosko. Navigating customer satisfaction - web-site issues. *28th International Conference on Information Technology Interfaces, 2006.*, pages 407–412, 2006.

[21] Laerd Statistics. Pearson Product-Moment Correlation. https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php.

[22] P. Longstreet. Evaluating Website Quality : Applying Cue Utilization Theory to WebQual. *Sciences-New York*, pages 1–7, 2010.

[23] D. Modjeska and A. Marsh. Structure and Memorability of Web Sites. 1997.

[24] C. J. Mueller, D. Tamir, O. V. Komogortsev, and L. Feldman. An Economical Approach to Usability Testing. In *33rd Annual IEEE International Computer Software and Applications Conference*, pages 124–129, 2009.

[25] J. Nielsen. Usability 101: Definition and Fundamentals - What, Why, How (Jakob Nielsen's Alertbox). http://www.useit.com/alertbox/20030825.html.

[26] J. Nielsen. Cost of User Testing a Site (Jakob Nielsen's Alertbox). http://www.useit.com/alertbox/980503.html, 1998.

[27] Powermapper Software. PowerMapper.com - Website Testing and Site Mapping Tools. http://www.powermapper.com/.

[28] S. Rosenbaum. Usability evaluations versus usability testing: When and why? *Professional Communication, IEEE*, 32(4), 1989.

[29] Silktide. About Nibbler - the free website testing tool. http://nibbler.silktide.com/about.

[30] A. Usability. Usability Goals: Memorability of a Website. http://www.affordableusability.com/usability/memorability.html.

[31] J. M. Zain, M. Tey, and G. Y. Soon. Using Aesthetic Measurement Application (AMA) to Measure Aesthetics of Web Page Interfaces. *2008 Fourth International Conference on Natural Computation*, pages 96–100, 2008.

[32] S. Zong, Y. Wang, and S. Zong. White space design and its application for website interface. *2008 9th International Conference on Computer-Aided Industrial Design and Conceptual Design*, pages 928–932, Nov. 2008.

Appendix A

Tables

CATEGORY	Great	Good	Fair	Poor
Fonts	The fonts are consistent, easy to read and point size varies appropriately for headings and text. **Use of font styles (italic, bold, underline) is used consistently and improves readability.**	The fonts are consistent, **easy to read** and point size varies appropriately for headings and text.	The fonts are **consistent** and point size varies **appropriately** for headings and text.	A wide variety of fonts, styles and point sizes was used.
Content	The site has a **well-stated** clear purpose and theme that is carried out throughout the site.	The site has a clearly stated purpose and theme, but may have **one or two elements that do not seem to be related to it.**	The purpose and theme of the site is somewhat **muddy or vague.**	The site lacks a purpose and theme.
Interest	The author has made **an exceptional attempt** to make the content of this Web site interesting to the people for whom it is intended.	The author **has tried** to make the content of this Web site interesting to the people for whom it is intended.	The author has put lots of information in the Web site but there is **little evidence** that the person tried to present the information in an interesting way.	The author has provided only **the minimum** amount of information and has not transformed the information to make it more interesting to the audience (e.g., has only provided a list of links to the content of others).

Layout	The Web site has an **exceptionally** attractive and usable layout. It is easy to locate all important elements. White space, graphic elements and/or alignment are used effectively to organize material.	The Web pages have an **attractive** and usable layout. It is easy to locate all important elements.	The Web pages have a usable layout, but may appear **busy or boring**. It is easy to locate most of the important elements.	The Web pages are **cluttered** looking or **confusing**. It is often **difficult** to locate important elements.
Navigation	Links for navigation are clearly labeled, **consistently placed**, allow the reader to easily move from a page to related pages (forward and back), and take the reader where s/he expects to go. A user does not become lost.	Links for navigation are **clearly labeled**, allow the reader to **easily move** from a page to related pages (forward and back), and internal links take the reader where s/he expects to go. A user rarely becomes lost.	Links for navigation take the reader where s/he expects to go, but **some needed links seem to be missing**. A user sometimes gets lost.	Some links do not take the reader to the sites described. **A user typically feels lost**.

Background	Background is **exceptionally** attractive, consistent across pages, adds to the theme or purpose of the site, and does not detract from readability.	Background is **attractive**, consistent across pages, **adds to the theme or purpose of the site**, and does not detract from readability.	Background is consistent across pages and does not detract from readability.	Background **detracts** from the readability of the site.
Color Choices	Colors of background, fonts, unvisited and visited links **form a pleasing palette**, do not detract from the content, and are consistent across pages.	Colors of background, fonts, unvisited and visited links do not detract from the content, and **are consistent across pages**.	Colors of background, fonts, unvisited and visited links do not detract from the content.	Colors of background, fonts, unvisited and visited links make the content hard to read or otherwise **distract** the reader.
Load Time	The page loads very **quickly** (10 seconds or less) on a 54k modem due to small graphics, good compression of sounds and graphics, and appropriate division of content.	The page loads **reasonably quickly** (10-15 seconds) on a 54k modem due to small graphics, good compression of sounds and graphics, and appropriate division of content.	The web page takes a little over 15 seconds to load. It's a **little bit frustrating**, but you don't have to wait long.	The web page takes more than 20-30 seconds to load. It is **very frustrating** how long it takes.

Table A.1: User Evaluation Rubric

URL	Background	Color Choices	Content	Fonts	Interest	Layout	Load Time	Navigation	Total User Score	Range of Dates Viewed
http://searchnu.com/	2	2	3	2	1.25	2.5	3.25	3.25	19.25	6/11/2012 - 7/5/2012
http://w3schools.com/	2.75	2.75	3.25	2.75	3	2.25	3.25	3	23	6/11/2012 - 7/5/2012
http://www.barnesandnoble.com/	3	3	3.33	3	3.67	3	4	3.33	26.33	6/11/2012 - 7/5/2012
http://www.bestbuy.com/	3.33	3.25	3.25	3.25	3.5	3	3	3	25.58	6/11/2012 - 7/5/2012
http://www.hp.com/	2.5	2.5	3.67	3.25	3.25	3	3.25	3	24.42	6/11/2012 - 7/5/2012
http://www.huffingtonpost.com/	2.67	2.5	3	2.75	3.5	2.5	2.75	2.75	22.42	6/11/2012 - 7/5/2012
http://www.seomoz.org/	3.5	3.33	3.75	3.25	4	3.75	3.75	3.25	28.58	6/11/2012 - 7/5/2012
http://www.weather.com/	3	3	3.33	2.67	3.33	2.67	3	3	24	6/11/2012 - 7/5/2012
http://www.zillow.com/	3	3	3.25	2.75	3.25	2.75	3	3	24	6/11/2012 - 7/5/2012
http://yfrog.com/	2.75	2.75	2.67	2.5	2.67	2.25	3	1.75	20.33	6/11/2012 - 7/5/2012
http://9gag.com/	2.67	2.67	3.67	2.33	3	2.67	3.67	3	23.67	6/14/2012 - 6/16/2012
http://abcnews.go.com/	2.67	2.33	3.33	2.67	2.5	2.67	3.33	2.67	22.17	6/14/2012 - 6/16/2012
http://www.bing.com/	3	3	3	3.33	2	3	3.33	3.33	24	6/14/2012 - 6/16/2012
http://www.cnn.com/	2.33	2.33	3.33	2	2.5	2	3	2.33	19.83	6/14/2012 - 6/16/2012
http://www.etsy.com/	2.67	3	3	3	2.67	2.67	3.33	3	23.33	6/14/2012 - 6/16/2012
http://www.nytimes.com/	1.67	1.67	2.67	2.33	2	1.67	3	1.67	16.67	6/14/2012 - 6/16/2012
http://www.slideshare.net/	2	2.67	3	2.67	2.33	2	3	2.33	20	6/14/2012 - 6/16/2012
http://www.ted.com/	2.67	3	3.33	3	3.33	2.67	3.33	3	24.33	6/14/2012 - 6/16/2012
http://www.wikimedia.org/	2.67	2.67	2.67	2.67	1	2.33	3.33	3	20.33	6/14/2012 - 6/16/2012
http://xfinity.comcast.net/	2.67	2.67	3.33	2.67	3	1.67	3	2	21	6/14/2012 - 6/16/2012
http://fc2.com/	1	2	1.5	2.25	1.25	1.5	3.25	2	14.75	6/14/2012 - 6/20/2012
http://fileserve.com/	2	2.5	2.75	2.75	2.25	2.5	2.75	2.75	20.25	6/14/2012 - 6/20/2012
http://hootsuite.com/	3	3	2.67	3	2.75	2.67	2.67	2.67	22.42	6/14/2012 - 6/20/2012
http://us.blizzard.com/en-us/	3.25	3.25	3.75	3	3.75	3.5	2.5	3.25	26.25	6/14/2012 - 6/20/2012
http://us.fotolia.com/	2.5	2.5	2.75	3	2.25	3.25	3.25	2.75	22.25	6/14/2012 - 6/20/2012
http://www.domaintools.com/	1.33	1.75	2.25	1.75	1.5	1.75	3.25	1.75	15.33	6/14/2012 - 6/20/2012
http://www.engadget.com/	2.67	2.25	2.75	2.5	3	2.5	3	2.25	20.92	6/14/2012 - 6/20/2012
http://www.expedia.com/	2	2.5	3.25	2.75	3	3	2.75	3.75	23	6/14/2012 - 6/20/2012
http://www.fedex.com/	2.33	2.75	3.25	2.25	2.75	3	3.25	3.25	22.83	6/14/2012 - 6/20/2012
http://www.linkedin.com/	2.67	3	2.75	2.75	3	3	3.25	3	23.42	6/14/2012 - 6/20/2012
http://ezinearticles.com/	2	2.33	3	2.67	2	1.67	3.33	2.33	19.33	6/14/2012 - 6/26/2012
http://hubpages.com/	3.33	3	3	3.67	3.33	3.33	3	3	25.67	6/14/2012 - 6/26/2012
http://mailchimp.com/	3	2.67	3.33	3.67	3.33	3.67	3.67	3.67	27	6/14/2012 - 6/26/2012
http://thepiratebay.se/	1.5	2.5	1	2	1	1.33	2	1.67	13	6/14/2012 - 6/26/2012
http://wigetmedia.com/	2.33	2.33	2.67	2.33	2.33	2	3.33	2.67	20	6/14/2012 - 6/26/2012
http://www.flickr.com/	3.33	2.67	3	3	3	3	3.33	3.33	24.67	6/14/2012 - 6/26/2012
http://www.informer.com/	2.33	2.33	1.67	2.33	1.67	1.67	2.67	1.33	16	6/14/2012 - 6/26/2012
http://www.mediafire.com/	3.33	3.33	3	3.33	3	3	4	3.33	26.33	6/14/2012 - 6/26/2012
http://www.slonewman.org/	2.67	2.67	3.33	3	3	3	3	2.67	23.33	6/14/2012 - 6/26/2012
http://www.softpedia.com/	3	2.5	2.67	2	2.33	2	4	1.67	20.17	6/14/2012 - 6/26/2012
http://imageshack.us/	2.67	2.75	3.25	2.75	2.5	2.75	3.5	3.25	23.42	6/14/2012 - 7/8/2012
http://imgur.com/	3.5	3.5	3.5	3.5	3.75	3.75	3.75	3.75	29	6/14/2012 - 7/8/2012
http://sandiego.craigslist.org/	2.33	2.25	2.75	2.5	2.75	2.5	3.5	2.75	21.33	6/14/2012 - 7/8/2012
http://soundcloud.com/	3.5	3.5	4	4	3.75	3.75	3.75	3.75	30	6/14/2012 - 7/8/2012

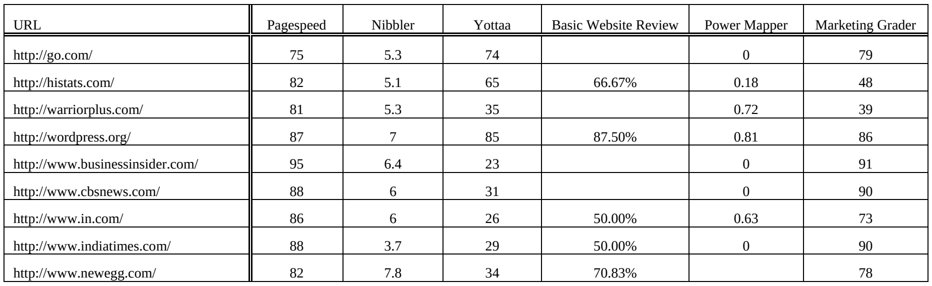
URL	Background	Color Choices	Content	Fonts	Interest	Layout	Load Time	Navigation	Total User Score	Range of Dates Viewed
http://sourceforge.net/	3.33	2.75	3.25	3.5	3.5	3.25	3.5	3	26.08	6/14/2012 - 7/8/2012
http://www.avg.com/us-en/homepage	3.5	3.5	3.25	3.25	3.25	3.5	3.75	3.75	27.75	6/14/2012 - 7/8/2012
http://www.babylon.com/	3.25	3.25	3.25	3.5	3.5	3.5	3.5	3.25	27	6/14/2012 - 7/8/2012
http://www.hostgator.com/	3	2.75	3.25	3	3	2.5	3.25	3.25	24	6/14/2012 - 7/8/2012
http://www.hulu.com/	3.67	3.5	3.5	3.5	3.5	3.25	3.25	3.25	27.42	6/14/2012 - 7/8/2012
http://www.ibm.com/us/en/	3.25	3.5	3.25	3.25	3.25	3	3.5	3	26	6/14/2012 - 7/8/2012
http://archive.org/	2	2.33	2.67	2	1.67	2.33	3.67	2.67	19.33	6/15/2012 - 6/17/2012
http://php.net/	1.67	2	3	2	1.5	1.67	3	2.33	17.17	6/15/2012 - 6/17/2012
http://pinterest.com/	3	3	3.33	3	3	2	3.67	2.67	23.67	6/15/2012 - 6/17/2012
http://www.apple.com/	2.67	3.33	3.33	3	3	3	4	3	25.33	6/15/2012 - 6/17/2012
http://www.ehow.com/	3	2.67	2.67	3.33	3	3.33	4	3.33	25.33	6/15/2012 - 6/17/2012
http://www.ikea.com/	3	2.33	3	2.67	2	2.67	3	3	21.67	6/15/2012 - 6/17/2012
http://www.msn.com/	2.67	2.67	3	2.67	3	2.33	3	2.67	22	6/15/2012 - 6/17/2012
http://www.myspace.com/	1.67	2.67	2	2.67	2	1.67	3.33	2.33	18.33	6/15/2012 - 6/17/2012
http://www.salesforce.com/	3	3.67	3	3.33	2	3	3.33	3.33	24.67	6/15/2012 - 6/17/2012
http://www.squidoo.com/	3	3	2.33	2.67	2.33	2.67	3.67	3.5	23.17	6/15/2012 - 6/17/2012
http://nedroid.com/	2.2	2.4	3	2.6	3	2.5	2.9	2.2	20.8	6/15/2012 - 6/26/2012
http://stackoverflow.com/	2.36	2.18	3	2.5	2.27	2	3.09	2.09	19.5	6/15/2012 - 6/26/2012
http://wordpress.com/	2.55	2.45	2.7	3	2.73	2.45	2.73	2.82	21.43	6/15/2012 - 6/26/2012
http://www.addthis.com/	2.91	3.18	3.09	3.36	2.91	3.18	2.91	2.73	24.27	6/15/2012 - 6/26/2012
http://www.aol.com/	2	1.91	2.18	2.27	2.09	1.82	3	2.36	17.64	6/15/2012 - 6/26/2012
http://www.dell.com/	2.6	2.5	2.9	2.9	2.9	2.6	3.1	3	22.5	6/15/2012 - 6/26/2012
http://www.imdb.com/	3	2.64	3.36	2.73	3.18	2.55	2.82	2.82	23.09	6/15/2012 - 6/26/2012
http://www.reddit.com/	2.27	2.09	2.45	2.36	2.18	2	3	2.36	18.73	6/15/2012 - 6/26/2012
http://www.surveymonkey.com/	2.64	2.91	3.09	3	2.45	2.82	3.27	2.82	23	6/15/2012 - 6/26/2012
http://www.yahoo.com/	2.67	2.67	2	2.75	2.33	1.67	3.17	2.33	19.58	6/15/2012 - 6/26/2012
http://livescore.com/	2.5	2.25	2.75	3	2	2.25	3.75	2.75	21.25	6/16/2012 - 6/23/2012
http://speedtest.net/	3.5	3.5	3	3.5	3.25	3.25	2.5	3	25.5	6/16/2012 - 6/23/2012
http://www.bet365.com/en/	2.75	2.75	3.25	2.75	3.25	3.25	3	3	24	6/16/2012 - 6/23/2012
http://www.putlocker.com/	3	3	3.25	3.25	3.25	2.75	3.5	3	25	6/16/2012 - 6/23/2012
http://www.quickmeme.com/	2.25	2.5	2.75	3.25	2.25	2	3.25	2.5	20.75	6/16/2012 - 6/23/2012
http://www.stumbleupon.com/	3.5	3.25	2.5	3.5	3.5	3.5	3.5	3	26.25	6/16/2012 - 6/23/2012
http://www.swagbucks.com/	3.5	3.5	2.5	3.5	2.75	3.5	3.5	3.25	26	6/16/2012 - 6/23/2012
http://www.tagged.com/	2.75	3.25	2.5	3.5	2.75	3.25	3.25	3	24.25	6/16/2012 - 6/23/2012
http://www.target.com/	2.5	3	3.25	3.25	3.25	3.25	3.25	3	24.75	6/16/2012 - 6/23/2012
http://www.wikipedia.org/	3	3.25	2.75	3.25	2.5	3.25	3.5	3.25	24.75	6/16/2012 - 6/23/2012
http://kat.ph/	2	1	2	3	1	2	4	2	17	6/18/2012 - 6/18/2012
http://photobucket.com/	2	2	3	2	2	2	3	3	19	6/18/2012 - 6/18/2012
http://vimeo.com/	3	4	3	3	3	2	3	3	24	6/18/2012 - 6/18/2012
http://www.4shared.com/	2	2	2	1	1	1	3	2	14	6/18/2012 - 6/18/2012
http://www.deviantart.com/	3	4	3	2	4	4	4	4	28	6/18/2012 - 6/18/2012
http://www.foxnews.com/	2	3	4	3	4	4	4	4	28	6/18/2012 - 6/18/2012
http://www.indeed.com/	1	1	4	3	4	4	4	4	25	6/18/2012 - 6/18/2012

URL	Background	Color Choices	Content	Fonts	Interest	Layout	Load Time	Navigation	Total User Score	Range of Dates Viewed
http://www.nccofc.org/	4	4	3	3	3	3	4	3	27	6/18/2012 - 6/18/2012
http://www.thefreedictionary.com/	2	2	2	2	2	1	4	2	17	6/18/2012 - 6/18/2012
http://fiverr.com/	3	3	3.67	3.33	3.33	3.33	3	3	25.67	6/26/2012 - 7/5/2012
http://go.com/	2.33	2.33	1.33	1.33	1.67	1.67	2	1.33	14	6/26/2012 - 7/5/2012
http://histats.com/	2.67	3	2.33	2.67	2.33	2.33	3	2.33	20.67	6/26/2012 - 7/5/2012
http://warriorplus.com/	2.33	2.33	2.67	2.67	2.33	2	3	2	19.33	6/26/2012 - 7/5/2012
http://wordpress.org/	3.33	3.33	3.67	3	3.33	3.67	3.67	3.33	27.33	6/26/2012 - 7/5/2012
http://www.businessinsider.com/	3	2	2.67	3	2.67	2	3	3	21.33	6/26/2012 - 7/5/2012
http://www.cbsnews.com/	3	3	3	3	3.33	2.33	2.67	3	23.33	6/26/2012 - 7/5/2012
http://www.in.com/	2.33	1.67	1.67	1.67	2	1.67	2.67	2	15.67	6/26/2012 - 7/5/2012
http://www.indiatimes.com/	2	2.33	2.33	2	2.33	1.67	2.67	2	17.33	6/26/2012 - 7/5/2012
http://www.newegg.com/	2.67	2.67	2.67	2.67	3.33	3.33	3	3.33	23.67	6/26/2012 - 7/5/2012

Table A.2: Human Rubric Scores

URL	Pagespeed	Nibbler	Yottaa	Basic Website Review	Power Mapper	Marketing Grader
http://searchnu.com/	86	5.2	93	50.00%	0	32
http://w3schools.com/	69	7.3	88		0.72	60
http://www.barnesandnoble.com/	76	6.6	24		0.09	87
http://www.bestbuy.com/	80	7.6	87		0	69
http://www.hp.com/	84	5.2	61		0.54	85
http://www.huffingtonpost.com/		6.4	31			74
http://www.seomoz.org/	81	7.4	49		0.63	92
http://www.weather.com/	81	5.2	33		0.18	75
http://www.zillow.com/	91	7.3	86	70.83%	0	92
http://yfrog.com/	80	5.8	54	66.67%		63
http://9gag.com/	95	5.3	38			91
http://abcnews.go.com/	77	7.9	30	62.50%	0	86
http://www.bing.com/	94	5.3	99	50.00%	0	68
http://www.cnn.com/	82	6.2	35	79.17%	0.27	91
http://www.etsy.com/	95	7.6	67		0.72	76
http://www.nytimes.com/	85	6.2	97	50.00%	0.09	88
http://www.slideshare.net/	95	6.7	18		0.9	88
http://www.ted.com/	79	7.4	61		0	90
http://www.wikimedia.org/	81	5.6	94	79.17%	0.09	44
http://xfinity.comcast.net/	76	5	82		0.09	79
http://fc2.com/	85	6	77		0.09	60
http://fileserve.com/	84	4.7	82		0.63	47
http://hootsuite.com/	93	6.1	63	75.00%	0.63	73
http://us.blizzard.com/en-us/	91	4.8	73	62.50%	0	80
http://us.fotolia.com/	90	6.2	53	66.67%	0.81	86
http://www.domaintools.com/						89
http://www.engadget.com/						89
http://www.expedia.com/	83	4.4	39	58.33%	0.63	77
http://www.fedex.com/	83	4.6	92		0.72	60
http://www.linkedin.com/	90	7.3	88	70.83%	0	62
http://ezinearticles.com/	90		73	62.50%	0.18	
http://hubpages.com/	95	7.5	59	83.33%	0.63	92
http://mailchimp.com/	95	7.8	71	91.67%	0.54	73
http://thepiratebay.se/	89	4.3	74	70.83%	0.81	
http://wigetmedia.com/	70	5	72		0.54	33
http://www.flickr.com/	97	6.2	72	79.17%	0	89
http://www.informer.com/	74	6.8	87	87.50%	0.09	50
http://www.mediafire.com/	91	5.7	49	70.83%	0.36	91
http://www.slonewman.org/	76	5.9	57	75.00%	0.72	55
http://www.softpedia.com/	81	5.8	31		0.09	72
http://imageshack.us/	80	5.8	60		0.36	56
http://imgur.com/	81	5.7	89	87.50%	0.45	90
http://sandiego.craigslist.org/	95	5.7	90		0	75
http://soundcloud.com/	94	6.7	56		0.72	80
http://sourceforge.net/	90	7	56		0	87

URL	Pagespeed	Nibbler	Yottaa	Basic Website Review	Power Mapper	Marketing Grader
http://www.avg.com/us-en/homepage	75		74	83.33%	0.81	86
http://www.babylon.com/	93	5.9	68		0.09	53
http://www.hostgator.com/	71	6.1	61	79.17%	0.81	89
http://www.hulu.com/	82	5.5	49	70.83%	0.36	78
http://www.ibm.com/us/en/	89		62	87.50%	0.36	84
http://archive.org/	76	4.6	72	87.50%	0.81	86
http://php.net/	58	6.5	75	83.33%	0.45	65
http://pinterest.com/	86	6.7	21		0	77
http://www.apple.com/	76	8.1	80	83.33%		80
http://www.ehow.com/	92	7.2	62	83.33%	0	82
http://www.ikea.com/	77	5.4	85		0.45	58
http://www.msn.com/		8.5	67	75.00%	0.09	85
http://www.myspace.com/	94	6.7	69	79.17%	0.72	76
http://www.salesforce.com/	81	7.7	35		0.81	94
http://www.squidoo.com/	94	6.9	70	79.17%	0.81	77
http://nedroid.com/	90	5.4	71	70.83%	0	84
http://stackoverflow.com/	96		66			
http://wordpress.com/	91	6.1	79		0	67
http://www.addthis.com/	89	7	81	70.83%		89
http://www.aol.com/	88	6.7	62	62.50%	0.18	89
http://www.dell.com/	86	6.1	67		0.63	86
http://www.imdb.com/	91	6.2	50	62.50%	0	79
http://www.reddit.com/	97	6.4	75		0.81	89
http://www.surveymonkey.com/	90	7.5	78	62.50%	0	88
http://www.yahoo.com/	85	4.5	68		0.09	66
http://livescore.com/	86	6.5	83		0.72	
http://speedtest.net/	96	6.8	81	70.83%	0	55
http://www.bet365.com/en/	68	4.7	85			61
http://www.putlocker.com/	86		72	66.67%		42
http://www.quickmeme.com/	91	6.3	44	70.83%	0	80
http://www.stumbleupon.com/	94	6.3	85	75.00%	0.11	81
http://www.swagbucks.com/	83	6.5	77	66.67%	0	92
http://www.tagged.com/	90	6.5	72	58.33%	0.36	83
http://www.target.com/	82	7.1	35		0.45	69
http://www.wikipedia.org/	74	6.8	74	70.83%	0.09	88
http://kat.ph/	87	4.6	64	37.50%	0.25	69
http://photobucket.com/	88	6.8	54		0.45	
http://vimeo.com/	95	7.5	65	79.17%	0	82
http://www.4shared.com/	83	6.5	39		0.36	90
http://www.deviantart.com/	87	5.2	35		0.09	93
http://www.foxnews.com/	80	5.5	36		0.18	93
http://www.indeed.com/	94	6.5	95		0.45	72
http://www.nccofc.org/	75	5.4	77	83.33%	0	40
http://www.thefreedictionary.com/	96	4.6	82	66.67%	0.36	81
http://fiverr.com/	69	6.5	28		0.63	89



URL	Pagespeed	Nibbler	Yottaa	Basic Website Review	Power Mapper	Marketing Grader
http://go.com/	75	5.3	74		0	79
http://histats.com/	82	5.1	65	66.67%	0.18	48
http://warriorplus.com/	81	5.3	35		0.72	39
http://wordpress.org/	87	7	85	87.50%	0.81	86
http://www.businessinsider.com/	95	6.4	23		0	91
http://www.cbsnews.com/	88	6	31		0	90
http://www.in.com/	86	6	26	50.00%	0.63	73
http://www.indiatimes.com/	88	3.7	29	50.00%	0	90
http://www.newegg.com/	82	7.8	34	70.83%		78

Table A.3: Automated Evaluator Scores