Chapter 9

Evaluating the Web Interface Profiles

9.1 Introduction

Chapter 8 demonstrated that the profiles of highly-rated Web interfaces could be used by the author to modify an example Web site. However, it is not clear whether it is possible for others to use these profiles to modify designs. Furthermore, it is not clear whether the resulting designs are of a higher quality than the original ones.

This chapter discusses a study conducted to determine whether changes made based on two profiles – the overall page quality and the good cluster models – improve design quality. The goal of the study was to determine if participants preferred the modified pages and sites above the original ones. Whether or not preferences reflect usability was not examined in this study, and consequently is not claimed. The study also demonstrates that two undergraduate students and a graduate student were able to use the models to revise study sites.

9.2 Study Design

A study was conducted between November 26, 2001 and November 27, 2001, in accordance with guidelines established by the Committee for the Protection of Human Subjects (project 2000-1-36). Thirteen participants completed a within-subjects experiment wherein they performed two types of tasks. The first task – page-level analysis – required participants to explore two alternative designs for a Web page and to select the design that they felt exhibited the highest quality; there were a total of fifteen comparisons for pages from three sites. The second task – site-level analysis – required participants to explore a collection of pages from a Web site and to rate the quality of the site on a 5-point scale. Participants rated alternative designs for two sites; there were a total of four site ratings.

Given that only a subset of pages were modified for each site, it was not feasible to have participants attempt to complete information-seeking tasks during this study. Instead, the page-level and site-level tasks were designed to be consistent with the perceived usability condition in the usability study discussed in Chapter 7.

9.2.1 Study Sites

For the analysis, five sites were randomly selected from various Yahoo categories, such as finance and education; the sites included the one discussed in Chapter 8. Similar to the example site in Chapter 8, the other sites were selected because they had valuable content but also exhibited some

Id	#Pages	Category	Description
Site-Level Analysis			
1	9	Education Health	Information about a community health-education museum and science center as well as its teaching programs. The design consists of small pages with a few graphical links and some colored and italicized body text. This site was discussed in Chapter 8.
2	7	Education	Information on the World Wide Web and computer use for K–12 teachers. The design consists of long pages of text and text links with a logo image at the top; horizontal rules are used extensively, but very little color is used.
Page-Level Analysis			
3	5	Community	Information about a bridge club, including competition results. The design consists of small pages with tables or lists of text links, few images, and several horizontal rules.
4	5	Finance	Links to sites about the Information Economy. The design consists of very long pages with lists of annotated text links, few images, few colors, and several horizontal rules.
5	5	Living	Job listings and information about careers and employment statistics. The design consists of long pages with lots of color and images.

Table 9.1: Descriptions of sites used for the study.

problematic design issues. None of the sites were included in the statistical profile development. Table 9.1 describes the sites used for the study, and Figures 9.1–9.5 depict example pages from each site.

The Site Crawler tool was used to download pages from the five sites; the default crawling options were used (i.e., download fifteen level-one pages and three level-two pages from each level-one page). Only five pages were selected from sites 3, 4, and 5 for the page-level comparisons. All nine of the available pages were used for site 1, and seven pages were selected for site 2.

The same process followed in Chapter 8 was also followed to create modified versions of the 31 pages. Specifically, output from the overall page quality and good page cluster models was used to iteratively make changes to pages so they would be more consistent with the models. Two undergraduate students (Deep Debroy and Toni Wadjiji) and a graduate student (Wai-ling Ho-Ching) modified sites 3, 4, and 5; the author modified sites 1 and 5 and made minor final revisions to sites 4 and 5. The students had little or no training in Web design and had very little experience with building Web sites. Furthermore, they did not have prior experience with the Analysis Tool, the quantitative measures, nor the profiles.

The students made straightforward changes directly based on the decision tree rules and cluster model results. Students had to rely on their own intuition in cases were design changes were not as straightforward. The students reported that they had some difficulty making changes that were suggested by the profiles, such as increasing the number of text columns or decreasing color usage. The students also emphasized that it was not enough to use the overall page quality

Figure 9.1: Example page taken from site 1 (http://www.hallofhealth.org/puppetshows.html; September 14, 2001); this is the content page discussed in Chapter 8. The page was rated poor overall and was 8.33 standard deviation units away from the small-page cluster centroid.

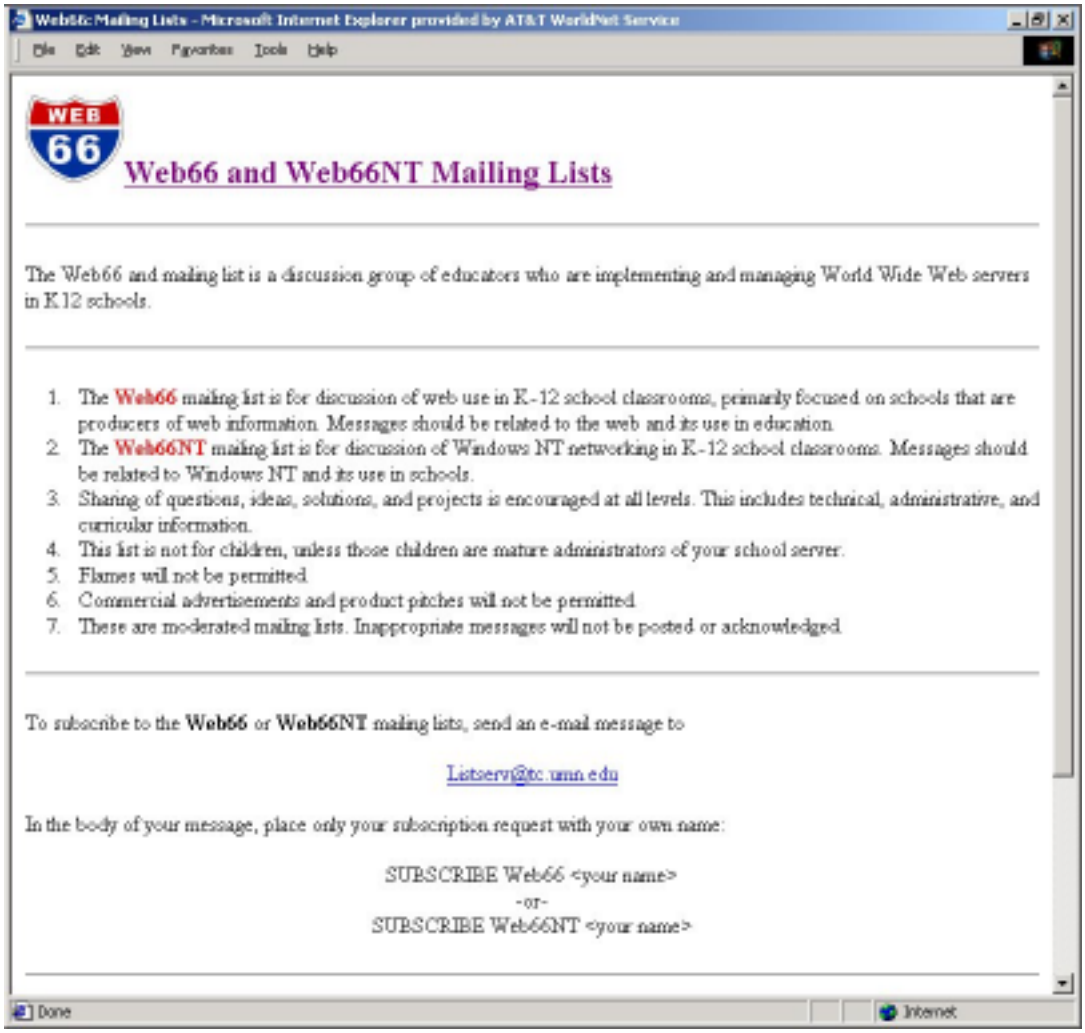

Figure 9.2: Example page taken from site 2 (http://web66.coled.umn.edu/List/Default.html; November 4, 2001). The page was rated poor overall and was 14.97 standard deviation units away from the small-page cluster centroid.

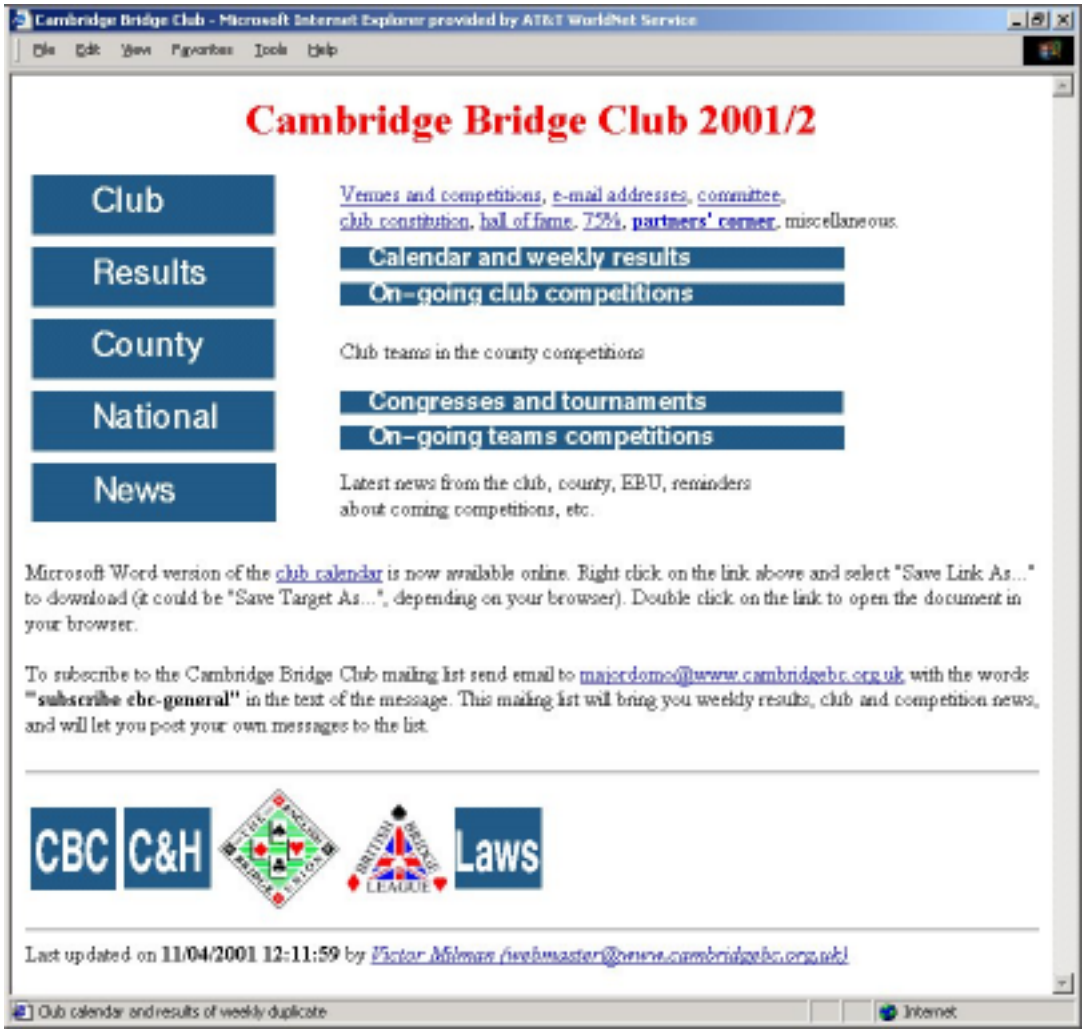

Figure 9.3: Example page taken from site 3 (http://www.cambridgebc.org.uk/CBC.html; November 4, 2001). The page was rated poor overall and was 26.46 standard deviation units away from the small-page cluster centroid.

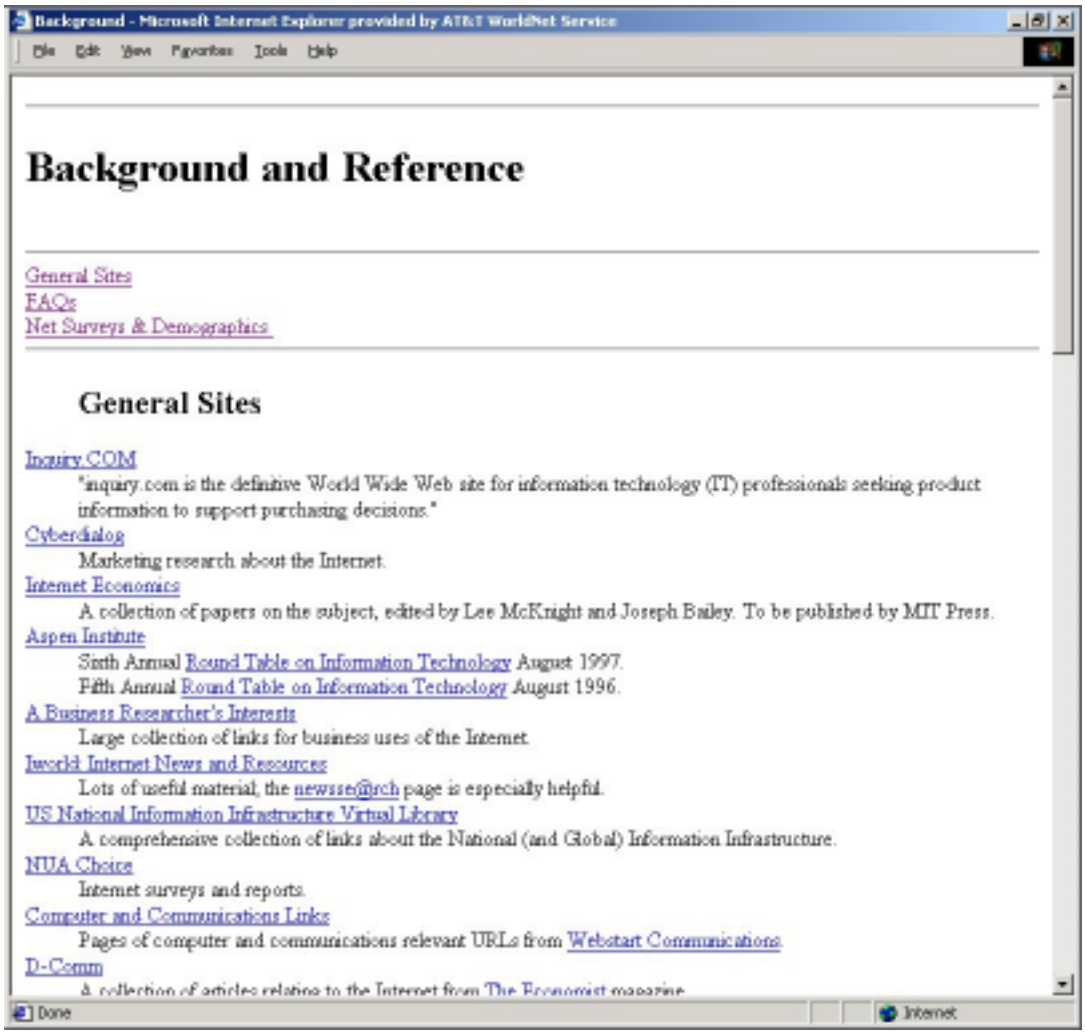

Figure 9.4: Example page taken from site 4 (http://www.sims.berkeley.edu/resources/infoecon/Background.html; November 4, 2001). The page was rated poor overall and was 32.78 standard deviation units away from the large-page cluster centroid.

Figure 9.5: Example page taken from site 5 (http://www.jobweb.com/employ/salary/default.cfm; November 25, 2001). The page was rated poor overall and was 5.99 standard deviation units away from the small-page cluster centroid.

model by itself; the cluster models needed to be used as well. Only the overall page quality and good cluster models were used to inform design changes, and the following rules were observed.

- Changes were made solely based on the Analysis Tool results. Pages were modified to conform as much as possible to the mapped good page cluster model (subgroups of good pages with similar characteristics – small-page, large-page, or formatted-page) first; then, if possible, pages were also modified to be consistent with the overall page quality model (model that does not consider the content category or page type in predictions). The median overall page quality was used as the site quality prediction.

- The content remained the same in the original and modified sites, except in instances where footers or headings were dictated by the model results. If the models dictated that the amount of content on a page needed to be reduced, then the page was split into multiple pages as necessary.

- One change was made at a time and its impact was subsequently assessed. Changes had to result in improved or equal quality (i.e., a reduced distance from the cluster centroid or a change in prediction from poor to good) or they were discarded.

- The modified pages had to have some noticeable differences from the original pages.

Table 9.2 summarizes the differences between the original and modified pages as measured by the overall page quality and cluster models. For the original pages, 6.5% were rated good overall, 6.5% were rated average, 87.1% were rated poor, and the median distance to the mapped cluster centroid was 15.3 standard deviation units. Changes similar to those made for site 1 in Chapter 8 were also made for the four other sites. The changes included reorganizing text or images to reduce scrolling, changing text formatting (e.g., removing italicized body text), reducing text clustering, and changing colors as needed. The models revealed that other changes related to the type of content (e.g., body and link text words) and the similarity in content between source and destination pages were needed, but these changes were not made. For the modified pages, 96.8% were rated good overall, 3.2% were rated average, and the median distance to the mapped cluster centroid was 6.31 standard deviation units. The predictions suggested that there were potentially noticeable differences between pages in most cases. Figures 9.6–9.10 depict modified versions of the example pages in Figures 9.1–9.5. Appendix D provides side-by-side comparisons of the original and modified versions of the five pages.

9.2.2 Participants

Study participants were recruited from Kaiser Permanente's Web Portal Services Group[1]. This group is responsible for designing, building, and maintaining numerous intranet and Internet sites; hence, designing quality Web interfaces is extremely important to people within the group. Thirteen participants completed the study; participants represented the three roles below.

- **Professional Web Designers** - have received formal training (i.e., earned a college or art school degree) in Web or graphic design and have actively designed Web sites. These participants were employed as designers; four of the thirteen participants were from this group.

[1]The author was employed as a member of this group at the time of the study and had working relationships with some of the study participants. Participants were not informed about the purpose of the study or the hypotheses being tested.

Pair Id	Original Page			Modified Page		
	Quality	Cluster	Distance	Quality	Cluster	Distance
Site 1						
–	Poor	Small-Page	22.74	Good	Small-Page	5.16
–	Poor	Small-Page	16.5	Good	Small-Page	4.61
–	Poor	Small-Page	18.74	Good	Small-Page	17.68
–	Poor	Small-Page	6.81	Good	Small-Page	3.88
–	Poor	Small-Page	6.26	Good	Small-Page	3.81
–	Poor	Small-Page	7.95	Good	Small-Page	5.11
–	Poor	Small-Page	10.88	Good	Small-Page	4.56
–	Poor	Small-Page	8.33	Good	Small-Page	5.04
–	Poor	Small-Page	15.94	Good	Small-Page	4.72
Site 2						
–	Poor	Small-Page	21.05	Good	Small-Page	10.86
–	Poor	Small-Page	14.97	Good	Small-Page	7.83
–	Poor	Small-Page	14.68	Good	Small-Page	6.31
–	Poor	Small-Page	20.42	Good	Small-Page	6.83
–	Poor	Small-Page	15.03	Good	Small-Page	6.58
–	Poor	Large-Page	18.14	Good	Large-Page	19.2
–	Poor	Small-Page	16.21	Good	Small-Page	7.59
Site 3						
1	Poor	Small-Page	26.46	Good	Small-Page	5.86
2	Average	Small-Page	14.6	Good	Small-Page	5.71
3	Poor	Small-Page	14.77	Good	Small-Page	6.05
4	Poor	Small-Page	15.3	Good	Small-Page	7.32
5	Poor	Small-Page	15.09	Good	Small-Page	7.52
Site 4						
6	Poor	Large-Page	32.78	Good	Small-Page	11.93
7	Poor	Large-Page	23.05	Good	Small-Page	394.19
8	Poor	Large-Page	49.3	Good	Small-Page	394.18
9	Good	Large-Page	38.61	Good	Large-Page	11.79
10	Poor	Small-Page	8.31	Good	Small-Page	7.19
Site 5						
11	Average	Small-Page	15.43	Average	Small-Page	15.48
12	Poor	Small-Page	11.23	Good	Small-Page	4.82
13	Poor	Large-Page	93.87	Good	Small-Page	4.39
14	Poor	Small-Page	5.56	Good	Small-Page	5.42
15	Good	Small-Page	5.99	Good	Small-Page	5.79

Table 9.2: Model predictions for the original and modified pages. The numbers in the first column are for the page-level analysis; the – indicates pages that are included in the site-level analysis. The quality predictions are from the overall page quality model. The reported clusters and cluster distances are from the good page cluster models; the distance reflects the number of standard deviation units of difference between metric values on a page and metric values at the centroid of a cluster. In most cases, the modified pages are closer to cluster centroids and are predicted to be of a higher quality than the original pages.

Figure 9.6: Modified page for site 1; this is the modified content page discussed in Chapter 8. The page was rated good overall and was 5.04 standard deviation units away from the small-page cluster centroid. The original page was rated poor overall and was 8.33 standard deviation units away from the small-page cluster centroid (see Figure 9.1).

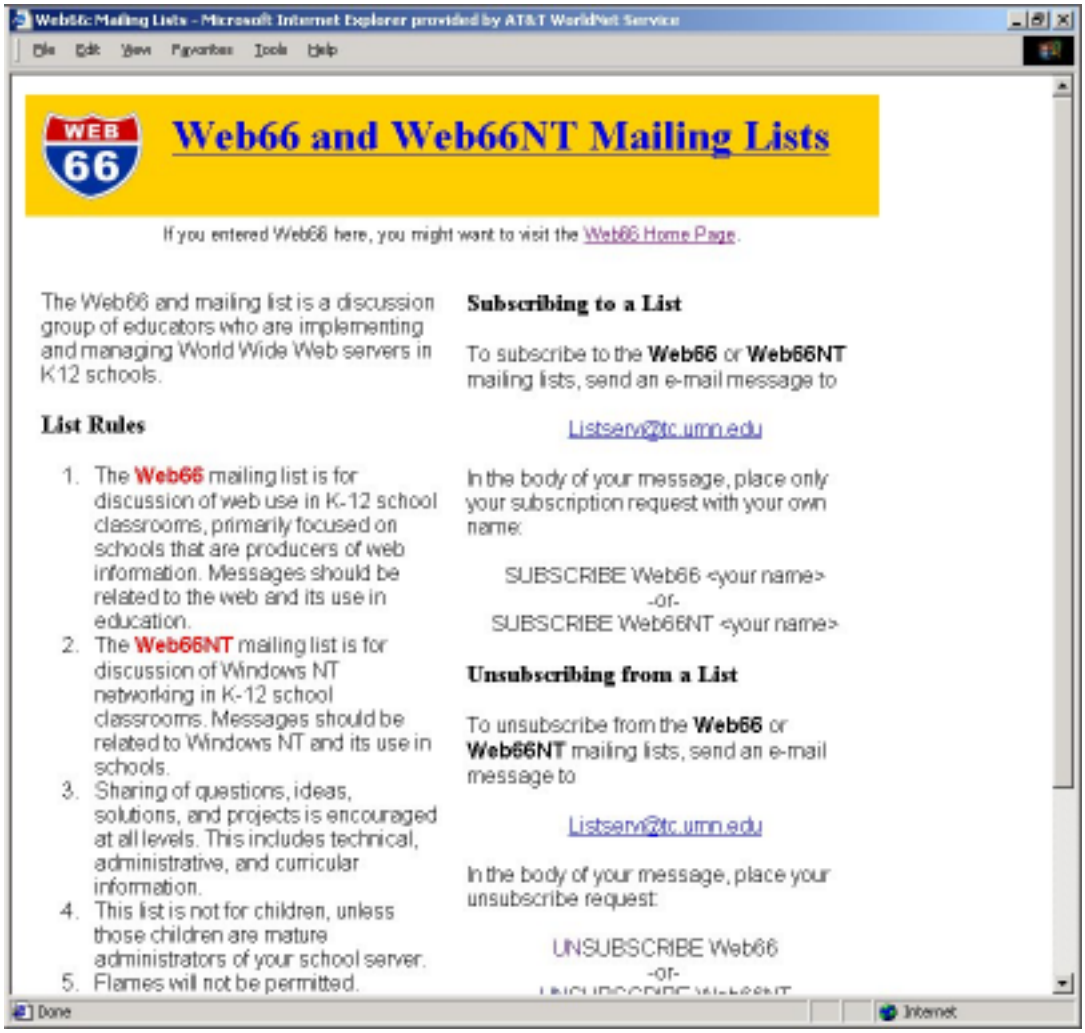

Figure 9.7: Modified page for site 2. The page was rated good overall and was 7.83 standard deviation units away from the small-page cluster centroid. The original page was rated poor overall and was 14.97 standard deviation units away from the small-page cluster centroid (see Figure 9.2).

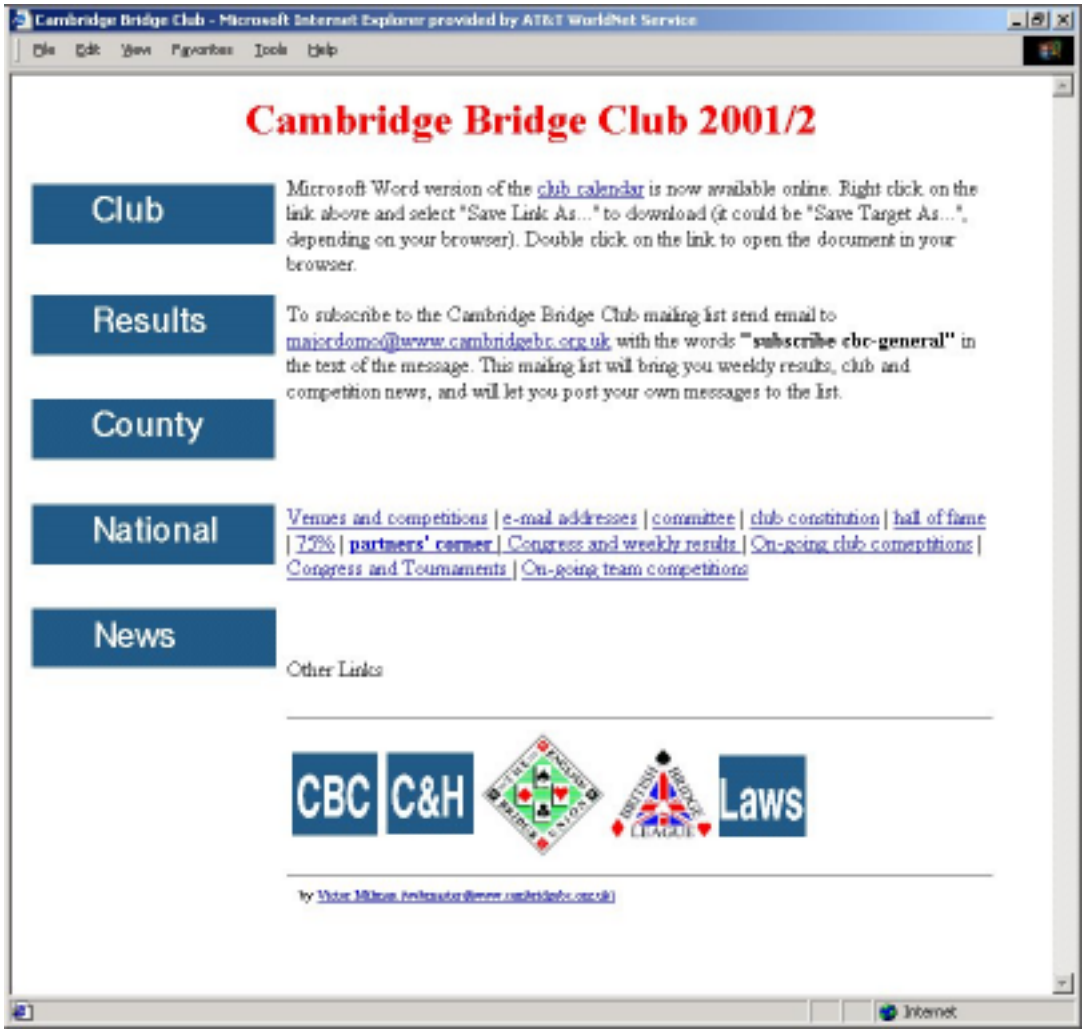

Figure 9.8: Modified page for site 3. The page was rated good overall and was 5.86 standard deviation units away from the small-page cluster centroid. The original page was rated poor overall and was 26.46 standard deviation units away from the small-page cluster centroid (see Figure 9.3).

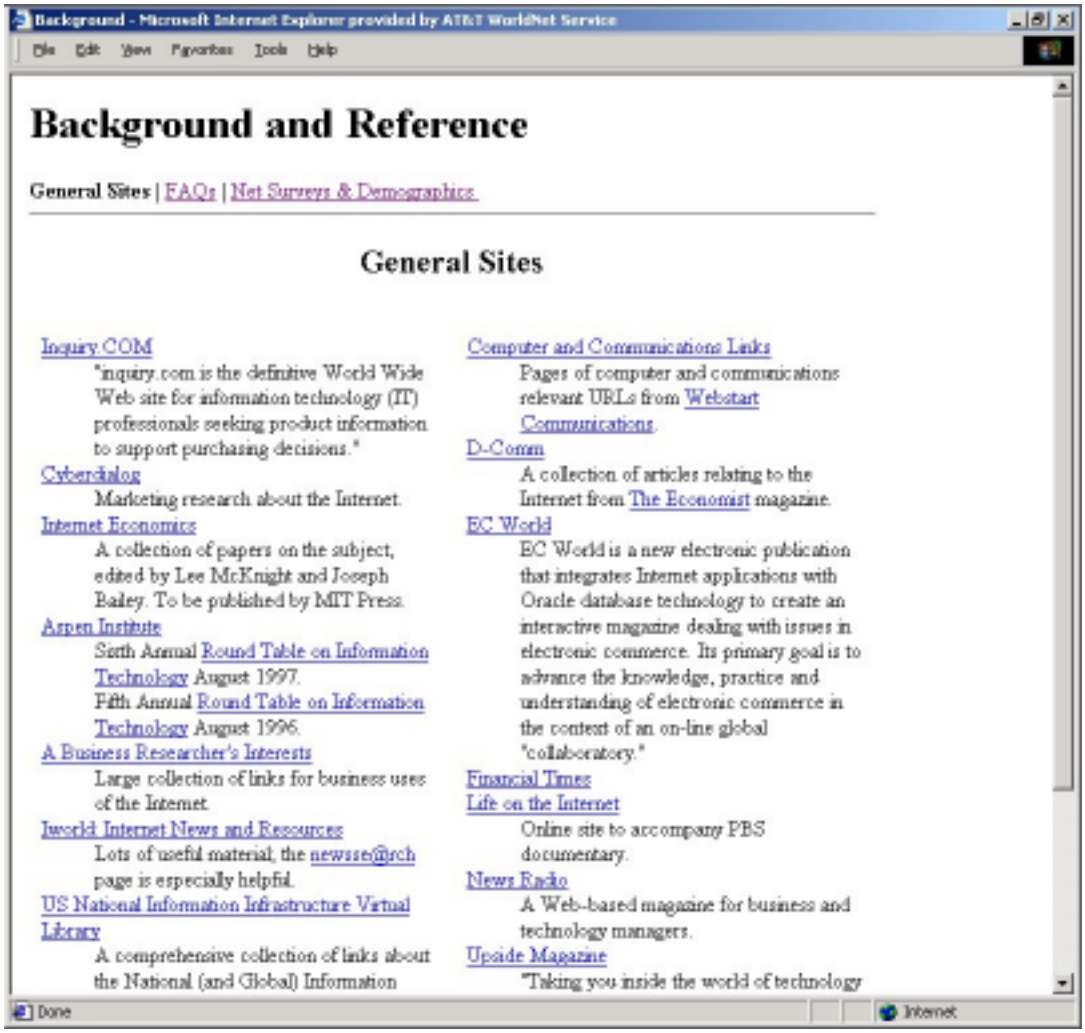

Figure 9.9: Modified page for site 4. The page was rated good overall and was 11.93 standard deviation units away from the small-page cluster centroid. The original page was rated poor overall and was 32.78 standard deviation units away from the large-page cluster centroid (see Figure 9.4).

Figure 9.10: Modified page for site 5. The page was rated good overall and was 5.79 standard deviation units away from the small-page cluster centroid. The original page was rated poor overall and was 5.99 standard deviation units away from the small-page cluster centroid (see Figure 9.5).

Id	Age	Gen	Education	CmpExp	IExp	IUse	EngExp	VI
1	46-55	Male	Some College	Beginner	Average	1–2	Average	No
2	26-35	Male	College Graduate	Expert	Expert	>10	Expert	No
3	26-35	Female	College Graduate	Average	Average	6–10	Expert	No
4	36-45	Female	Post Graduate	Average	Expert	>10	Expert	No
5	>55	Female	College Graduate	Expert	Expert	1–2	Expert	No
6	26-35	Female	College Graduate	Average	Average	>10	Average	No
7	36-45	Male	Some College	Expert	Expert	>10	Average	No
8	46-55	Male	Post Graduate	Average	Expert	>10	Expert	No
9	36-45	Male	Post Graduate	Expert	Expert	>10	Expert	No
10	26-35	Female	College Graduate	Expert	Expert	6–10	Expert	No
11	26-35	Male	College Graduate	Expert	Expert	>10	Expert	No
12	46-55	Female	Post Graduate	Expert	Expert	>10	Expert	No
13	26-35	Female	College Graduate	Average	Expert	3–5	Expert	No

Table 9.3: Summary of participants' demographic information. Participants provided their age and gender (Gen) and described their proficiency with computers (CmpExp), the Internet (IExp), and the English language (EngExp). Participants also reported the number of hours they spend using the Internet weekly (IUse) and whether or not they had a visual impairment that possibly interfered with their ability to assess Web design quality (VI).

- **Non-Professional Web Designers** - have not received formal training (i.e., no degree) in Web or graphic design, but have played a role in designing Web sites or creating Web pages. These participants were employed as Web Coordinators; three of the thirteen participants were from this group.
- **Non Web Designers** - have not received formal training, have not designed Web sites, and have not created Web pages. These participants were typically developers and managers; six of the thirteen participants were from this group.

Participants answered demographic questions prior to starting the study. Specifically, participants provided their age, gender, as well as information about their education background, computer and Internet experience, the number of hours they use the Internet weekly, English experience, and whether they had a visual impairment that could interfere with their ability to assess the quality of Web interfaces. Table 9.3 summarizes responses to these questions. Participants were also asked several questions about their role and experience with designing Web sites, including the number of sites they have designed. Table 9.4 summarizes responses to these questions.

There were seven female and six male participants. The typical participant was 26–35, a college graduate, an expert computer and Internet user, spent more than ten hours online weekly, was an expert with the English language, and had no visual impairments. None of the participants had visual impairments. All of the professional Web designers have designed more than ten sites; half of them felt that they were experts at creating quality sites, while the other half felt that they were average. Two of the non-professional Web designers have also designed more than 10 sites and all of them felt that they were experts at creating quality sites. Most of the non Web designers had designed from zero to three sites and felt that they were beginners at creating quality sites. Based on the demographic information, all of the participants appear to have been appropriate for this study.

Id	Role	DesignExp	#Sites Designed
1	Non Web Designer	Beginner	0–3
2	Professional Web Designer	Expert	>10
3	Non Web Designer	Beginner	0–3
4	Non Web Designer	Average	0–3
5	Non-Professional Web Designer	Expert	0–3
6	Non Web Designer	Beginner	0–3
7	Professional Web Designer	Average	>10
8	Non Web Designer	Beginner	0–3
9	Non-Professional Web Designer	Expert	>10
10	Professional Web Designer	Expert	>10
11	Non-Professional Web Designer	Expert	>10
12	Non Web Designer	Average	4–10
13	Professional Web Designer	Average	>10

Table 9.4: Summary of participants' Web design roles and experience. Participants described their proficiency with creating quality Web sites (DesignExp) and reported the number of sites they had designed prior to the study.

9.2.3 Testing Interface

A testing interface was developed using HTML forms, JavaScript, and PERL. In addition, a script was developed to generate five randomized experiment designs. The order of page comparisons was randomized and controlled such that pages from the same site were not compared consecutively; the presentation of the two page designs was also randomized. Similarly, the order of site ratings was randomized and controlled such that the two versions of a site were not rated consecutively.

9.2.4 Testing Session

The study consisted of a 1-hour session wherein participants completed 15 page-level comparisons and four site-level evaluations. Participants were initially given an instruction sheet (see Figure 9.11) that provided an overview of the study procedure. After reviewing the instruction sheet, participants completed a statement of informed consent and moved to a computer station for completing the study.

The study interface requested demographic information and assigned the participant to one of the 5 experiment designs. The interface subsequently guided participants through two types of tasks as discussed below.

- **Page-Level Analysis.** Participants were instructed to initially explore alternative versions of a Web page. Then, participants were asked to select the version that they felt exhibited the best quality. Participants were also asked to explain why they selected the design. Figure 9.12 depicts the screens for performing the page-level tasks.

- **Site-Level Analysis.** Participants were instructed to initially explore the pages from a Web site. Then, participants were asked to rate the quality of the site using a 5-point Likert scale; ratings ranged from very poor (1) to very good (5). Participants were also asked to explain why they rated the site as they did. Figure 9.13 depicts the screens for performing the site-level tasks.

Web Site Quality Study Instructions

I. Overview of the Study
The purpose of this study is to get user feedback on the quality of a collection of web pages and sites. Basically, you will explore the pages and sites and provide responses to a number of statements about their quality. There is no risk to you, and you will be compensated with a free lunch for your participation.

Please read and sign the informed consent form and return to the tester.

II. Overview of the Study Procedure

General goal:
Rate web pages and sites. There will be two kinds of tasks. (a) Compare two alternative designs for a web page (b) Explore pages on a site and rate the site.

Tasks
(a) **Compare two alternative designs for a web page.**
In the first part of the task, you will be asked to explore the two designs. Look at each design and select the design that you feel exhibits the best quality. **Please do not spend more than a few minutes exploring the designs.**
You are encouraged to comment on why you selected one design over the other one.

(b) **Explore pages on a site and rate the site.**
You will be presented with web pages from a site and asked to explore the pages. Then rate the site. **Please do not spend more than a few minutes exploring the pages.**

Rating the site:
You will be asked to rate the site on a 5-point scale (Very Poor - Very Good). You are encouraged to comment on why you rated the site as you did.

Figure 9.11: Instructions given to study participants.

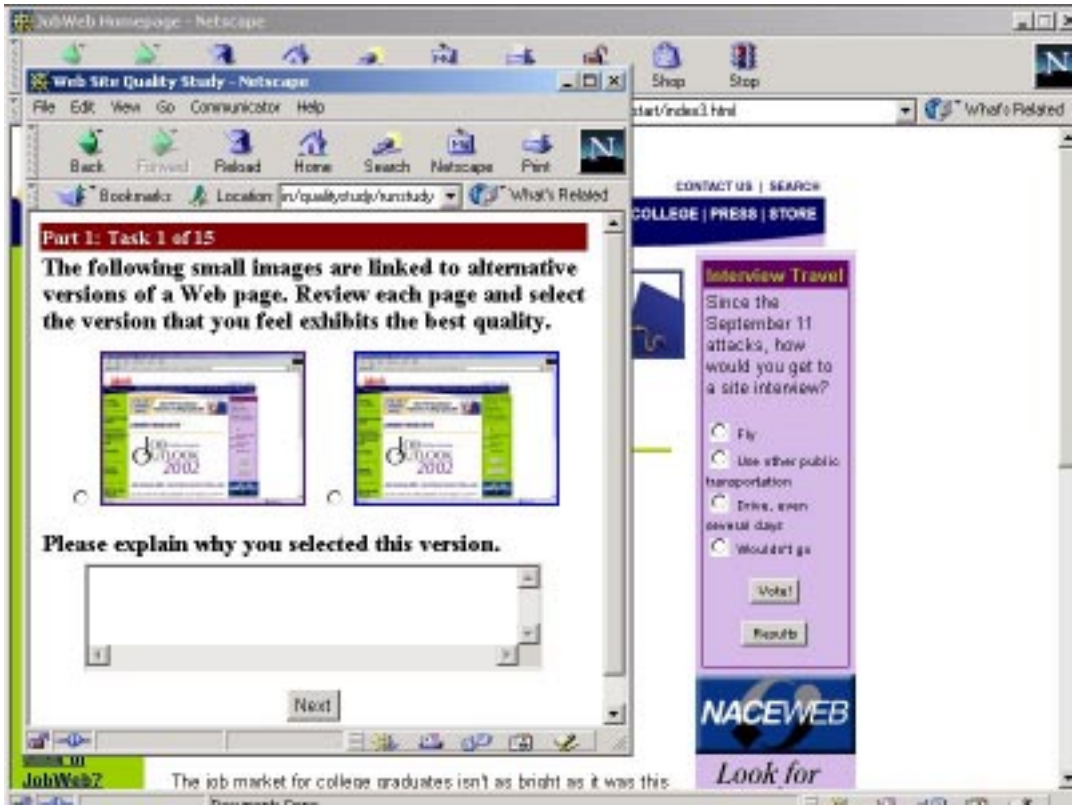

Figure 9.12: Testing interface for completing the page-level analysis tasks. The smaller browser window provides links to pages that are displayed in the larger window.

During the testing session, participants completed the page-level tasks and then completed the site-level tasks. The testing interface tracked whether participants explored all of the pages in both sections of the study. A message window appeared whenever participants did not explore all of the pages; participants were also restricted from moving forward in the study until they explored all pages. All of the links on pages were redirected to an error page reminding participants to stay focused on the individual pages.

9.2.5 Testing Environment

Participants completed the study in either the Kaiser Permanente Web Portal Services Group's computer training room or at their office computer; participants worked individually in both scenarios. All computer stations had PCs running Microsoft Windows NT 4.0 with 128 MB of RAM. Stations also had 15" or 17" monitors and high speed LAN connections. Participants used the Netscape Navigator 4.7 browser. The testing interface resized the browser window to 800 x 600 pixels (equivalent to a 15" monitor) and sites were not cached to provide a realistic experience. User surveys have shown that over 50% of Internet users access sites with 800 x 600 monitor resolution and 56K and slower connection speeds [DreamInk 2000].

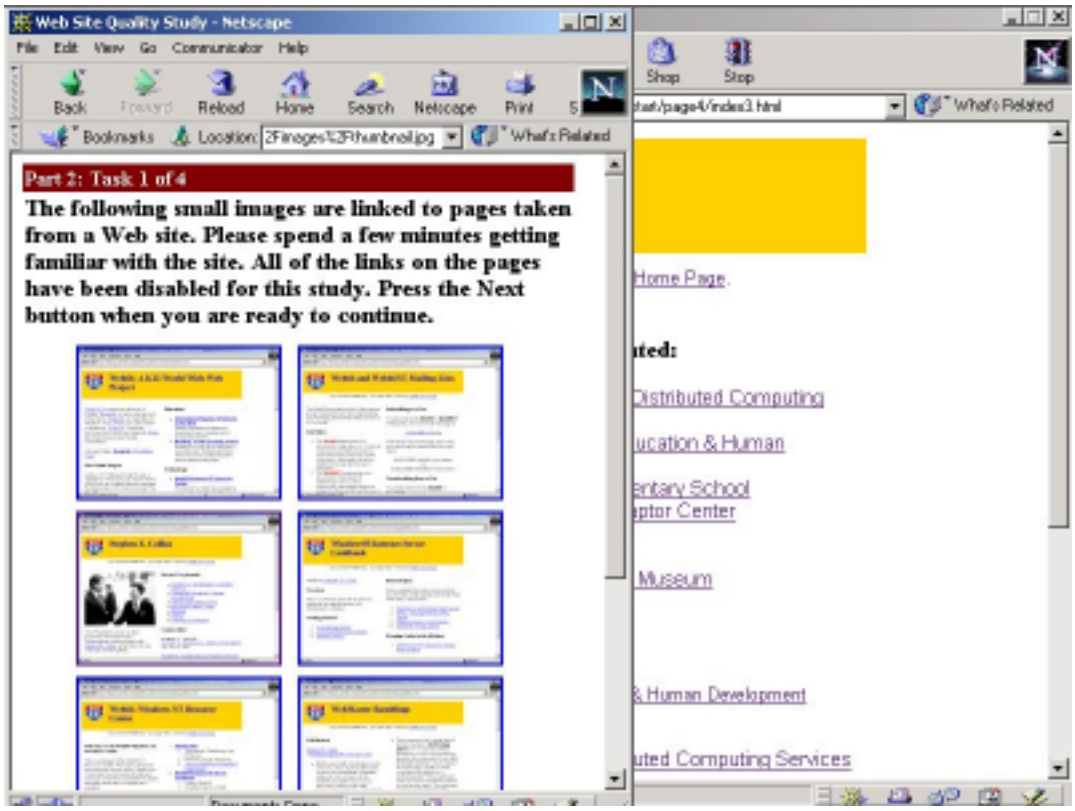

Figure 9.13: Testing interface for completing the site-level analysis tasks. The smaller browser window provides links to pages that are displayed in the larger window.

9.3 Data Collection

The analysis data consisted of 195 page-level (i.e., page preference) responses and 26 site-level (i.e., ratings of the original and corresponding modified sites) responses. Participants provided freeform comments in most cases. The site-level ratings were screened to replace small outliers with the next largest value as appropriate; the resulting data followed a normal distribution and exhibited equal variances.

9.4 Page-Level Results

The page-level analysis focused on testing the hypothesis that pages modified based on the overall page quality and the good cluster models are of a higher quality than the original pages. Recall that remodeling did not entail changing the content on pages, except for redistributing content over multiple pages, adding headings, etc. as dictated by the models. Thus, most of the differences between the original and modified pages were minimal and focused on the page layout and text formatting. The results showed that modified pages were preferred 57.4% of the time, while the original pages were preferred 42.6% of the time. The Chi-Square test revealed this difference to be significant ($\chi^2 = 4.3$, asymptotic significance of .038). ANOVAs revealed that there were no differences in preferences among the three Web design roles.

Figure 9.14 depicts results for the fifteen page-level comparisons. Participants preferred the modified pages in ten of the fifteen comparisons. Their comments about why they preferred the modified pages supported the changes made based on the profiles. In particular, participants felt that the modified pages were easier to read, required less scrolling, were cleaner, used better color schemes, made better use of whitespace, used headings, eliminated italics, and used better fonts. Several comments are provided below.

> "2 columns with shorter lines of te[x]t [is] easier to read. Primary navigation under the headline gives i[t] more prominence. Overall composition is better." (page pair 6, professional Web designer)

> "all page fits in window." (page pair 2, non-professional Web designer)

> "The text is easy to read. The text on the other page is to[o] long." (page pair 9, non Web designer)

Comments also revealed that in some cases, mainly with page pairs 1, 3, 4, 5, and 11, participants responded negatively to changes made based on the profiles. For example, modifications for the first page pair (site 3) included reorganizing graphical links and texts to minimize vertical scrolling. However, participants preferred the original version of this page because the width of text was restricted to fit within the browser window (i.e., the page did not require horizontal scrolling). This horizontal scrolling was introduced when the undergraduate student modified the page; this was an unfortunate mistake that arose because of the student's inexperience with HTML and Web design. This change did not impact the overall quality prediction, nor was it among the top ten deviations from the cluster model; most of the top ten deviations were associated with text and text formatting measures, such as the sans serif word count and text cluster and column counts.

Figure 9.14 shows that for four of the five pages on site 3 (page pairs 1–5), the original pages were preferred over the modified pages. As the Chi-Square test showed, this was not an expected outcome. Examining the model output for the modified pages revealed that other changes, such

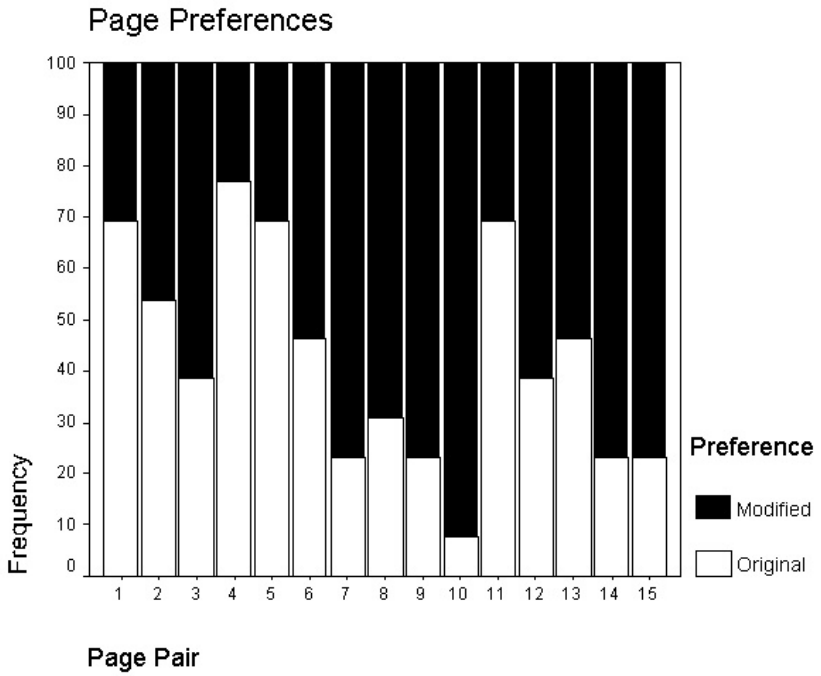

Figure 9.14: Results for the 15 page-level comparisons. Comparisons 1–5 were for pages from site 3. Comparisons 6–10 were for pages from site 4. Comparisons 11–15 were for pages from site 5.

as using sans serif fonts for text, were not implemented by the undergraduate student. Hence, it appears that the modified pages were insufficiently remodeled for comparison. If responses for pages on this site are excluded, then the results for the 2 remaining sites show that modified pages were preferred 66.9% of the time, while the original pages were preferred 33.1% of the time; this difference was highly significant ($\chi^2 = 14.9$, asymptotic significance of .000).

Several participants reported that in some cases, they could not tell the difference between the two pages and selected the first one; pair thirteen is one example. Given the restrictions on page modifications, it was not always possible to produce radically different designs. However, several participants also verbally commented on how subtle changes made a lot of difference.

9.5 Site-Level Results

The site-level analysis focused on testing the hypothesis that sites with pages modified based on the overall page quality and the good page cluster models are of a higher quality than the original sites. Similarly to the page-level analysis, the profiles were used to modify individual pages in the site. The overall site quality model was not used to assess the site given the discrepancies discussed in Section 8.6. Instead, the median overall page quality was used for site level assessments; as the quality of the individual pages improved, so did the median overall page quality. Table 9.5 showed that almost all of the original pages were rated poor overall, while almost all of the modified pages were rated good overall; the corresponding median overall page quality was poor and good, respectively.

Figures 9.15 and 9.16 depict the distributions of ratings for the original and modified sites; both distributions are nearly normal. Participants rated the quality of the original sites as 3.0 on average ($\sigma = 1.36$); however, they rated the quality of modified sites as 3.5 on average ($\sigma = 1.03$). A paired samples t-test revealed that this difference was significant ($p = .025$); this means that

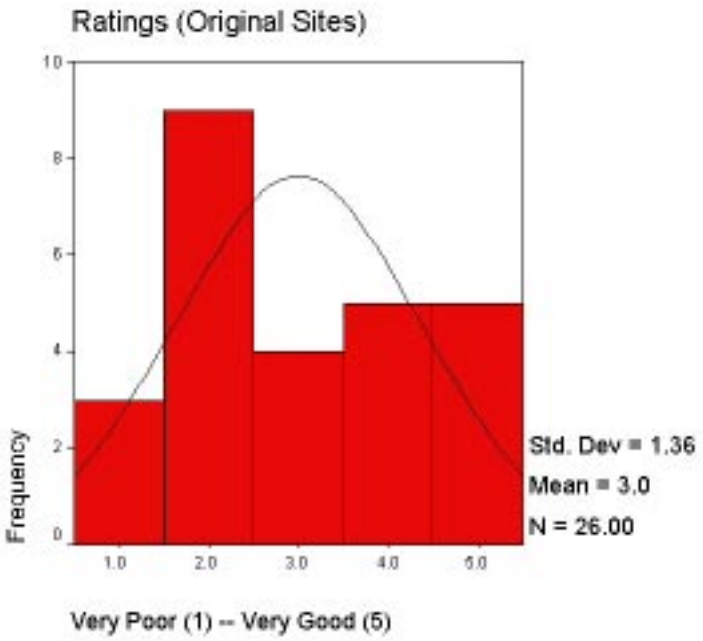

Figure 9.15: Distribution of ratings for the original versions of the 2 sites.

Design Role	Mean		Std. Dev.		Sig.
	Orig.	Mod.	Orig.	Mod.	
Professional Web Designer	2.63	3.38	1.30	0.92	0.020
Non-professional Web Designer	3.00	3.83	1.26	0.98	0.042
Non Web Designer	3.25	3.42	1.48	1.16	0.674

Table 9.5: Site ratings for the three groups of participants: 4 professional Web designers; 3 non-professional Web designers; and 6 non Web designers. Paired t-tests were used to compute the significance values.

each participant tended to rate the modified version higher than the original version. Similarly to the page-level analysis, participant comments provided support for many of the changes made based on the profiles.

There were differences in ratings among participants in the three roles. Table 9.5 shows a wider difference in mean ratings for the professional and non-professional Web designers; these differences are also significant. However, the table shows little difference in average ratings for the non Web designers, and the difference is not significant. Thus, it appears that the site level results are somewhat skewed possibly by the non Web designers' inability to gauge differences between the two versions of the sites. The comments showed that they often questioned whether they were rating the same site again. Some of these participants even stated that they rated both versions of the site the same.

For the analysis above, ratings for both sites were aggregated; however, participants' ratings for the individual sites were also examined. The modified versions of both sites were rated higher than the original versions, but the differences were not significant in most cases (see Table 9.6). Similarly to the discussion above, there were differences in ratings among the three participant groups. With respect to the example site discussed in Chapter 8 (site 1), Table 9.6 shows that in all cases participants rated the modified version slightly higher than the original site, although the

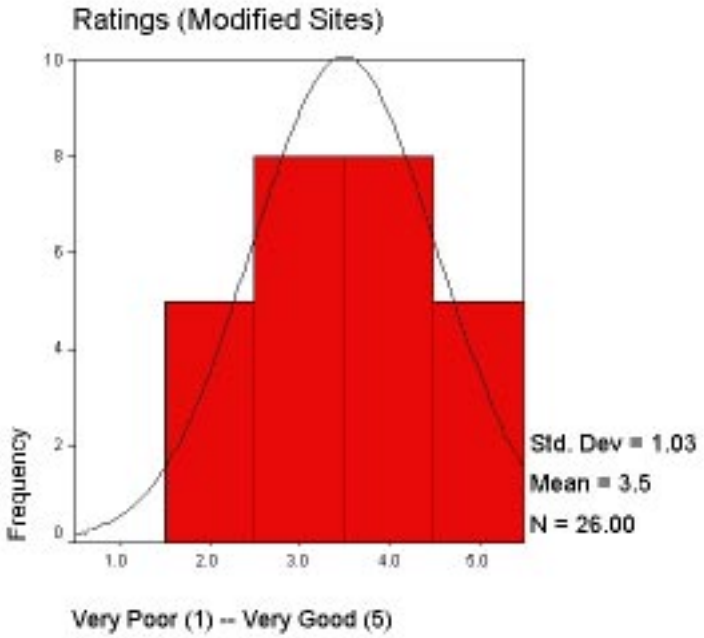

Figure 9.16: Distribution of ratings for the modified versions of the 2 sites.

difference was not significant in any case.

The intent of site remodeling was not to produce the best possible site and no claim is made about the modified sites exhibiting high quality. Due to the restrictions on page modifications (see Section 9.2.1), there were many issues that were not addressed, such as changing color schemes and adding images and links. Some of these issues were reported by the models and others, such as illegible text in images, were not. In many cases, participants' ratings and comments reflected their negative responses to these outstanding issues; it is likely that their responses also skewed the results. For example, Table 9.6 shows that professional Web designers rated sites lower than non-professional and non Web designers even though they rated the modified sites higher. The intent of site remodeling was to improve the quality of the designs in spite of the inherent problems; the results suggest that this goal was accomplished.

9.6 Summary

This chapter presented results from a user study of five Web sites wherein thirteen participants (four professional, three non-professional, and six non Web designers) completed two types of tasks: 1. explore alternative versions of Web pages and select the ones exhibiting the highest quality; and 2. explore pages from sites and rate the quality of the site. In preparation for the study, three students – two undergraduates and one graduate – used the overall page quality and good page cluster models to modify pages. The students were given access to an interactive appendix that summarized all of the quantitative measures and were able to ask the author questions about the models and measures as needed. The study demonstrated that it was possible for people other than the author to interpret and apply the models. The analysis focused on answering the following questions.

- Do participants prefer the modified pages over the original pages?
- Do participants rate the modified sites higher than the original sites?

Site	Mean		Std. Dev.		Sig.
	Orig.	Mod.	Orig.	Mod.	
All Roles					
1	3.62	3.92	1.26	1.04	0.219
2	2.38	3.08	1.19	0.86	0.069
Professional Web Designers					
1	3.50	3.75	1.29	1.26	0.391
2	1.75	3.00	0.50	0.00	0.015
Non-professional Web Designers					
1	4.00	4.67	1.00	0.58	0.423
2	2.00	3.00	0.00	0.00	–
Non Web Designers					
1	3.50	3.67	1.52	1.03	0.695
2	3.00	3.17	1.55	1.33	0.822

Table 9.6: Ratings for the 2 study sites. Paired t-tests were used to compute the significance values; the test did not return a significance value for non-professional designers' ratings for site 2.

Analysis of page-level data showed that participants preferred the modified versions of pages 57.4% of the time and preferred the original versions of pages 42.6% of the time; this difference was significant. Analysis of site-level data showed that participants rated the modified sites higher than the original sites, and this difference was significant. The site-level analysis also showed that the modified version of the site discussed in Chapter 8 was rated slightly higher than the original site; however, the difference was not significant. Participants' comments in both parts of the study provided support for many of the changes made to pages in the example site in Chapter 8 (and the other four) based on the Web interface profiles.

Modifications made to the pages and sites were very conservative for this first study. This is mainly because of the amount of effort required to manually make changes and because it was not clear if changes made based on the models actually improved quality; the latter was examined by the study itself. It is possible that less conservative changes would have resulted in larger differences in page preferences and site ratings. Future studies will be conducted to re-examine this question after recommendations and possibly modifications have been automated in some manner.