Evolution of Web Site Design Patterns

MELODY Y. IVORY and RODRICK MEGRAW
University of Washington

The Web enables broad dissemination of information and services; however, the ways in which sites are designed can either facilitate or impede users' benefit from these resources. We present a longitudinal study of web site design from 2000 to 2003. We analyze over 150 quantitative measures of interface aspects (e.g., amount of text on pages, numbers and types of links, consistency, accessibility, etc.) for 22,000 pages and over 1,500 sites that received ratings from Internet professionals. We examine characteristics of highly rated sites and provide three perspectives on the evolution of web site design patterns: (1) descriptions of design patterns during each time period; (2) changes in design patterns across the three time periods; and (3) comparisons of design patterns to those that are recommended in the relevant literature (i.e., texts by recognized experts and user studies). We illustrate how design practices conform to or deviate from recommended practices and the consequent implications. We show that the most glaring deficiency of web sites, even for sites that are highly rated, is their inadequate accessibility, in particular for browser scripts, tables, and form elements.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation/methodology, Screen design, Style guides*; H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia—*User issues*; I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems; K.4.2 [**Computers and Society**]: Social Issues

General Terms: Design, Human Factors

Additional Key Words and Phrases: World Wide Web, automated usability evaluation, web site design, design guidelines, accessibility, usability, empirical studies

1. INTRODUCTION

The World Wide Web plays an important role in our society—enabling broader dissemination of information and services than was previously available. Nonetheless, there is evidence that many sites have inadequate usability and accessibility [Forrester Research 1999; Jackson-Sanborn et al. 2002]. Clearly, the ways in which sites are designed can either facilitate or impede users' benefit from the vast resources that are available on the Web. What are the

WebTango research was supported by a Hellman Faculty Fund award, a Microsoft Research grant, a Gates Millennium Fellowship, a GAANN fellowship, and a Lucent Cooperative Research Fellowship Program grant. The Information School also provided funding for WebTango research.
Authors' address: University of Washington, 330C Mary Gates Hall, Box 352840, Seattle, WA 98119-2840; email: {myivory,remegraw}@u.washington.edu.

design characteristics of the most effective sites? Have these characteristics changed over time, perhaps due to technological advances, or are they still the same?

Our goal in this article is to examine these questions. Consequently, we present a longitudinal study of web site design patterns for the time period that spans from 2000 to 2003. In prior work, we developed a comprehensive set of over 150 quantitative measures to assess web interface aspects (e.g., amount of text on pages, numbers and types of links, consistency, accessibility, etc.) [Ivory 2001]. During our study period, we computed these measures for over 22,000 pages and over 1,500 sites that received ratings from Internet professionals. We used the measures and ratings to build accurate statistical models to distinguish highly rated interfaces [Ivory et al. 2000, 2001; Ivory 2001; Ivory and Hearst 2002a, 2002b].

We expand upon our prior work by using these measures to examine characteristics of interfaces that were highly rated. In early studies, we discussed characteristics of highly rated interfaces during the 2000 time frame [Ivory 2001, 2003b]. In this article, we provide three perspectives on the evolution of web site design: (1) descriptions of designs during each time period; (2) descriptions of changes in designs across the three time periods; and (3) comparisons of design patterns to those that are recommended in the relevant literature (i.e., texts by recognized experts and user studies; e.g., National Cancer Institute [2001]; Nielsen [2000]; Spool et al. [1999]; van Duyne et al. [2002]; and W3C [1999]). We illustrate how design practices conform to or deviate from recommended practices and the consequent implications. For instance, we show that the most glaring deficiency of web sites, even for sites that are highly rated, is their inadequate accessibility, in particular for browser scripts, tables, and form elements.

We begin with a discussion of related studies on web site design practices and design guidelines. As background to our analyses, we discuss our prior work with respect to building models to facilitate automated web site evaluation (Section 3). Our model-building work and our analyses of design patterns are complementary, because they both use the same measures and datasets. We provide an overview of our design pattern analyses within Section 4. Sections 5–7 present: characteristics of Web site designs during 2000, 2002, and 2003; design changes during the three years; and comparisons of design patterns to recommended practices.

2. RELATED WORK

Our primary aim in this article is to: (1) examine changes in web site design patterns during our study period and (2) contrast those changes to recommended design practices. We discuss relevant work along these two strands in this section. We provide an overview of design guidelines and patterns, problems faced by designers in applying them, and some approaches for helping to mitigate the problems that they face. We then discuss related studies on web site design.

2.1 Design Guidelines and Patterns

A plethora of design guidance exists (e.g., Brinck et al. [2001]; Computer Science and Telecommunications Board [1997]; Comber [1995]; Detweiler and Omanson [1996]; Flanders and Willis [1998]; Fleming [1998]; Koyani et al. [2003b]; Nielsen [2000]; Spool et al. [1999]; van Duyne et al. [2002]; and W3C [1999]). Most resources provide prescriptive guidance about how to design sites and offer the advantage of being applicable throughout the entire development process. Prescriptive guidance is often voluminous, vague, conflicting, or divorced from the context in which sites are being developed, thus making it difficult to apply.

Many guidelines have not been validated empirically and there is little overlap across guideline sets [Ratner et al. 1996]. A noteworthy exception is the research-based guidelines that researchers at the National Cancer Institute developed [Koyani et al. 2003a, 2003b; National Cancer Institute 2001]. The resource consists of 187 peer-reviewed guidelines, with each guideline containing: an overarching principle, research or supporting information, citations to relevant literature, a score to indicate the "strength of evidence" that supports the guideline, a score to indicate the "relative importance" of the guideline to the overall success of a site, and graphical depictions of guideline conformance in practice.

Designers may not even be able to successfully use validated guidelines. Historically, designers have found it difficult to apply design guidelines [Borges et al. 1996; Lowgren and Nordqvist 1992; Smith 1986; Souza and Bevan 1990]. Our four studies of web designers show that both novice and professional designers experience difficulties in applying guidelines effectively [Chevalier and Ivory 2003a, 2003b]; consequently, such resources are underused [Barry and Lang 2001; Vora 1998]. Our 2002 survey of web practitioners (designers, usability engineers, information architects, webmasters, etc.) revealed that only 36 percent of them always use guidelines when designing sites [Ivory et al. 2003].

One way to help designers conform to guidelines is to incorporate them into design tools. For instance, Damask is a sketch-based design tool that enables designers to leverage common web design patterns like L-shaped navigation bars or specific task completion sequences [Lin and Landay 2002]. The tool incorporates sketches of design patterns that van Duyne et al. [2002] developed. Damask does not have semantics associated with sketch objects other than text areas (page titles, headers, or body text); associating semantics with design objects may be helpful in providing feedback to designers and in evaluating early design representations.

In our work, we take another approach to improving guideline conformance: automated web site evaluation (Section 3). What distinguishes our methodology from other design guidelines, design patterns, and similar automated evaluation tools is our use of: (1) quantitative measures, (2) empirical data to develop guidelines, and (3) profiles as a comparison basis. Our approach makes it possible to conduct context-sensitive analysis (e.g., based on the site's genre or page's style). Another distinction of our approach is that it is possible to leverage the

empirical data to derive thresholds for quantitative measures and to compare these thresholds to design guidelines. We demonstrate this capability within this article.

2.2 Studies on Web Site Design

There have been numerous studies on web site design, mainly examining accessibility issues (see Schmetzke [2004a] for a survey). Most studies entail evaluating the accessibility of sites within academic, government, and library environments (e.g., accessibility of university home pages, department sites, library sites, and government sites). A major limitation of the existing studies is that their analyses are based solely upon errors reported by Bobby [WatchFire 2002], an automated web site evaluation tool. We include Bobby-reported errors as only five of the 157 quantitative measures that we use to characterize and analyze web interfaces.

These studies provide valuable insight about accessibility issues, which we confirm through our analysis. For example, one study of forty-five sites from three genres—education, government, and e-commerce—found that 47, 53, and 0 percent, respectively, were approved by the Bobby tool [Jackson 1999]. A study of 549 sites from six categories, which included government, education, job listing, and most popular sites, showed that 66 percent of the sites were not approved by Bobby [Jackson-Sanborn et al. 2002]. An interesting finding was that, of the sites that were considered the most popular, only 15 percent were approved by Bobby. A longitudinal study from 1999 to 2001 of web pages at the University of Wisconsin's thirteen campuses revealed a gradual improvement in accessibility [Schmetzke 2004b]; however, the highest Bobby approval rating was only 53 percent. Our study suggests that this positive trend may not be the case for sites in general. Furthermore, we found much lower approval rates (maximum of 24 percent); the lower rate is most likely attributable to the larger sample sizes that we used.

Related longitudinal studies of web site design examine the rate of content and other design changes [Brewington and Cybenko 2000; Cho and Garcia-Molina 2003; Ntoulas et al. 2004]. Although these studies have different motivations (understanding rates and types of design changes and their implications for Internet search engines), their results confirm that web site designs do change. For example, in one study, researchers analyzed 720 thousand pages on a daily basis over four months and found that 40 percent of the pages changed within a week, while 23 percent of the business pages (i.e., .com domain) changed daily [Cho and Garcia-Molina 2000]. In another study, researchers analyzed 151 million pages on a weekly basis over eleven months. During page analyses, researchers recorded several statistics, including two measures that we use—the length of pages (i.e., byte sizes) and number of words [Fetterly et al. 2003, 2004]. About 67 percent of pages did not change during the study period. These studies suggest that our yearly (as opposed to a more frequent) time-frame of analysis is appropriate for examining the evolution of web site design.

3. WEBTANGO METHODOLOGY

Our methodology entails deriving design guidance (i.e., prevalent design patterns) by examining well-designed interfaces [Ivory 2001, 2003a]. Our objective was to enable designers to compare their designs to those that are well-designed, to determine whether their designs exhibit similar properties, and, if not, to determine how their designs differ. We discuss briefly how we accomplish this objective.

3.1 Overview

The WebTango approach involves five steps.

(1) Identifying an exhaustive set of quantitative interface measures.
(2) Computing measures for a large sample of rated interfaces.
(3) Deriving statistical models from the measures and ratings.
(4) Using the models to predict ratings for new interfaces.
(5) Validating model predictions.

The methodology comprises two distinct, yet related phases: (1) establishing an interface quality baseline and (2) analyzing interface quality. For phase one, we complete all five of the steps discussed above. For phase two, we complete steps two and four; measures computed during step 2 are only for the site that is being evaluated. During the first phase, we couple the page- and site-level measures with expert ratings of sites and apply data mining algorithms (i.e., automated knowledge discovery) [Witten and Frank 1999] to the aggregate data to build profiles (predictive models) of highly rated interfaces. Profiles encapsulate key quantitative measures, thresholds for these measures, and effective relationships among measures; they represent an interface quality baseline.

During the second phase, we compare page- and site-level measures for a site to the developed profiles, as a way to assess its quality. (Evaluated sites are usually not the same sites that we use to develop profiles.) After we develop profiles, designers can use them in an ongoing manner to evaluate sites. We periodically repeat (e.g., annually or semi-annually) the interface quality baseline phase to ensure that profiles reflect current web design practices. In this article, we summarize the measures and datasets that we used for our model-building efforts.

3.2 Identification of Web Interface Measures

The first step of the WebTango process entails identifying an exhaustive set of quantitative measures to assess as many aspects of web interfaces as possible. Based on our extensive survey of design recommendations from recognized experts and usability studies (e.g., [Flanders and Willis 1998; Fleming 1998; Nielsen 1998, 1999, 2000; Rosenfeld and Morville 1998; Sano 1996; Schriver 1997; Shedroff 1999; Shneiderman 1997; Spool et al. 1999), we developed 157 page- and site-level measures. Our objective was to first quantify features discussed in the literature and then to determine their importance in producing high-quality designs.

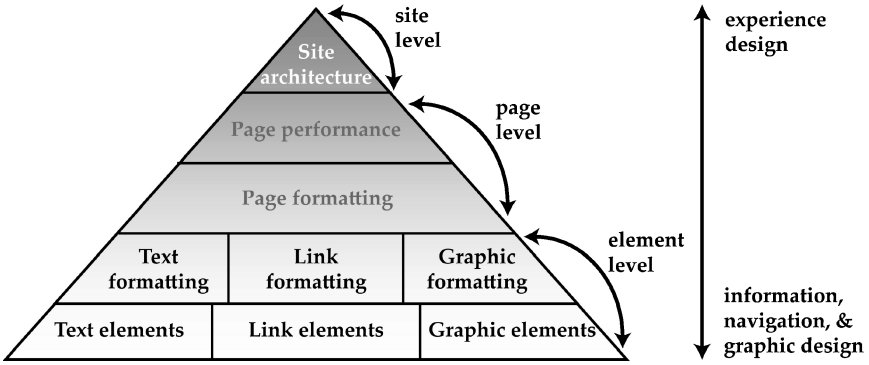

Fig. 1. Conceptual model of web interfaces. Text, link, and graphic elements are the building blocks. Page- and site-level features use these elements to improve the user's experience.

We begin with a conceptual model of web interfaces and then present a summary of the 157 measures. The reader can find an in-depth discussion of them in Ivory [2001, Chapter 5]. We also developed an interactive appendix to illustrate all measures; the reader can access HTML and PowerPoint versions on the WebTango Project's web site (http://webtango.ischool.washington.edu).

3.2.1 *Conceptual Model of Web Interfaces.* We originally developed the WebTango approach mainly for information-centric web sites (i.e., sites whose primary tasks entail locating specific information) as opposed to functionally oriented ones (i.e., sites wherein users follow explicit task sequences). An information-centric web interface is a mix of text, link, and graphic elements, formatting of these elements, and various aspects that affect its usability, accessibility, and quality. Web interface design entails a complex set of activities for addressing these diverse aspects—information, navigation, graphic, and experience design (see Ivory and Hearst [2002a] for a discussion). They deal with information organization, navigation mechanisms, visual presentation or layout, and overall experience, respectively.

Information, navigation, graphic, and experience design can be further refined into the aspects depicted in Figure 1. The figure shows that text, link, and graphic elements are the building blocks of web interfaces; all other aspects are based on them. The next level of Figure 1 addresses formatting of these building blocks, and the subsequent level addresses page-level formatting. The top two levels address the performance of pages and the architecture of sites, including the consistency, breadth, and depth of pages within sites. The bottom three levels of Figure 1 are associated with information, navigation, and graphic design activities, while the top two levels—Page performance and Site architecture—are associated with experience design activities. (All levels influence the users' experience with a site.) We developed measures to assess as many of these interface features as possible.

3.2.2 *Web Interface Measures.* After conducting an extensive survey of the web design literature, we enumerated 62 design features of web interfaces, including: the amount of text on a page, fonts, colors, consistency of page layout

Table I. Summary of 157 Measures Used for Web Interface Evaluation. Each Category Corresponds to a Block in Figure 1

Category	Number of Measures	Aspects Measured
Text elements	31	Amount of text, type, quality, and complexity. Includes visible and invisible text.
Link elements	6	Number and type of links.
Graphic elements	6	Number and type of images.
Text formatting	24	How body text is emphasized; whether some underlined text is not in text links; how text areas are highlighted; font styles and sizes; number of text colors; number of times text is repositioned.
Link formatting	3	Colors used for links and whether there are text links that are not underlined.
Graphic formatting	7	Minimum, maximum, and average image width and height; page area covered by images.
Page formatting	27	Color use, fonts, page size, use of interactive elements, page style control, and so on. Key measures include evaluating the quality of color combinations (for text and panels) and predicting the functional type of a page.[a]
Page performance	37	Page download speed; page accessibility for people who have impairments; presence of HTML errors; and "scent" strength.[b]
Site architecture	16	Consistency of page elements, element formatting, page formatting and performance, and site size (number of pages or documents).[c]

[a]The decision tree for predicting page type—home, link, content, form, or other—exhibited 75 percent accuracy on average for 1,770 pages.
[b]The linear regression model for predicting download speed exhibited 86 percent accuracy on average. We use output from Bobby 3.2 [WatchFire 2002] for our accessibility measures. We report the total number of HTML errors determined by Weblint 1.02 [Bowers 1996]. To assess scent quality, we report word overlap between: the source page's text and the destination page's text, the source link's text and the destination page's text, and the source and destination pages' titles.
[c]Consistency measures are based on coefficients of variation (standard deviation normalized by the mean) across measures for pages within the site. The site size measure only reflects the portion traversed by our crawler. We compute these measures only if there is page-level data for at least five of the site's pages.

across the site, use of framesets, and others [Ivory 2001]. We then developed 157 quantitative measures to assess 56 of the 62 features (90%); not all features could be assessed in an automated manner. Table I describes the measures; the measures adhere to guidance on developing effective performance metrics [Jain 1991]: all but one of the measures are not ratios of two or more measures (low variability), no measures convey the same information (non redundancy), and measures reflect most aspects of web interfaces (completeness). Because most site-level measures examine variation across pages, we only compute them when there are page-level measures for at least five of a site's pages. We determined the five-page criterion during early experiments with the site-level measures.

We assessed the accuracy of computed measures by comparing manual and automated computations for a sample of fourteen web pages, which had different design characteristics. With three exceptions, all measures were 84 percent accurate on average. The least accurate measures—number of changes in text

alignment from flush left (text positioning count) and areas highlighted with color, rules, lists, and so on (cluster counts)—require image processing to improve accuracy. We have begun to explore the use of image and document processing techniques to improve measure accuracy [Harrison and Shin 2003].

3.3 Data Collection

We developed a Crawler Tool and a Metrics Computation Tool to facilitate building a corpus of interfaces for model building [Ivory 2001]. We run the crawler on a set of pre-determined URLs to create a local copy of each site; the crawler downloads a random subset of information-centric pages from each site. All pages that could be read with a standard HTML browser, including pages generated dynamically by server-side scripts, were crawled for the data collection.

We run the metrics tool on the downloaded pages to compute the 157 page- and site-level measures. Recall from Section 3.2.2 that site-level measures are computed only for sites from which the crawler was able to download at least five information-centric pages. We normalize data independently within each of the three classes (good, average, and poor interfaces; described below) to eliminate outliers and to induce a normal distribution to the data.

In 2000, 2002, and 2003, we used identical quantitative measures and processes to produce datasets for model building. Our process included use of the Webby Awards as a source for rated sites [The International Academy of Arts and Sciences 2000]. The Webby Awards is a unique resource, as it is the largest collection of sites that are rated along one set of criteria and organized into over 27 content categories or site genres (e.g., finance, art, education, and news). A panel of over 100 Internet experts (web designers, writers, artists, etc.) use a 10-point scale to rate sites on overall experience and five specific criteria: content, structure and navigation, visual design, functionality, and interactivity. Judging takes place in three stages: review, nominating, and final; only the list of nominees for the final round are available to the public. Anyone can nominate any site to the review stage, and at least two reviewers inspect and rate each site.

Sites submitted for the review stage represented the full spectrum of interface quality (i.e., from poorly to highly rated sites) and captured the heterogeneity of web sites and pages. Further, our early usability test of 57 reviewed sites from the 2000 dataset suggested judges' ratings to be somewhat similar to usability assessments [Ivory 2001, Chapter 7]. (Our findings in the remaining sections suggest that similarity between judges' ratings and usability assessments may no longer be the case, especially when accessibility is considered.) Hence, we used sites and ratings from review stages to build our datasets (Table II). Before building our 2003 dataset, we modified the crawler to address HTML coding issues that previously caused it to malfunction. Table II shows that the repairs increased the median number of pages that the crawler was able to retrieve from sites.

Given that review stage sites range in ratings from very low to very high (i.e., 1 to 10), we used ratings for the overall experience criterion to group sites into three classes: good (top 33 percent of ratings), average (middle 34 percent

Table II. Summary of Datasets Used for Model Building. Median Pages per Site is Based on the Number of Pages Analyzed on Each Site

	Dataset		
Statistic	2000	2002	2003
Total sites analyzed	333	570	668
Total pages analyzed	5,346	4,483	12,599
Median pages per site	5	6	9

Table III. Summary of Datasets Used for Model Building by Quality Ratings (Classes). Percentages are Based on the Total Pages and Sites Analyzed (Table II)

	Dataset		
Rating	2000	2002	2003
Pages by rating			
Good	1,906 (36%)	1,409 (31%)	5,067 (40%)
Average	1,835 (34%)	1,644 (37%)	3,927 (31%)
Poor	1,605 (30%)	1,430 (32%)	3,605 (29%)
Sites by rating			
Good	121 (36%)	183 (32%)	245 (37%)
Average	118 (35%)	204 (36%)	217 (32%)
Poor	94 (29%)	183 (32%)	206 (31%)

of ratings), and poor (bottom 33 percent of ratings). We also applied a site's rating to its associated pages; the assumption is that if a site is rated as good, average, or poor, then all pages in the site are of that same rating. Our early study of web site usability and model-building efforts showed the assumption to be reasonable [Ivory 2001]; however, our findings in in the remaining sections suggest that this applicability assumption may no longer be valid, especially for recent datasets.

Table III shows a breakdown of the datasets based on quality ratings. Datasets contain roughly equal proportions of good, average, and poor interfaces, except for the 2003 page sample. One of our main model-building objectives was for assessments to consider the context in which sites and pages are designed. We consider this context by segmenting the data based on content category and page type and building separate models for these subsamples. When a site is submitted to the Webby Awards, the submitter assigns one or more content categories to it. We use this metadata to build a model for each content category. Similarly, we use predicted page types (mentioned in Table I in Section 3.2.2) to build a model for each of the five page types—home page, link page, form page, content page, or other page. We incorporate accurate prediction models into an online tool (Analysis Tool) to enable designers to compare aspects of their designs to highly rated ones.

4. DERIVATION OF DESIGN PATTERNS

To understand how designs evolved between 2000 and 2003, we analyzed quantitative measures for the three datasets and report on specific changes in the remaining sections. First, we discuss the salient features of highly rated

interfaces within each dataset (Section 5). By salient features, we mean quantitative measures (and related design patterns), which played an important role in distinguishing pages and sites within the good, average, and poor classes. Second, we discuss salient features of interfaces across the three datasets (Section 6). For this analysis, we identified quantitative measures (and related design patterns), which played an important role in distinguishing pages and sites within each dataset. Finally, we discuss design patterns for the amount of text on pages, numbers of links and graphics, consistency across pages, and other aspects (Section 7). In this final analysis, we contrast observed design patterns within the three datasets to recommended practices. By showing trends in design guideline conformance, we expand upon our early discussions [Ivory et al. 2000, 2001; Ivory 2001, 2003b; Ivory and Hearst 2002a, 2002b].

In these sections, we use the term *good pages* as a shorthand for pages that came from sites that were rated the highest by the Webby judges. Similarly, we use the terms *good sites*, *poor pages*, and so on. We make no claim that good pages or sites, nor the design patterns that we extracted from them, correspond to highly usable or accessible interfaces. Furthermore, we make no claim that the derived thresholds are completely accurate; measures may not be computed accurately in all instances, which could affect range values. Our intention is to demonstrate how the patterns changed, aligned, or deviated from recommended practices.

5. CHARACTERISTICS OF WEB SITE DESIGNS

To characterize page designs for each dataset, we identified measures that played a significant role in distinguishing pages within the good, average, and poor classes. We computed a one-way analysis of variance (ANOVA) [Easton and McColl 1997] based on class to determine each measure's effect in distinguishing pages. We ranked measures in descending order by their F ratios; hence, measures that had the ten largest F ratios and significance at least at the $p = .05$ level represent the top ten predictors. The top ten predictors for the three datasets comprised 27 different measures; only three measures were predictors for more than one dataset. This lack of overlap in measures suggests major design changes during the time period.

We conducted a similar analysis of the top five predictors for distinguishing sites within the good, average, and poor classes. This analysis revealed a 33 percent overlap in predictors; three of the seven identified measures were key predictors for at least two datasets. In the page- and site-level analyses, key predictors revealed interesting design patterns. In Section 7, we revisit these patterns, present thresholds for measures, and compare the thresholds to design guidance.

5.1 2000 Designs

The top ten predictors revealed major differences in designs for pages within the good, average, and poor classes. There were no significant differences for site-level measures. We describe the key characteristics of pages in the good class below. Figure 2 depicts an example page that we use to organize our discussion.

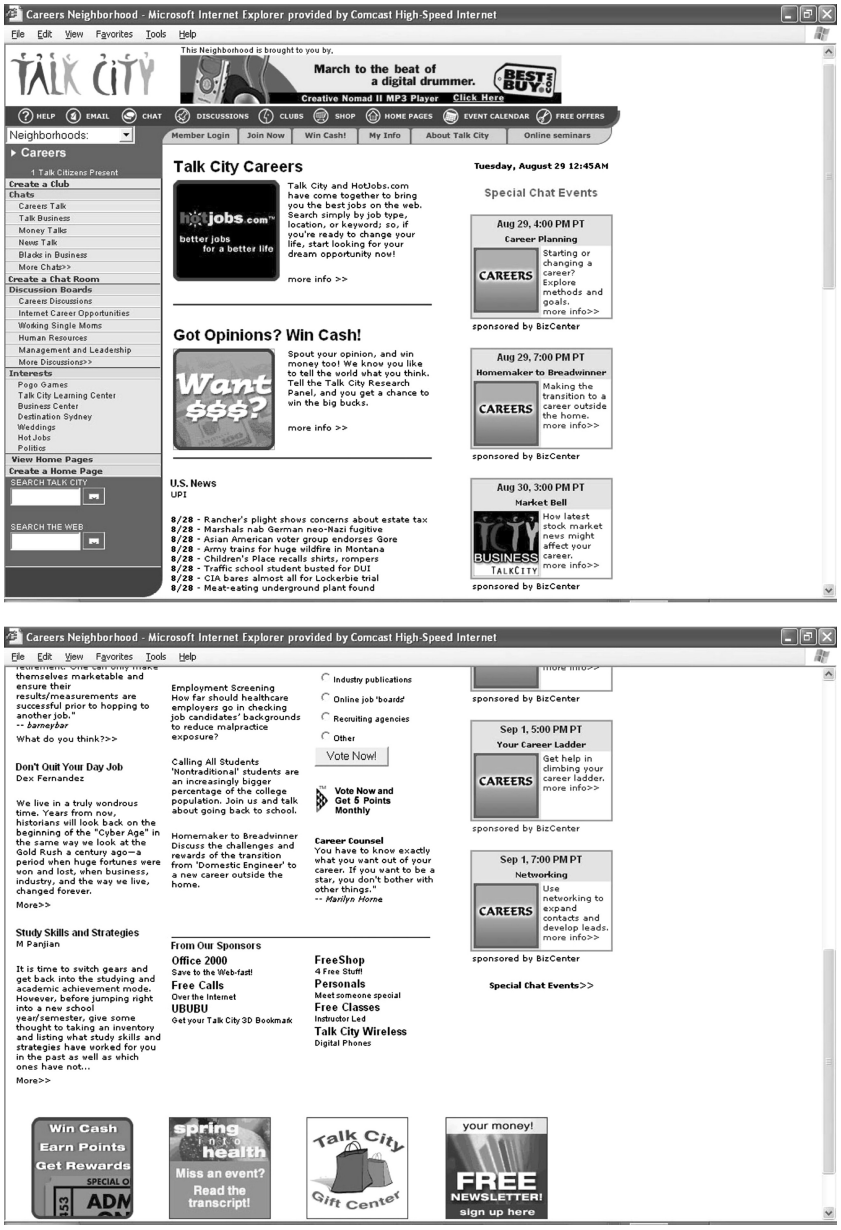

Fig. 2. Example page from the archived 2000 dataset (good class). The page depicts typical design features: large numbers of textual and internal links, graphical and animated ads, interactive objects, and use of link clustering. The page uses a fixed screen width.

—**Links.** Good pages tended to use text for hyperlinks, rather than graphics. Links were organized into clusters (e.g., navigation bars or bulleted lists of links). Most links pointed to pages that were internal to the site versus external to it. The example page in Figure 2 contains 87 links (84 are internalto

the site) that are described with 149 words (i.e., most links are textual). Most links are grouped into four clusters (navigation areas). The page also depicts the L-shaped navigation pattern that is used on many web sites [van Duyne et al. 2002].

—**Text Formatting.** We discovered that the good pages were not likely to contain italicized body text (i.e., text that is not in headings or links). Italicized text, especially long passages, may be difficult to read on computer screens [Schriver 1997]. Good pages were also likely to use small font sizes for some text (mainly for footers) and at least one accent color (i.e., a color that is used sparingly). The example page contains only two italicized body text words, uses a 9-pt font size for most of the text (which is not necessarily an effective design practice), and uses an accent color.

—**Advertisements.** Good pages tended to contain graphical ads, possibly as a sign of the site's success or credibility [Fogg et al. 2001]. We discovered that pages within the poor class also contained graphical ads; the difference was that the ads were not for easily recognizable companies (i.e., businesses that are popular or have off-line presences). In both cases, graphical ads tended to be animated (which is not necessarily an effective design practice). The example page contains an ad from Best Buy (a recognizable company) at the top of the page. Ads that are at the top and middle of the page (top screen) are animated. Three of the four ads that are at the bottom of the page (bottom screen) are animated. The page has thirteen graphical ads; 38 percent are animated.

—**Interactivity.** Interactive objects enable users to send information to the site as opposed to just receiving it. Good pages tended to have interactive objects (i.e., text fields, buttons, and other form elements). Typically, interactive objects appeared on every page as compared to appearing on certain types of pages (e.g., form pages). For instance, search functionality or mailing list sign-up components were embedded into header areas or navigation bars on every page. The example page contains twelve interactive objects: search functionality and a drop-down list which appear in the left column of every page and a form which users can use to respond to a poll (near the right of the page in the bottom screen).

—**HTML Coding.** Good pages tended to have a large number of HTML coding errors reported by Weblint (which is not necessarily an effective design practice). The example page contains 95 HTML coding errors. Reported errors include: improper order of tags (e.g., <p> tag in <head> element), use of unknown attributes, improper use of meta-characters (e.g., use of > versus >), and absence of closing tags. Most errors were repetitive, but the way in which Weblint reported them (i.e., each instance reported as a separate error) leads to inflated error counts. The Bobby tool uses a different reporting approach; it reports unique errors along with the number of times they occur (i.e., each instance). In our studies of automated web site evaluation tools [Ivory and Chevalier 2002; Ivory et al. 2003], we discussed the implications of such error reporting for designers.

5.2 2002 Designs

All the patterns that we described in the preceding section were also present within 2002 interfaces; however, the 2000 patterns were less salient. For instance, pages within the average class had textual links, link clusters, animated graphical ads, and other features to the same degree as pages within the good class. Only two patterns remained salient—use of italicized body text and interactive objects. The other predictors revealed major differences in the designs of pages and sites. We describe the key characteristics of good pages and sites below. Figure 3 depicts pages from an example site that we use to organize our discussion.

—**Text Formatting.** Similarly to the 2000 pages, good pages did not contain italicized body text. They used serif fonts to format text to a lesser degree than 2000 pages and 2002 poor pages. Studies suggest that on computer screens, sans serif fonts may be more legible than serif fonts [Bernard and Mills 2000; Nielsen 2000; Schriver 1997]. Pages in Figure 3 use the Arial font (sans serif).

—**Graphics.** Pages contained redundant or repeated graphics, for instance as list bullets or spacers to control page layouts. Pages also had small minimum graphic heights and widths (e.g., one pixel), which is indicative of spacers. For example, the top page in Figure 3 contains 56 redundant graphics, mainly to control the page's layout; the minimum graphic width and height is one pixel.

—**Color.** We measured the use of color within a page's HTML and stylesheets; however, we did not measure the use of color within images, applets, scripts, and other elements. Within the scope of what we measured, we found that good pages used several colors, for instance to cluster text and links so that they stand out. Pages in Figure 3 depict the use of multiple color clusters; furthermore, different color schemes are used across pages.

The Analysis Tool uses findings from Murch's study of color on computer screens to assess the quality of color combinations; Murch's study revealed high-contrast, low-contrast, and somewhat neutral color combinations [Murch 1985]. The tool evaluates color combinations used for text (i.e., the foreground and background colors used for text or thin lines) and panels (i.e., the foreground and background colors used for regions like navigation bars or thick lines).

Good pages used high-contrast text color combinations but low-contrast panel color combinations (which is not necessarily a good design pattern). The example page at the top of the figure uses six high-contrast text color combinations (black text on light-green, white, medium-beige, medium-green, and light-beige backgrounds); the remaining text color combinations (e.g., medium-green text on light-green background) are reported as being neutral (i.e., neither effective nor ineffective). The page also uses four panel color combinations that are considered to have low contrast: light-green, medium-beige, and medium-green regions on white backgrounds and white regions on a medium-green background.

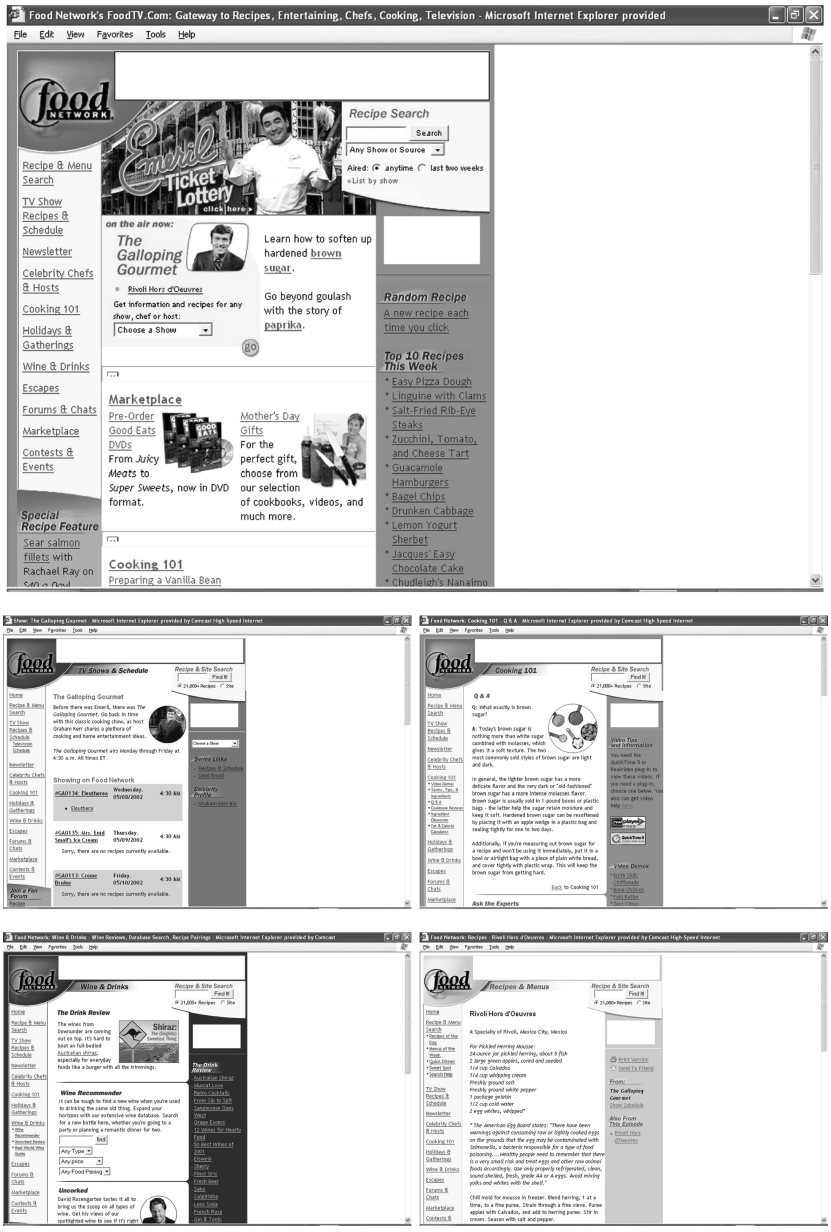

Fig. 3. Example pages from the archived 2002 dataset (good class). Pages depict typical design features: use of sans serif fonts, use of color to highlight links and text, and variation in page formatting. White, rectangular areas are placeholders for graphical ads; the crawler was not able to download the ads due to browser scripts. Pages use fixed screen widths.

—**Interactivity.** Similarly to the 2000 pages, 2002 pages contained interactive objects. The example page at the top of the figure has search functionality and a drop-down list for navigation; however, interactive elements did not appear on every page.

—**HTML Coding.** Good pages tended to have Priority 1 coding errors reported by Bobby (which is not necessarily an effective design practice). The example page contains two Bobby Priority 1 errors—alternative text not provided for two images and missing titles for frames (inline frames for ads).

—**Formatting Consistency.** Sites in the good class had some variation in the formatting of links and page layouts. Pages in Figure 3 demonstrate this variation. For instance, pages have similar layouts, but variation in color use, text columns, and text formatting. There is also some variation in link formatting, mainly for links that are in the middle of pages. Typically, pages did not have major structural changes like using completely different page layouts.

5.3 2003 Designs

Patterns from the 2000 and 2002 interfaces were also present within 2003 designs, but they were less salient. Only the use of interactive objects was salient across all three datasets. Four of the 2002 patterns were salient in 2003—use of color, graphics, repeated graphics, and formatting variation. The other predictors revealed differences in the designs of pages and sites. We describe the key characteristics of good pages and sites below. Figure 4 depicts pages from an example site, which we refer to in our discussion.

—**Text Formatting.** Good pages were likely to use multiple columns for text (i.e., vertical positions where text is placed). Unlike page designs for prior years, some positioning was not in alignment with delineated text columns (which is not necessarily an effective design practice). For instance, the top screen in Figure 4 contains a textual navigation bar at the top of the page and indented links on the left of the page; a total of five text columns are used.

—**Color.** Good pages tended to use numerous colors for navigation bars, body text, and links. Similarly to 2002 designs, they used high-contrast text color combinations. They also used high-contrast panel color combinations, unlike the 2002 designs.

—**Graphics.** Similarly to 2002 designs, good pages tended to use a large number of graphics; many of them were redundant graphics (e.g., spacers). For instance, the example page at the top of the figure has 34 graphics and 14 are repeated spacer images. The remaining graphics are used for navigation, advertisement, and other ornamentation.

—**Interactivity.** Similarly to the 2000 and 2002 pages, 2003 pages contained interactive objects. The example page at the top of the figure has search functionality (top of page) and a form, which is not depicted. Search functionality appears in the top navigation bar on every page within the site.

—**HTML Coding.** Good pages had browser-compatibility errors reported by Bobby (which is not necessarily an effective design practice). The example page contains 17 errors; most errors referred to compatibility with the Lynx 2.7 Browser (e.g., unknown or missing attributes in HTML tags).

Fig. 4. Example pages from the archived 2003 dataset (good class). Pages depict typical design features: text starting at various vertical positions, use of color and graphics, use of good text and panel color combinations, and variation in page formatting. Most pages do not use fixed screen widths.

—**Formatting Consistency.** Similarly to 2002 sites, 2003 sites had some variation in page layouts. Pages in Figure 4 demonstrate this variation. Unlike the 2002 designs, 2003 sites had structural changes in page layouts. Forinstance, the left navigation bar does not appear on all pages, page widths

change, and the right column is not used on all pages within the example site. Such structural changes, inconsistent navigation in particular, can impede usability and accessibility [Flanders and Willis 1998; Fleming 1998].

6. SIGNIFICANT DESIGN CHANGES

The discussion in the preceding section illustrated design differences among interfaces in the good, average, and poor classes for each dataset. A visual comparison of the design examples (Figures 2–4) does not reveal subtle and code-level changes. To identify significant changes across years, we computed a one-way ANOVA based on model year (2000, 2002, or 2003). We then ranked the F ratios and report the main significant differences. We report trends that are independent of interface quality (i.e., good, average, or poor interfaces); however, we indicate differences for good interfaces.

6.1 Text Formatting

The smallest font size that was used for text increased slightly from 9pt (2000) to 10pt (2003). Larger font sizes improve legibility, especially for users who have vision impairments [Bernard et al. 2001; Flanders and Willis 1998; Schriver 1997]. In addition, there was a shift from using serif font faces to sans serif ones (mainly Arial). Studies suggest that sans serif fonts, in particular for small font sizes, may be more legible than serif fonts [Bernard and Mills 2000; Nielsen 2000; Schriver 1997].

6.2 Link Formatting

There was a shift from having textual links appear with lines underneath them to having then appear without lines underneath them. This practice may cause users to consider the links to be text as opposed to hyperlinks [Sawyer and Schroeder 2000]. Consequently, usability may impeded.

6.3 Graphics

The use of graphics doubled from 2000 (average of 25) to 2003 (average of 50); there was a slightly larger increase for good pages (from 26 to 62 graphics a page on average). Our measure does not distinguish between visible and invisible graphics (i.e., spacers), but over half of the graphics are redundant, which suggests that some graphics are used for lists or spacers. The number of distinct image files also increased from an average of 15 to 21. This increase reflects a shift from textual links to graphical links within navigation bars and the use of organization and ornamental graphics (bullets, form buttons, icons, etc.) [Scanlon and Schroeder 2000b]. For instance, the 2003 design examples in Figures 4 and 5 (bottom page) use textual and graphical links and other graphic elements.

6.4 Tables

There were several changes in page layouts, most of which are not visually apparent. The number of tables on pages doubled from 7 in 2000 to 14 in 2003, with most tables being used to control page layouts. Increased table use could impede accessibility and degrade performance [W3C 1999]. Figure 5 compares

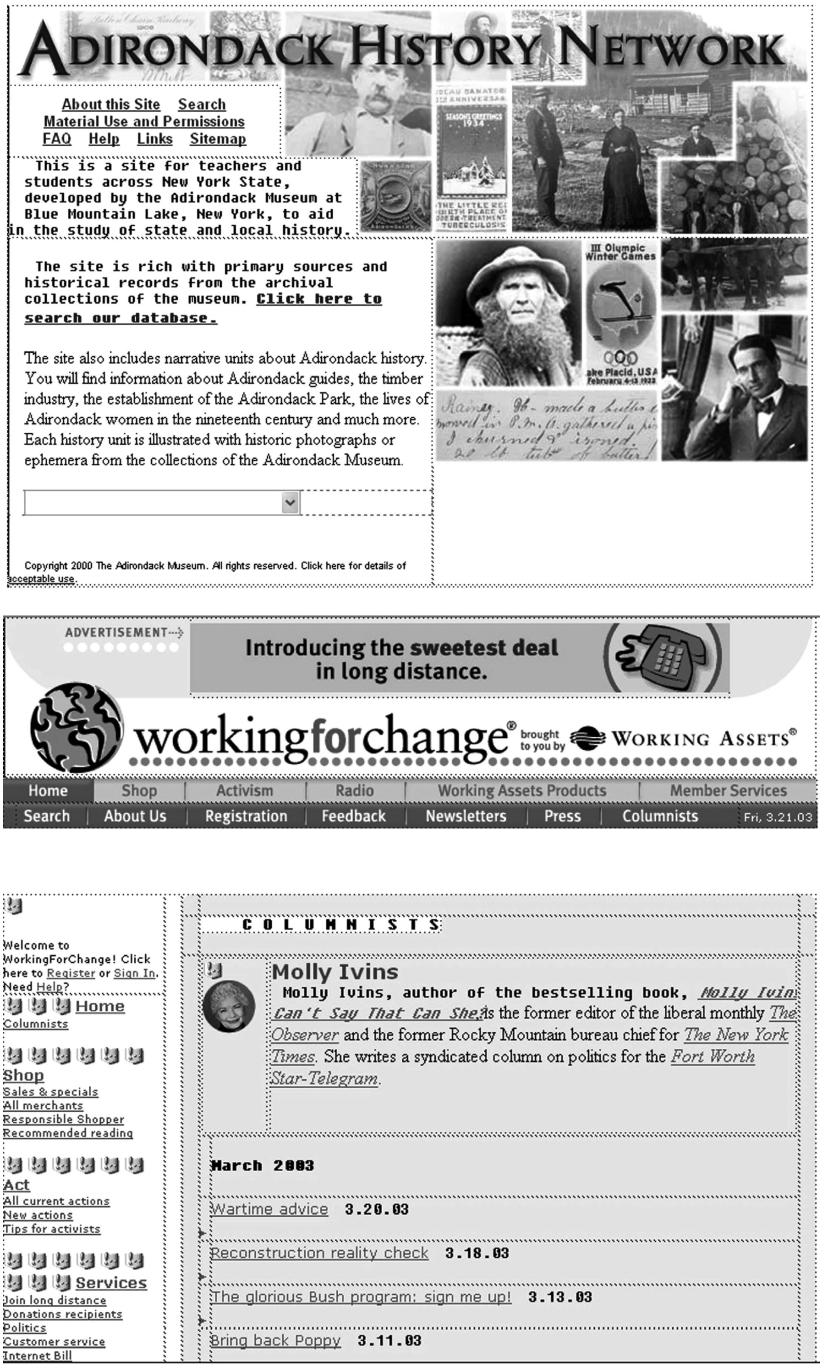

Fig. 5. Tables used on example pages from the archived 2003 dataset (good class). Dotted lines show table boundaries (rows, columns, and entire tables). The top page uses one table, while the bottom page (portions not depicted) uses 264 tables (the largest number of tables for pages within the good class). We captured pages from within Macromedia Dreamweaver.

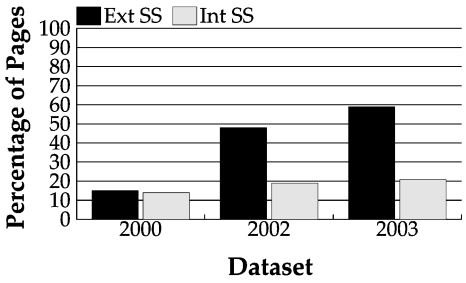

Fig. 6. Use of external (Ext SS) and internal (Int SS) stylesheets on web pages. Percentages are based on the total number of pages within each dataset (Table II).

two page layouts from the 2003 dataset. The bottom page uses 264 tables to control the layout, because each article in the middle column is formatted with a separate table. As demonstrated by the top page, the bottom page could have used one table to control the page layout, without sacrificing design aesthetics. For instance, the designer could have used bulleted lists with the same bullet graphic to organize the articles. Neither page provides summaries for each table, which is a practice that impedes accessibility.

6.5 Stylesheets

External cascading stylesheets (i.e., stylesheets that are stored in separate files and linked to from within the HTML page) improve page layout consistency [Nielsen 2000]. Nonetheless, they were not prevalent until the 2003 designs (Figure 6). There was a corresponding decrease in the degree to which pages could be rendered with just the base HTML file and linked image files versus external stylesheets, browser scripts, and other objects (referred to as self-containment). There was also a corresponding increase in the number of HTML and CSS bytes (nearly doubled from 14KB to 26KB). These changes can potentially degrade site performance, especially for graphics-heavy pages that are transferred over narrowband connections [Harwood and Rainie 2004; Madden and Rainie 2003].

6.6 Scripts

Browser scripts (i.e., code that is embedded or linked to from within HTML pages) became prevalent in 2003; scripts include JavaScript and applets, but not cascading stylesheets. We quantify the use of browser scripts by counting each script tag and the use of external script files; we also count the the total number of objects (scripts, applets, audio, video, etc.). The 2000 pages had no script files on average, but the 2002 and 2003 pages had one and two script files, respectively. There was a corresponding increase in the number of scripts, applets, and other objects embedded within pages (from 2 to 5 objects in 2000 and 2003, respectively); there was a slightly larger increase for good pages (from 2 to 7 objects on average).

Our Analysis Tool does not measure whether noscript alternatives are provided. We inspected ten randomly selected pages that contained scripts; the sample included pages from the 2003 dataset (poor, average, and good classes).

Our inspection revealed that none of the pages provided noscript alternatives. In some cases, users would not be able to navigate sites if script processing was disabled within their browsers.

Figure 7 depicts an example page that uses scripts extensively (29 scripts). Most scripts are embedded within the page's HTML code, but 16 unique script files are also used; the page has the highest number of script files and objects for the 2003 dataset. Most of the top and left navigation is lost when scripts are disabled. For instance, each item in the row of elements at the top of the page has a dynamic menu that appears when the user positions the mouse over it. As another example, the form, which enables users to sign up for a newsletter (to the right of the search functionality), is rendered with scripts. Many page elements are not downloaded by the crawler or analyzed, because our tools do not process scripts.

6.7 HTML Coding

As discussed in the preceding sections, there was a steady increase in the number of errors that are reported by Bobby (browser compatibility) and Weblint (coding). We used outdated software versions (Bobby 3.2 and Weblint 1.02) for consistency with the 2000 dataset. More recent tool versions will most likely yield additional errors due to new guidelines. Pages that do not comply with HTML coding guidelines may cause usability and accessibility problems.

6.8 Scent Quality

We measure the overlap between the text that is on source pages (i.e., the page that contains a hyperlink) and on destination pages (i.e., the page to which a hyperlink points). This overlap is often referred to as "scent" [Chi et al. 2000; Furnas 1997]. The amount of text overlap between source and destination pages increased from an average of 29 words in 2000 to 43 words in 2003; stop words were not considered. Increased scent between pages may improve usability.

6.9 Site Consistency

As depicted in Figures 3 and 4, page layouts became less consistent. In 2002, designs typically had changes in color schemes. In 2003, page structure was changed slightly. The literature discusses the consistency of page layouts across the site. Some sources advocate consistency [Flanders and Willis 1998; Fleming 1998; Mahajan and Shneiderman 1997], while others advocate some variation [Nielsen 2000; Sano 1996; Sawyer et al. 2000]. From a usability perspective, the types of changes that we observed within 2003 designs, changes in navigation in particular (see middle left page in Figure 4), could be problematic.

7. DESIGN REALITY VS. RECOMMENDATIONS

The discussion in the preceding sections provided descriptions of web site designs during the time period that we studied. We expand on the preceding discussion by presenting concrete thresholds (i.e., ranges of observed metric values) for ten aspects of web interfaces and by contrasting these thresholds to established guidelines from recognized experts and user studies. As discussed in Section 2.1, there is a large number of web design guidelines available in

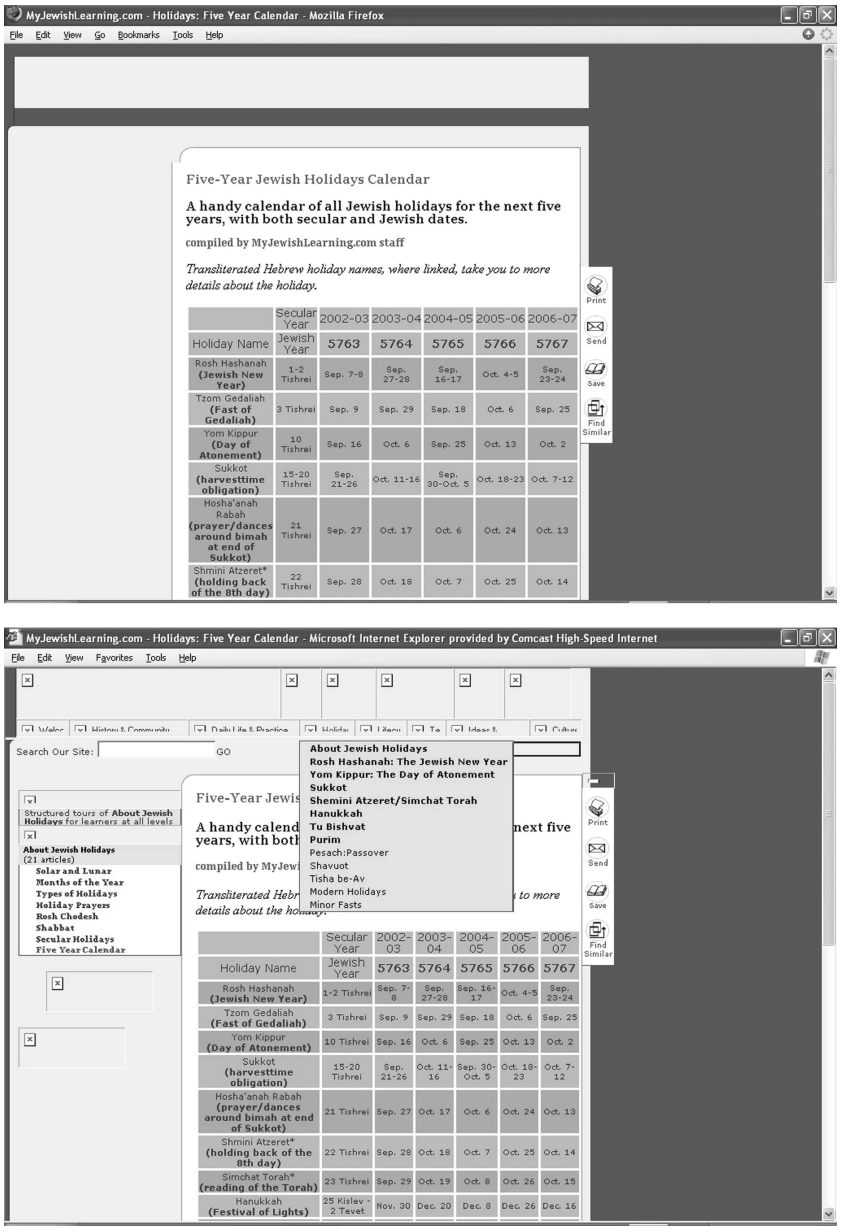

Fig. 7. Scripts used on an example page from the archived 2003 dataset (good class). The top screen shot shows the page rendered in the Mozilla Firefox browser with scripting disabled. The bottom screen shot shows the portion of the page downloaded by the crawler; missing elements were embedded within scripts, which the crawler did not process. The page uses a fixed screen width.

Table IV. Ranges for the Amount and Formatting of Text on Good Pages

	Dataset			
Measure	2000	2002	2003	Significance
Word Count	74–667	62–627	0–1,270	$F(2, 8381) = 77.58$, $p = .00$
Body Word Count	29–491	29–468	0–1,083	$F(2, 8381) = 73.04$, $p = .00$
Display Word Count	0–34	0–29	0–80	$F(2, 8381) = 34.87$, $p = .00$
Link Word Count	6–123	2–105	0–282	$F(2, 8381) = 69.40$, $p = .00$
Text Cluster Count	0–4	0–4	0–10	$F(2, 8381) = 50.60$, $p = .00$
Vertical Scrolls	1–3	1–3	1–8	$F(2, 8381) = 97.66$, $p = .00$

print and on the Web. Oftentimes, these recommendations are vague, contradictory, and not validated empirically. We can use thresholds, which we derive from highly rated interfaces, to augment design guidance that is provided to designers.

We used the three datasets to derive thresholds for measures that relate to ten design aspects, including: the amount of text, numbers of links and graphics, and consistency. More specifically, we report ranges for each measure such that the ranges represent two standard deviation units (i.e., mean plus and minus the standard deviation) around the mean values for interfaces that are within the good class. Thresholds for individual measures are not intended to be used in isolation. To illustrate these relationships, for each design aspect, we report groups of measures that have significant correlations within all three datasets. In some cases, thresholds support current design guidelines; in other cases, they counter them.

7.1 Amount of Page Text

The literature contains contradictory heuristics on the ideal amount of page text. Sources suggest that users prefer all relevant content to appear on one page [Landesman and Schroeder 2000], and others suggest that content should be broken up into smaller units and distributed across multiple pages [Flanders and Willis 1998; Nielsen 2000]. Another recommendation is to use an amount of text that is relevant to the page's function (e.g., make home and link pages shorter) [Koyani et al. 2003b]. Irrespective of the recommendation, there is no concrete guidance on how much text is enough or too much.

We depict several relevant measures in Table IV; ranges for the three datasets were all significantly different based on a one-way ANOVA. Ranges in Table IV appear to support the suggestion to place all text on one page, even for home pages (see discussion below). Although the measures provide some guidance on the amount of text, they do not provide insight on whether text is broken up into multiple pages. Ranges suggest that web pages are becoming increasingly text-heavy (word count); the bulk of the word count was attributable to body text (i.e., text that is not used in headings or links; body word count). There was a proportional amount of text for headings and textual links (display and link word counts). Text formatting also appeared to be proportional to the amount of text. Table IV shows that text clustering (i.e., areas of text that are highlighted with color, rules, lists, etc.; text cluster count) increased over the years. The number of vertical scrolls increased and was proportional to the amount of text.

Table V. Ranges for the Amount of Link Text on Good Pages

	Dataset			
Measure	2000	2002	2003	Significance
Average Link Words	2–3	2–4	1–4	$F(2, 8381) = 17.29$, $p = .00$
Average Good Link Words	1–3	1–3	1–3	$F(2, 8381) = 16.61$, $p = .00$

We contrasted the ranges in Table IV to ranges derived for just the home pages; we only included pages that were the first pages crawled on sites and were predicted to be home pages by our page type model (see Section 3.2.2). Ranges for home pages were very similar to Table IV; the only difference for the most recent dataset was that home pages used fewer text clusters (0–7).

Similarly to the use of graphics and other elements, the increase in the amount of text is most likely attributable to the increase in Internet connection speeds (e.g., broadband). Despite the availability of higher speed connections, the majority of users accessed the Web via narrowband connections during the study period [Harwood and Rainie 2004]. Hence, it is possible that the increased amount of text, graphics, and so on may have decreased web site usability for these users.

7.2 Length and Quality of Link Text

Design guidelines are contradictory with respect to the ideal amount of text to include within textual links. One suggestion is to use 2–4 words [Nielsen 2000]. Another suggestion is to use 7–12 "useful" words (i.e., words that provide hints about the content on a destination page) [Sawyer and Schroeder 2000]. The average link words measure (number of words used in textual links / number of textual links) shows that textual links contained from two to four words (Table V). Furthermore, the average good link words measure shows that one to three of these words were not stop words (i.e., common words) or the word 'click.' This finding suggests that the link text on good pages may have been useful.

Ranges suggest that the link text that is on good pages is consistent with Nielsen's heuristic. There is one caveat to this finding: Our metrics tool does not distinguish between links that are within the body text or outside of the body text (e.g., in a navigation bar). There could be differences in the number of words used for these two types of links.

7.3 Number and Types of Links

There is an ongoing debate about the appropriate number and types of links to use on web pages. Sources suggest that: (1) the number of links should be minimized, (2) certain types of links (e.g., graphical, repeated, or within-page links) should be avoided, and (3) multiple links should be provided to the same content and in different forms (e.g., text, text embedded within a graphic, or graphic) [Flanders and Willis 1998; Sano 1996; Nielsen 2000; Sawyer and Schroeder 2000; Scanlon and Schroeder 2000b; Spool et al. 1999, 2000]. Table VI provides ranges for several link element measures: number of links (link count),

Table VI. Ranges for the Number, Types, and Formatting of Links on Good Pages

	Dataset			
Measure	2000	2002	2003	Significance
Link Count	12–69	10–57	3–107	$F(2, 8381) = 167.72$, $p = .00$
Text Link Count	4–50	2–38	0–87	$F(2, 8381) = 117.91$, $p = .00$
Link Graphic Count	3–21	2–21	3–30	$F(2, 8381) = 177.19$, $p = .00$
Internal Link Count	10–60	8–50	0–97	$F(2, 8381) = 146.32$, $p = .00$
Page Link Count	0–5	0–1	0–8	$F(2, 8381) = 30.69$, $p = .00$
Redundant Link Count	0–15	0–12	0–25	$F(2, 8381) = 96.38$, $p = .00$
Link Text Cluster Count	0–3	0–2	0–7	$F(2, 8381) = 136.18$, $p = .00$

numbers of textual (text link count) and graphical (link graphic count) links, numbers of links that point to pages that are within the same page (page link count) or site (internal link count), and number of repeated links on a page (redundant link count).

The table shows that a large number of links were used on pages, especially within the 2003 dataset. One way to compensate for a large number of links is to organize them into clusters (i.e., regions on web pages in which links are grouped together with colors, lists, or borders). The table shows that the use of link text clusters was proportional to the number of links.

The table shows that redundant or repeated links are used on good pages and graphical links are not avoided as suggested in the literature. It is possible that the graphical links have text within them, but the text is not currently detected by our metrics tool. Given the ranges for redundant links along with the ranges for textual and graphical links, it seems that some links were repeated in both text and graphic formats. The measures do not currently reveal how often redundant links correspond to links that are also graphical or textual links.

Table VI shows that good pages contained few within-page links, as suggested by the literature. Most links pointed to pages that were within the site, as opposed to pages that were external to the site. This trend seems somewhat inconsistent with page ranking algorithms that are used by search engines like Google [Brin and Page 1998]. These algorithms rely on cross-linking to determine a page's popularity and use these rankings to order search results.

7.4 Number and Types of Graphics

Images are a key element of a page's graphic design, and the literature extensively discusses their use [Ambuhler and Lindenmeyer 1999; Flanders and Willis 1998; Sano 1996; Schriver 1997; Spool et al. 1999; Stein 1997]. The consensus seems to be that the number of images should be minimized to improve download speed. Even with the shift to broadband connections [Madden and Rainie 2003], the majority of users still use narrowband connections, especially at home [Harwood and Rainie 2004]. Sources also suggest that certain types of images should be avoided [Flanders and Willis 1998; Nielsen 2000; Scanlon and Schroeder 2000b]: (1) images that contain text (content graphics), (2) images that are used for navigation (navigation graphics), and (3) images that are animated.

Table VII. Ranges for the Number, Types, and Formatting of Graphics on Good Pages

	Dataset			
Measure	2000	2002	2003	Significance
Graphic Count	5–45	7–64	6–115	$F(2, 8381) = 487.40,\ p = .00$
Link Graphic Count	3–21	2–21	3–30	$F(2, 8381) = 177.19,\ p = .00$
Graphic Link Count	1–17	2–19	2–25	$F(2, 8381) = 150.38,\ p = .00$
Animated Graphic Count	0–2	0–1	0–2	$F(2, 8381) = 37.39,\ \ p = .00$
Redundant Graphic Count	0–22	0–35	0–84	$F(2, 8381) = 453.40,\ p = .00$
Average Graphic Height	11–59px	11–57px	0–63px	$F(2, 8381) = 305.00,\ p = .00$
Average Graphic Width	45–184px	49–178px	24–186px	$F(2, 8381) = 13.96,\ \ p = .00$

Table VIII. Ranges for the Number and Type of Graphical Ads on Good Pages

	Dataset			
Measure	2000	2002	2003	Significance
Graphic Ad Count	0–3	0–2	0–5	$F(2, 8381) = 111.95,\ p = .00$
Animated Graphic Ad Count	0–1	0–1	0–1	$F(2, 8381) = 98.32,\ \ p = .00$

Table VII suggests that the use of graphics was not minimized on pages. There was a sharp increase in the number of graphics used on the 2003 pages; however, there was a corresponding increase in the number of repeated or redundant graphics on pages. One limitation of our graphic count measure is that it does not distinguish between visible and invisible (i.e., spacer) graphics. The redundant graphic count suggests the use of invisible graphics and possibly the use of organizer graphics like rules, bullets, and so on. There were negative correlations between graphic size (average width and height) and graphic and redundant graphic counts; correlations also suggest the use of spacer and organizer graphics, due to their smaller sizes. Based on the average graphic width and height, it seems that graphics tended to be rectangular, with longer widths than heights.

As discussed in the preceding section and shown in Table VII, navigation graphics were not avoided. The graphic link (number of graphics that are also links) and link graphic (number of links that are also part of graphics) counts are similar with respect to quantifying the use of navigation graphics. Link graphic count will be higher than graphic link count when image maps are used on pages, because every clickable area of an image map is counted as a separate link graphic. Range differences for these two measures suggest that pages used image maps; this pattern is contrary to design guidance. The table shows that animated graphics were minimized, as suggested by the literature.

7.5 Use of Nonanimated and Animated Graphical Ads

Sources suggest that graphical ads (animated or not) should be minimized or avoided [Flanders and Willis 1998; Klee and Schroeder 2000; Nielsen 2000]. On the contrary, there is at least one source that suggests that ads increase credibility [Kim and Fogg 1999]. Table VIII shows that good pages were likely to contain no more than five graphical ads. Use of graphical ads more than doubled on pages within the 2003 dataset, which suggests that design practices are deviating from recommendations. This deviation is interesting given the

Table IX. Ranges for the Number, Type, and Size of Fonts on Good Pages

	Dataset			
Measure	2000	2002	2003	Significance
Font Style	Sans Serif	Sans Serif	Sans Serif	$F(2, 8381) = 46.66, \quad p = .00$
Font Count	4–8	2–7	3–9	$F(2, 8381) = 100.75, p = .00$
Serif Font Count	0–1	0–1	0–1	$F(2, 8381) = 93.73, \quad p = .00$
Sans Serif Font Count	0–1	0–1	1–2	$F(2, 8381) = 596.29, p = .00$
Minimum Font Size	8–8pt	9–11pt	9–11pt	$F(2, 8381) = 406.42, p = .00$
Average Font Size	10–12pt	10–12pt	10–13pt	$F(2, 8381) = 79.49, \quad p = .00$

ongoing debate as to whether or not users ignore graphical ads (e.g., banner ads) [Curry 2002; Dahlen 2001; Donatelli 2003; The Poynter Institute 2004].

Good pages in all datasets tended to have zero or one animated graphical ad, which suggests that animation was used sparingly. There was a large positive (and significant) correlation between animated graphics (discussed in the preceding section) and animated graphical ads, as well as between graphical ads and animated graphical ads. These findings suggest that: (1) when animated graphics were used on pages, they tended to be graphical ads; and (2) when graphical ads were used on pages, they tended to be animated.

7.6 Use of Font Styles and Sizes

A font is a combination of four features: a font face (typeface), a font size, whether text is bolded, and whether text is italicized [Schriver 1997]. The literature suggests the use of serif typefaces, sans serif typefaces, or a combination of both (e.g., serif typeface for larger text) [Bernard and Mills 2000; Bernard et al. 2001; Nielsen 2000; Schriver 1997]. Guidance varies with respect to minimum font sizes to use: Recommendations range from 9pt to 14pt, with larger sizes suggested for older adults. Finally, sources recommend the use of only a few font sizes from one or two typeface families.

Our analysis revealed that sans serif is the predominant font style used on good pages (Table IX); studies suggest that these typefaces are more legible online than serif typefaces. Although the pages used variations in fonts (e.g., using bold text or different font sizes; font count), they did not use a large number of font faces. Pages used no more than one serif font face (serif font count) and no more then two sans serif font faces (sans serif font count). Our measures do not assess whether larger font sizes are used for serif typefaces and vice versa for sans serif typefaces; however, correlations between the numbers of font faces and font sizes suggest that design practices were consistent with the recommendations. There was a slight increase in the minimum and average font size in the most recent datasets; these increases are consistent with design guidance. One caveat about the ranges is that they do not reflect font faces and styles that are used within graphics.

7.7 Use of Unique Colors and Color Combinations

The literature recommends using a small number of colors, browser-safe colors, default link colors, color combinations with adequate contrast, and so on [Flanders and Willis 1998; Kaiser 1998; Murch 1985; Nielsen 2000; Spool et al.

Table X. Ranges for the Number and Types of Colors on Good Pages

	Dataset			
Measure	2000	2002	2003	Significance
Body Color Count	1–3	1–3	1–4	$F(2, 8381) = 122.86$, $p = .00$
Display Color Count	1–2	0–2	0–3	$F(2, 8381) = 97.92$, $p = .00$
Link Color Count	2–4	2–5	2–5	$F(2, 8381) = 110.88$, $p = .00$
Standard Link Color Count	0–2	0–3	0–3	$F(2, 8381) = 22.45$, $p = .00$
Color Count	5–11	5–11	5–14	$F(2, 8381) = 181.60$, $p = .00$
Browser-Safe Color Count	4–7	3–7	4–8	$F(2, 8381) = 58.15$, $p = .00$
Good Text Color Combination	1–6	1–6	1–7	$F(2, 8381) = 122.85$, $p = .00$
Good Panel Color Combinations	0–1	0–1	0–3	$F(2, 8381) = 354.91$, $p = .00$

1999]. We developed over twelve measures to assess color use within the HTML code and stylesheets. One limitation of the measures is that colors that are used within images, browser scripts, and other objects are not considered by our tools. Nonetheless, the current measures provide some valid insight on color use. We discuss a subset of the measures in this section (Table X).

Our analysis revealed that good pages tended to use from one to four colors for body text and up to three colors for headings; the number of heading colors was proportional to the amount of text or formatting on pages. Although it is not clear whether different colors were used for body and heading text on these pages, unique color counts (color count) suggest that this may be the case. The 2003 dataset reflects a small increase in the number of colors used overall; this trend is not consistent with guidance in the literature. We found that good pages tended to use high-contrast text color combinations. Good pages used high-contrast panel color combinations to a lesser degree than text color combinations.

Good pages used two to five colors for links (link color count); default browser colors (i.e., red, blue, and purple; standard link color count) were not always used. Ranges for link colors suggest that good pages did not closely follow the guidance in the literature. Ranges for the total number of unique colors show that good pages did not adhere to the guidance of using no more than six discriminable colors [Murch 1985]. Furthermore, they did not strictly use browser-safe colors (browser-safe color count), as recommended by the literature.

7.8 Use of Non-HTML Technology

The literature discourages the use of applets, controls, scripts, video, sound, plug-ins, and so on. [Ambuhler and Lindenmeyer 1999; Flanders and Willis 1998; Nielsen 2000; Rosenfeld and Morville 1998; Spool et al. 1999; Stein 1997]. Table XI shows ranges for several measures related to the use of browser scripts (i.e., JavaScript or applet code that is embedded or linked to from within HTML pages) and other objects. The object count includes all script, video, and related tags that appear within a web page; however, the object bytes and object file count measures exclude scripts. Contrary to recommendations, there was a steady increase in the use of non-HTML technology, especially the use of browser scripts. In severe cases, pages did not render properly or navigation elements were lost when scripting was disabled within browsers (Figure 7). These practices can impede usability and accessibility.

Table XI. Ranges for Scripts and other Objects on Good Pages

	Dataset			
Measure	2000	2002	2003	Significance
Object Count	0–4	0–6	0–13	$F(2, 8381) = 592.72$, $p = .00$
Object Bytes	0–6B	0–143KB	0–138KB	$F(2, 8381) = 32.61$, $p = .00$
Object File Count	0–0	0–2	0–1	$F(2, 8381) = 11.02$, $p = .00$
Script Bytes	0–3KB	0–7KB	0–31KB	$F(2, 8381) = 656.95$, $p = .00$
Script File Count	0–2	0–1	0–4	$F(2, 8381) = 438.31$, $p = .00$

Table XII. Ranges for HTML Coding Errors on Good Pages

	Dataset			
Measure	2000	2002	2003	Significance
Bobby Approved	No	No	No	$F(2, 8381) = 53.71$, $p = .00$
Bobby Priority 1 Errors	0–2	0–2	1–2	$F(2, 8381) = 52.07$, $p = .00$
Bobby Priority 2 Errors	3–5	2–5	3–5	$F(2, 8381) = 75.26$, $p = .00$
Bobby Priority 3 Errors	2–2	2–2	2–2	$F(2, 8381) = 44.01$, $p = .00$
Bobby Browser Errors	6–21	9–25	11–32	$F(2, 8381) = 473.27$, $p = .00$
Weblint Errors	0–70	0–70	0–184	$F(2, 8381) = 232.53$, $p = .00$

7.9 Validity of HTML Coding

Our analysis revealed that Bobby and Weblint errors were more prevalent on good pages than on average and poor pages, despite the guidance that web designers adhere to accessibility principles [Clark and Dardailler 1999; Cooper 1999; Nielsen 2000; W3C 1999] and avoid making HTML errors [Bowers 1996; Kim and Fogg 1999; Fogg et al. 2000]. Table XII shows that good pages were typically not Bobby approved.

Similarly to the accessibility studies that we discussed in Section 2.2, we analyzed the Bobby approval percentages for pages. We conducted a univariate ANOVA to examine the effect of the dataset's year and the interface's class (good, average, or poor) on Bobby approval. We found that both the dataset's year and interface's class had an effect; there was also an interaction effect. Overall, approval rates were lowest for the 2003 dataset, possibly due to the larger number of pages; approval rates for the good, average, and poor classes were 12, 16, and 17 percent, respectively. Approval rates were highest for the 2002 dataset, with rates for the good, average, and poor classes being 21, 16, and 24 percent, respectively. These approval rates are in stark contrast to prior studies, possibly due to the much smaller sample sizes and specific genres that these studies used.

Table XII shows that pages contained several accessibility errors (Priority 1, 2, and 3) and a large number of browser-compatibility and Weblint errors. Both the accessibility and Weblint errors were proportional to the amount of formatting on pages. More specifically, the errors correlated with the use of tables, interactive elements (i.e., form elements), and browser scripts. Common coding problems included: improper labeling of form elements, missing table summaries, missing row and column headings in data tables, and missing no-script tags for browser scripts. Even though HTML coding guidance has been available for quite some time, HTML coding issues steadily increased during the time period.

Table XIII. Ranges for Consistency Measures on Good Sites. All Measures are Percentages

	Dataset			
Measure	2000	2002	2003	Significance
Page Formatting Variation	0–21	5–55	7–63	$F(2, 546) = 45.62,\ p = .00$
Page Title Variation	17–180	19–139	9–218	$F(2, 546) = 8.20,\ \ p = .00$

7.10 Consistency Across Pages

Some sources advocate consistent use of design elements across web pages [Flanders and Willis 1998; Fleming 1998; Mahajan and Shneiderman 1997], while others claim that such elements become invisible [Sano 1996; Sawyer et al. 2000]. The literature also suggests that page titles (i.e., title tag) should be varied [Nielsen 2000]; in other words, each page should have a different title. We derived twelve measures to reflect the percentage of variation for groups of related page-level measures [Ivory 2001]; a large variation percentage reflects less consistency and vice versa. In this section, we discuss two measures that experienced the most change during the time period.

Our analysis showed an increase in changes to page layouts (page formatting variation) during the time period (Table XIII). It appears that the trend is to introduce more variation into page layouts (e.g., Figure 4), as opposed to keeping them consistent. This trend is somewhat disturbing, because some studies show that performance improves with consistent interfaces.

Page title variation suggests that titles vary considerably on good sites. This trend is consistent with design guidance. This practice can help users to keep track of their current location within a site's information architecture.

8. LIMITATIONS AND FUTURE RESEARCH

We summarize limitations and future work with respect to the datasets, WebTango methodology, and design patterns in this section.

8.1 Datasets

To our knowledge, the Webby resource is the largest sample of sites that are evaluated on consistent interface criteria. Hence, all three of our datasets were derived from sites submitted for evaluation during the Webby Awards' review stage. Based on judges' ratings of these sites, we segmented pages and sites into good, average, and poor classes. We assumed that a site's rating applied to all pages within it.

Our findings with respect to design patterns (Sections 5–7) suggest that the Webby resource may not be an ideal rating source and that our assumption about site rating applicability may not be valid. For instance, we showed instances of good pages in which important page elements and navigation were lost when scripts were disabled. We also showed how the structure of pages varied across pages within the same site. To develop models of effective design practices, we need designs that exhibit those practices. Consequently, more rigor needs to go into the review of Webby sites. In particular, more attention needs to be paid to accessibility issues.

We suggest that the Webby committee, or any other organization that rates sites, adopt a rigorous evaluation procedure. The Webby committee follows a systematic process with respect to how sites advance through judging stages; however, the judges are given very little guidance on how to conduct evaluations. Essentially, they are given descriptions of heuristics, but no guidance on how to apply them.

We developed one such heuristic evaluation approach to guide the assessment of web sites designed for courses [Ivory 2004]. This approach consists of 12 high-level criteria and 46 sub-criteria. Each criterion has a protocol associated with it to guide evaluators in assessing the site's conformance to the criterion and in assigning a rating for the criterion. This methodology could be adapted to facilitate the evaluation of generic web sites.

8.2 WebTango Methodology

Even though we demonstrated that it is possible to find correlations between values for measures and expert ratings, we make no claim about the profiles representing causal links. It is possible that the highly rated sites are highly rated for reasons other than what is assessed with the measures, such as the quality of the content on the site. The current models and tools cannot improve on poor content. However, our empirical studies provided preliminary evidence that they can provide insight on how to take good content that is poorly presented and improve its presentation, thus improving users' experiences in accessing that content [Ivory 2001; Ivory and Hearst 2002a, 2002b]. And, because it is possible to empirically find commonalities among the presentation elements of the highly rated sites, this provides strong evidence that the presentational aspects of highly rated sites that differ from those of poorly rated sites are in fact important for good design.

Another major limitation of our approach is that the quantitative measures do not capture users' subjective preferences. For example, one study has shown that perceived download speed is more important than actual download speed [Scanlon and Schroeder 2000a]. Although we can measure actual download speed, it may not be possible to assess perceived speed. Nonetheless, the methodology can be viewed as a reverse engineering of design decisions that were presumably informed by user input.

When the project first began, the Web was very different from what it is currently. Browser scripting, animation, and multimedia content were far less prevalent on the Web and rarely integral to the core features and content of sites. Today, these elements are no longer novelties. They are essential to many sites, and, consequently, they should be considered during assessments of usability and accessibility. Model-building results suggest that we need to improve current measures and our approach to building models. We also need to develop new measures and techniques (e.g., image processing algorithms) to address the current state of the Web. We have begun to explore the use of image and document processing techniques to improve measure accuracy [Harrison and Shin 2003].

8.3 Design Patterns

The thresholds, which we presented in Section 7, complement the models that we discussed briefly. For instance, our Analysis Tool augments model predictions with the derived ranges to show designers how their designs are similar to or different from good interfaces. In this article, we only presented a subset of the thresholds for good interfaces, irrespective of content categories and page types. We have derived a complete set of thresholds for good interfaces in general and similar context-specific thresholds; all thresholds are embedded within the Analysis Tool. The thresholds and patterns are meant to be descriptive, rather than prescriptive. Ideally, through iterative cycles of design, evaluation, and interpretation, the designer can become more knowledgeable about design patterns that may not be apparent.

Similarly to other automated evaluation approaches, the models cannot evaluate design patterns for which there are no heuristics. Currently, we do not display ranges for all measures. We will explore doing so in the future, as a way to mitigate model limitations. We also plan to investigate ways in which to distinguish design patterns that designers should apply versus ignore (e.g., ineffective coding of tables), even if the patterns are prevalent on highly rated interfaces.

9. CONCLUSIONS

We demonstrated through three separate analyses that web site design is a moving target. We described aspects of web site design during three time periods (2000, 2002, and 2003). We then discussed significant design changes for the time periods. We found that web site designs became increasingly graphical in nature, reliant on browser scripts, and less consistent. We discussed how these practices and others mesh with established design guidance. In many cases, such as the numbers and types of links, graphical ads, color usage, and accessibility, design practices contradict heuristics from the literature. Nonetheless, in other cases, such as the length of link text, font styles, and page titles, design practices support heuristics from the literature. The most glaring deficiency of web sites, even for highly rated interfaces, is their inadequate accessibility. Current practices with respect to tables, forms, and browser scripts could impede accessibility and usability. Some of these design practices need to undergo initial or further empirical studies.

ACKNOWLEDGMENTS

We thank everyone who has contributed to the WebTango Project, including: Marti Hearst, Rashmi Sinha, Alissa Harrison, Young-Mi Shin, Shiquing Yu, Mary Deaton, Nicole Elger, Tina Marie, Wenchun Wang, Deep Debroy, Toni Wadjiji, Chantrelle Nielsen, Wai-ling Ho-Ching, Stephen Demby, David Lai, Judy Ramey, Jennifer Turns, Beth Kolko, David Farkas, and Aline Chevalier. We thank Maya Draisin and Tiffany Shlain at the International Academy of Digital Arts and Sciences for making the Webby Awards data available. We thank Lincoln Stein for his assistance with an early version of the Metrics Computation Tool and Tom Phelps for his assistance with the extended Metrics Computation Tool.

REFERENCES

AMBUHLER, R. AND LINDENMEYER, J. 1999. Measuring accessibility. In *Proceedings of the Eighth International World Wide Web Conference*. Foretec Seminars, Inc., Toronto, Canada, 48–49. http://www.weboffice.ethz.ch/www8.

BARRY, C. AND LANG, M. 2001. A survey of multimedia and web development techniques and methodology usage. *IEEE MultiMedia 8,* 3, 52–60.

BERNARD, M., LIAO, C. H., AND MILLS, M. 2001. The effects of font type and size on the legibility and reading time of online text by older adults. In *Proceedings of the Conference on Human Factors in Computing Systems*. Vol. 2. Seattle, WA, 175–176.

BERNARD, M. AND MILLS, M. 2000. So, what size and type of font should I use on my website? *Usability News Summer*. http://wsupsy.psy.twsu.edu/surl/usabilitynews/2S/font.htm.

BORGES, J. A., MORALES, I., AND RODRIGUEZ, N. J. 1996. Guidelines for designing usable World Wide Web pages. In *Proceedings of the Conference on Human Factors in Computing Systems*. Vol. 2. ACM Press, Vancouver, Canada, 277–278.

BOWERS, N. 1996. Weblint: quality assurance for the World Wide Web. In *Proceedings of the Fifth International World Wide Web Conference*. Elsevier Science Publishers, Paris, France. http://www5conf.inria.fr/fich_html/papers/P34/Overview.html.

BREWINGTON, B. E. AND CYBENKO, G. 2000. How dynamic is the web? In *Proceedings of the 9th international World Wide Web conference on Computer networks: the international journal of computer and telecommunications netowrking*. North-Holland Publishing Co., Amsterdam, The Netherlands, 257–276.

BRIN, S. AND PAGE, L. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems 30*, 1–7, 107–117.

BRINCK, T., GERGLE, D., AND WOOD, S. D. 2001. *Usability for the Web: Designing Web Sites That Work*. Morgan Kaufmann Publishers, San Francisco.

CHEVALIER, A. AND IVORY, M. Y. 2003a. Can novice designers apply usability criteria and recommendations to make web sites easier to use? In *Proceedings of the 10th International Conference on Human-Computer Interaction*. Theory and Practice (Part I), vol. 1. Crete, Greece, 58–62. http://ubit.ischool.washington.edu/pubs/hcii03/hcii03a.pdf.

CHEVALIER, A. AND IVORY, M. Y. 2003b. Web site designs: Influences of designer's experience and design constraints. *Int. J. Hum. Comput. Studies 58*, 1, 57–87. http://ubit.ischool.washington.edu/pubs/ijhcs03/designers.pdf.

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

CHO, J. AND GARCIA-MOLINA, H. 2000. The evolution of the web and implications for an incremental crawler. In *Proceedings of the 26th International Conference on Very Large Data Bases*. Morgan Kaufmann Publishers Inc., 200–209.

CHO, J. AND GARCIA-MOLINA, H. 2003. Estimating frequency of change. *ACM Trans. Inter. Tech. 3*, 3, 256–290.

CLARK, D. AND DARDAILLER, D. 1999. Accessibility on the web: Evaluation & repair tools to make it possible. In *Proceedings of the CSUN Technology and Persons with Disabilities Conference*. Los Angeles, CA. http://www.dinf.org/csun_99/session0030.html.

COMBER, T. 1995. Building usable web pages: An HCI perspective. In *Proceedings of the First Australian World Wide Web Conference*, R. Debreceny and A. Ellis, Eds. Norsearch, Ballina, Australia, 119–124. http://www.scu.edu.au/sponsored/ausweb/ausweb95/papers/hypertext/comber/.

COMPUTER SCIENCE AND TELECOMMUNICATIONS BOARD. 1997. *More Than Screen Deep: Toward Every-Citizen Interfaces to the Nations Information Infrastructure*. National Academy Press, Washington, DC.

COOPER, M. 1999. Universal design of a web site. In *Proceedings of the CSUN Technology and Persons with Disabilities Conference*. Los Angeles, CA. http://www.dinf.org/csun_99/session0030.html.

CURRY, S. 2002. Marketing—e-marketing evolution—some say banner ads are a waste of time and money, but contrary to popular belief, online marketing is finally growing up take a look at

how three companies have used electronic marketing campaigns to raise revenues—and profits. *Sales & Marketing Management 154*, 6, 32.

DAHLEN, M. 2001. Banner ads through a new lens. *Journal of Advertising Research 41*, 4, 8.

DONATELLI, B. 2003. Bandwagon—online adviser: Hunting where the ducks are: Online banner ads in 2002. *Campaigns & Elections 24*, 4, 3.

EASTON, V. J. AND MCCOLL, J. H. 1997. Statistics glossary v1.1. http://www.stats.gla.ac.uk/steps/glossary/index.html.

FETTERLY, D., MANASSE, M., NAJORK, M., AND WIENER, J. 2003. A large-scale study of the evolution of web pages. In *Proceedings of the 12th International Conference on World Wide Web*. ACM Press, Budapest, Hungary, 669–678.

FETTERLY, D., MANASSE, M., NAJORK, M., AND WIENER, J. L. 2004. A large-scale study of the evolution of web pages. *Softw. Pract. Exper. 34*, 2, 213–237.

FLANDERS, V. AND WILLIS, M. 1998. *Web Pages That Suck: Learn Good Design by Looking at Bad Design*. SYBEX, San Francisco, CA.

FLEMING, J. 1998. *Web Navigation: Designing the User Experience*. O'Reilly & Associates, Sebastopol, CA.

FOGG, B., MARSHALL, J., OSIPOVICH, A., VARMA, C., LARAKI, O., FANG, N., PAUL, J., RANGNEKAR, A., SHON, J., SWANI, P., AND TREINEN, M. 2000. Elements that affect web credibility: Early results from a self-report study. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM Press, The Hague, The Netherlands, 287–288. http://www.webcredibility.org/WebCredEarlyResults.ppt.

FOGG, B. J., MARSHALL, J., LARAKI, O., OSIPOVICH, A., VARMA, C., FANG, N., PAUL, J., RANGNEKAR, A., SHON, J., SWANI, P., AND TREINEN, M. 2001. What makes web sites credible?: a report on a large quantitative study. In *Proceedings of the ACM CHI'01 Conference on Human Factors in Computing Systems*. Vol. 1. Seattle, WA, 61–68.

FORRESTER RESEARCH. 1999. Why most web sites fail. http://www.forrester.com/Research/ReportExcerpt/0,1082,1285,00.html.

FURNAS, G. W. 1997. Effective view navigation. In *Proceedings of Conference on Human Factors in Computing Systems*. Vol. 1. ACM Press, Atlanta, GA, 367–374.

HARRISON, A. AND SHIN, Y.-M. 2003. Using document image analysis to evaluate web page design. In *Proceedings of the Sixth Annual UW Undergraduate Research Symposium*. Poster Session. University of Washington. http://webtango.ischool.washington.edu/pages/showAbstract.php?absid=urp03\&abstype=Publication.

HARWOOD, P. AND RAINIE, L. 2004. Use of the Internet in places other than home or work: A Pew Internet project data memo. http://www.pewinternet.org/pdfs/PIP_Other_Places.pdf.

IVORY, M. Y. 2001. An empirical foundation for automated web interface evaluation. Ph.D. thesis, University of California, Berkeley. Ph.D. thesis.

IVORY, M. Y. 2003a. *Automated Web Site Evaluation: Researchers' and Practitioners' Perspectives*. Human-Computer Interaction Series, vol. 4. Kluwer Academic Publishers, Dordrecht, The Netherlands.

IVORY, M. Y. 2003b. Characteristics of web site designs: Reality vs. recommendations. In *Proceedings of the 10th International Conference on Human-Computer Interaction*. Theory and Practice (Part I), vol. 1. Crete, Greece, 773–777.

IVORY, M. Y. 2004. SmartSites toolkit for evaluating course web sites. Tech. Rep. IS-TR-2004-12-02, Information School, University of Washington. http://hdl.handle.net/1773/2030.

IVORY, M. Y. AND CHEVALIER, A. 2002. A study of automated web site evaluation tools. Tech. Rep. 02-10-01, University of Washington, Department of Computer Science and Engineering. http://ubit.ischool.washington.edu/pubs/tr02/toolstudy.pdf.

IVORY, M. Y. AND HEARST, M. A. 2002a. Improving web site design. *IEEE Internet Computing 6*, 2, 56–63. http://ubit.ischool.washington.edu/pubs/ieee02/ieee-final.pdf.

IVORY, M. Y. AND HEARST, M. A. 2002b. Statistical profiles of highly-rated web site interfaces. In *Proceedings of the Conference on Human Factors in Computing Systems*. CHI Letters, vol. 4. Minneapolis, MN, 367–374. http://ubit.ischool.washington.edu/pubs/chi02/chi2002.pdf.

IVORY, M. Y., MANKOFF, J., AND LE, A. 2003. Using automated tools to improve web site usage by users with diverse abilities. *IT&Society 1*, 3, 195–236. http://www.stanford.edu/group/siqss/itandsociety/v01i03/v01i03a11.pdf.

IVORY, M. Y., SINHA, R. R., AND HEARST, M. A. 2000. Preliminary findings on quantitative measures for distinguishing highly rated information-centric web pages. In *Proceedings of the 6th Conference on Human Factors & the Web*. Austin, TX. http://ubit.ischool.washington.edu/pubs/hfw00/hfw00.pdf.

IVORY, M. Y., SINHA, R. R., AND HEARST, M. A. 2001. Empirically validated web page design metrics. In *Proceedings of the Conference on Human Factors in Computing Systems*. Vol. 1. Seattle, WA, 53–60. http://ubit.ischool.washington.edu/pubs/chi01/chi2001.pdf.

JACKSON, A. 1999. Web page design: A study of three genres. Master's paper, University of North Carolina - Chapel Hill.

JACKSON-SANBORN, E., ODESS-HARNISH, K., AND WARREN, N. 2002. Website accessibility: A study of ADA compliance. Tech. Rep. TR-2001-05, University of North Carolina - Chapel Hill, School of Information and Library Science. http://ils.unc.edu/ils/research/reports/accessibility.pdf.

JAIN, R. 1991. *The Art of Computer Systems Performance Analysis*. Wiley-Interscience, New York.

KAISER, J. 1998. Browser-safe colors. *Web Design June 15*. http://webdesign.about.com/compute/webdesign/library/weekly/aa061598.htm.

KIM, N. AND FOGG, B. J. 1999. World Wide Web credibility: What effects do advertisements and typos have on the perceived credibility of web page information? Ph.D. thesis, Stanford University.

KLEE, M. AND SCHROEDER, W. 2000. Report 2: How business goals affect site design. In *Designing Information-Rich Web Sites*. User Interface Engineering, Bradford, MA.

KOYANI, S., ALLISON, S., BAILEY, R., CHAPARRO, B., IVORY, M., AND WHEELER, S. 2003a. Use of research-based guidelines in the development of websites. In *Proceedings of the Conference on Human Factors in Computing Systems*. Extended Abstracts. New York, Fort Lauderdale, FL, 696–697. http://ubit.ischool.washington.edu/pubs/chi03/sigchi03.pdf.

KOYANI, S. J., BAILEY, R. W., AND NALL, J. R. 2003b. *Research-Based Web Design & Usability Guidelines*. National Institutes of Health.

LANDESMAN, L. AND SCHROEDER, W. 2000. Report 5: Organizing links. In *Designing Information-Rich Web Sites*. User Interface Engineering, Bradford, MA.

LIN, J. AND LANDAY, J. A. 2002. Damask: A tool for early-stage design and prototyping of multi-device user interfaces. In *Proceedings of The 8th International Conference on Distributed Multimedia Systems*. San Francisco, CA, 573–580.

LOWGREN, J. AND NORDQVIST, T. 1992. Knowledge-based evaluation as design support for graphical user interfaces. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM Press, Monterey, CA, 181–188.

MADDEN, M. AND RAINIE, L. 2003. America's online pursuits: The changing picture of who's online and what they do. http://www.pewinternet.org/pdfs/PIP_Online_Pursuits_Final.PDF.

MAHAJAN, R. AND SHNEIDERMAN, B. 1997. Visual and textual consistency checking tools for graphical user interfaces. *IEEE Trans. Soft. Eng. 23*, 11, 722–735.

MURCH, G. M. 1985. Colour graphics—blessing or ballyhoo? *Computer Graphics Forum 4*, 2, 127–135.

NATIONAL CANCER INSTITUTE. 2001. Research-based web design & usability guidelines. http://usability.gov/guidelines/.

NIELSEN, J. 1998. Web usability: Why and how. *Users First! September 14*. http://www.zdnet.com/devhead/stories/articles/0,4413,2137433,00.html.

NIELSEN, J. 1999. User interface directions for the web. *Comm. ACM 42*, 1, 65–72. http://www.acm.org:80/pubs/citations/journals/cacm/1999-42-1/p65-nielsen/.

NIELSEN, J. 2000. *Designing Web Usability: The Practice of Simplicity*. New Riders Publishing, Indianapolis, IN.

NTOULAS, A., CHO, J., AND OLSTON, C. 2004. What's new on the web?: the evolution of the web from a search engine perspective. In *Proceedings of the 13th international conference on World Wide Web*. ACM Press, New York, NY, USA, 1–12.

RATNER, J., GROSE, E. M., AND FORSYTHE, C. 1996. Characterization and assessment of HTML style guides. In *Proceedings of the Conference on Human Factors in Computing Systems*. Vol. 2. ACM Press, Vancouver, Canada, 115–116.

ROSENFELD, L. AND MORVILLE, P. 1998. *Information Architecture for the World Wide Web*. O'Reilly & Associates, Sebastopol, CA.

SANO, D. 1996. *Designing Large-Scale Web Sites: A Visual Design Methodology*. Wiley Computer Publishing, John Wiley & Sons, Inc., New York, NY.

SAWYER, P., DANCA, R., AND SCHROEDER, W. 2000. Report 6: Myths of page layout. In *Designing Information-Rich Web Sites*. User Interface Engineering, Bradford, MA.

SAWYER, P. AND SCHROEDER, W. 2000. Report 4: Links that give off scent. In *Designing Information-Rich Web Sites*. User Interface Engineering, Bradford, MA.

SCANLON, T. AND SCHROEDER, W. 2000a. Report 1: What people do with web sites. In *Designing Information-Rich Web Sites*. User Interface Engineering, Bradford, MA.

SCANLON, T. AND SCHROEDER, W. 2000b. Report 7: Designing graphics with a purpose. In *Designing Information-Rich Web Sites*. User Interface Engineering, Bradford, MA.

SCHMETZKE, A. 2004a. Web accessibility survey homepage. `http://library.uwsp.edu/aschmetz/Accessible/websurveys.htm`.

SCHMETZKE, A. 2004b. Web page accessibility on University of Wisconsin campuses: 2004 survey and six-year trend data. `http://library.uwsp.edu/aschmetz/Accessible/UW-Campuses/Survey2004/contents2004.htm`.

SCHRIVER, K. A. 1997. *Dynamics in Document Design*. Wiley Computer Publishing, John Wiley & Sons, Inc., New York, NY.

SHEDROFF, N. 1999. Recipe for a successful web site. `http://www.nathan.com/thoughts/recipe`.

SHNEIDERMAN, B. 1997. Designing information-abundant web sites: Issues and recommendations. *Int. J. Hum.-Comput. Studies 47*, 1, 5–29.

SMITH, S. L. 1986. Standards versus guidelines for designing user interface software. *Behaviour and Information Technology 5*, 1, 47–61.

SOUZA, F. D. AND BEVAN, N. 1990. The use of guidelines in menu interface design: Evaluation of a draft standard. In *Proceedings of the Third IFIP TC13 Conference on Human-Computer Interaction*, G. Cockton, D. Diaper, and B. Shackel, Eds. Elsevier Science Publishers, Cambridge, UK, 435–440.

SPOOL, J. M., KLEE, M., AND SCHROEDER, W. 2000. Report 3: Designing for scent. In *Designing Information-Rich Web Sites*. User Interface Engineering, Bradford, MA.

SPOOL, J. M., SCANLON, T., SCHROEDER, W., SNYDER, C., AND DEANGELO, T. 1999. *Web Site Usability: A Designer's Guide*. Morgan Kaufmann Publishers, Inc., San Francisco, CA.

STEIN, L. D. 1997. The rating game. `http://stein.cshl.org/~lstein/rater/`.

THE INTERNATIONAL ACADEMY OF ARTS AND SCIENCES. 2000. The webby awards 2000 judging criteria. `http://www.webbyawards.com/judging/criteria.html`.

THE POYNTER INSTITUTE. 2004. Poynter online - design / graphics. `http://www.poynter.org/subject.asp?id=11`.

VAN DUYNE, D. K., LANDAY, J. A., AND HONG, J. I. 2002. *The Design of Sites: Patterns, Principles, and Processes for Crafting a Customer-Centered Web Experience*. Addison-Wesley, Boston.

VORA, P. R. 1998. Design/methods & tools: Designing for the web: A survey. *interactions 5*, 3, 13–30.

W3C. 1999. Web content accessibility guidelines 1.0. `http://www.w3.org/TR/WAI-WEBCONTENT/`.

WATCHFIRE. 2002. Welcome to bobby worldwide. `http://bobby.watchfire.com/bobby/html/en/index.jsp`.

WITTEN, I. H. AND FRANK, E. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco.

Received November 2004; revised July 2005; accepted July 2005