Chapter 7

Linking Web Interface Profiles to Usability

7.1 Introduction

Chapter 6 demonstrated that profiles of highly-rated Web interfaces could be developed based on key quantitative measures. However, it is not clear what these profiles represent – highly usable, aesthetically-pleasing or perhaps merely popular pages. Two studies were conducted by the author to provide some insight about what the profiles represent. The first study (discussed in this chapter) evaluates what went into developing the profiles of Web interfaces, namely the expert ratings. The second study (discussed in Chapter 9) evaluates the results of applying the Web interface profiles.

This chapter discusses a usability study conducted to determine the relationship between Webby judges' scores and ratings assigned by participants (non experts) who used sites to complete tasks. The goal of this study was to determine if judges' scores were consistent with usability ratings for sites at the extremes of the rating scale (i.e., sites with overall scores in the top and bottom 33% of Webby sites).

The study produced usability ratings for 57 Web sites that were used in the development of profiles in Chapter 6. Thirty participants completed the study and rated sites in two conditions: after simply exploring the site (referred to as perceived usability); and after exploring and completing three information-seeking tasks on the site (referred to as actual usability). Participants rated sites in both conditions using the WAMMI usability scale [Kirakowski and Claridge 1998]. The analysis focused on answering the following questions.

- Are Webby judges' scores consistent with perceived usability ratings?
- Are Webby judges' scores consistent with actual usability ratings?

Analysis of study data suggests some consistency between the Webby judges' scores and both the actual and perceived usability ratings. One of the major limitations of this study is that it was conducted at least six months after sites were reviewed by Webby judges. Hence, sites may have undergone major changes in the interim. It may have been possible to use the Internet Archive[1] to determine whether sites had changed; however, the site evaluation dates were unknown and this information was needed for the assessment. Despite measured consistency between judges' and participants' ratings, a strong conclusion about judges' scores reflecting usability cannot be

[1] Available at http://www.archive.org/.

made from this study. The methodology described in this chapter could be repeated in a more ideal setting to enable stronger conclusions to be made.

7.2 User Study Design

A Web site usability study[2] was conducted between November 11, 2000 and November 17, 2000, in accordance with guidelines established by the Committee for the Protection of Human Subjects (project 2000-1-36). Thirty participants completed a between-subjects experiment wherein they explored and rated 22 sites in two conditions: simply exploring the site; and completing information-seeking tasks after exploring the site. Participants rated each site in only one condition. However, they rated eleven sites after exploring them and the other eleven sites after exploring and completing tasks on them. A total of 57 sites were evaluated by participants using the WAMMI (Website Analysis and MeasureMent Inventory) [Kirakowski and Claridge 1998] usability scale.

7.2.1 Study Sites and Tasks

For the analysis, 60 sites were selected from the Webby Awards 2000 dataset studied in Chapter 6. Half of the sites were from the good sample (top 33% of reviewed sites), and the other half were from the poor sample (bottom 33% of reviewed sites). All of the sites fell within the top/bottom cutoffs discussed in Section 6.3.1; this was the case for the overall Webby score and the Webby factor (variable derived via principal components analysis to summarize the six rating criterion) across all content categories as well as within each content category. There was equal representation among the six categories (Community, Education, Health, Finance, Living, and Services). Furthermore, the selected sites met the following criteria: the site used the English language; the site was not implemented with Macromedia Flash; and the site did not require login. Three of the sites became unavailable or malfunctioned during the course of this study; thus, results are only reported for 57 sites. In some cases participants experienced technical difficulties; hence, responses were eliminated in these situations as well.

Three information-seeking tasks were developed for sites. First, sites within each of the six categories were explored to develop two general tasks, such as finding information about how to contact the company or the major product/service offered through the site. Table 7.1 contains the two general tasks developed for each content category. General tasks required participants to locate information that has been noted as essential to good Web design practices in the literature [Fogg *et al.* 2000; Nielsen 2000; Sano 1996]. Each site was then explored to identify a site-specific information-seeking task. These tasks were non obvious, required participants to follow at least five links through the site, and were comparable to tasks chosen for the other sites in the content category. Tables 7.2 and 7.3 summarize site-specific tasks.

A testing interface was developed using HTML forms, JavaScript, and PERL. Figure 7.1 depicts the screens for performing information-seeking tasks in the actual usability condition. In addition, a script was developed to generate 30 randomized experiment designs (i.e., the original 60 sites where randomly assigned to these experiments). Each experiment consisted of 22 sites, two of which were for training. The order of site exploration was randomized as well as the presentation of the three tasks. Two pilot studies were conducted to improve the final study design, testing interface, and testing materials.

[2] Rashmi Sinha and Marti Hearst provided valuable input into the study design, assisted with recruiting participants, and facilitated several testing sessions. Sinha also conducted a preliminary analysis of the study data.

Content Category	Information to locate in the site
Community	How to contact the company What topics are discussed on this site
Education	What educational products/services are offered through this site What topics/disciplines does this site address
Finance	What is the major financial product/service offered through this site Get a sense for whom this site is meant for
Health	What topics are discussed on this site Find contact information for the site
Living	Get a sense for whom this site is meant for What is the major product/service offered through this site
Services	What is the major product/service offered through this site How to contact the company

Table 7.1: General information-seeking tasks for each content category.

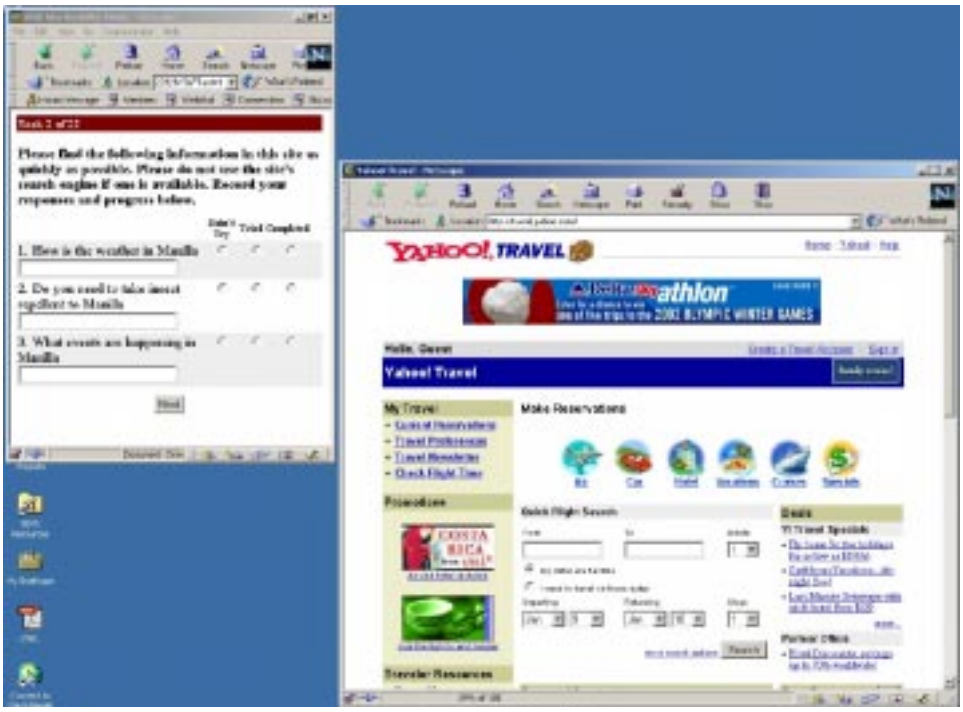

Figure 7.1: Testing interface for completing the information-seeking tasks in the actual usability condition. The smaller browser window provides instructions, and the larger window displays the site.

Id	Rating	Information to locate in the site
Community		
1_242	Good	Article on computer privacy
1_250	Good	A good age to start teaching children to swim
1_255	Good	The height of the second tallest mountain in California
1_261	Good	How to find donated items for your local charity
1_262	Good	Discussions on rising heat costs
1_003	Poor	Details about the movie The Cell
1_006	Poor	Details about the Age of Attention
1_007	Poor	Details about recent findings in Gamla
1_009	Poor	List of items you need for playing rugby
1_015	Poor	Details about horsepower
Education		
2_205	Good	Different kinds of computers
2_208	Good	Causes of drought
2_209	Good	School voucher issues
2_217	Good	Courses on snowboarding
2_218	Good	Details about Madagascar
2_012	Poor	Benefits of water birth
2_013	Poor	Leadership case studies
2_016	Poor	The student who designed the Iceman poster
2_017	Poor	Requirements for being a student anchor
2_018	Poor	Details about the Brazil training stress
Finance		
3_087	Good	Status of car sales in the US
3_098	Good	Basic investment options
3_101	Good	Details about student loans
3_102	Good	Investment strategies for college students
3_103	Good	Details about Roth IRAs
3_002	Poor	How the market affects this company's funds
3_014	Poor	How a loan officer evaluates your credit report
3_018	Poor	How Theresa Pan became a billionaire
3_019	Poor	Return for the best stock picked by Sam Isaly
3_022	Poor	How education influences compensation of financial executives

Table 7.2: Site-specific information-seeking tasks (Community, Education, and Finance).

Id	Rating	Information to locate in the site
Health		
4_127	Good	Techniques for coping with stress
4_128	Good	The vaccine for Lyme disease
4_129	Good	Flu shots for elders
4_130	Good	Impact of second-hand smoke on asthma sufferers
4_132	Good	Latest findings on job-related depression
4_005	Poor	Links to online health statistics
4_008	Poor	Details about e-pharmacy.MD
4_018	Poor	Issues that are best suited for online therapy
4_020	Poor	Treatment for common colds
4_021	Poor	Advice for seeking a second medical opinion
Living		
5_137	Good	Details about The Zone diet
5_138	Good	Using sesame seed in recipes
5_139	Good	Taking a virtual tour of New York
5_145	Good	Cost for copyrighting software
5_002	Poor	Menu for an inexpensive dinner party
5_005	Poor	Recipe for Chili Cornmeal Biscuits
5_008	Poor	Price list for California wines
5_014	Poor	Creating a wish list
5_021	Poor	Whether you can recycle the tray from your food package
Services		
6_115	Good	How to receive faxes online
6_116	Good	Cost for a recycled front door window for a 1992 Honda Civic DX Hatchback
6_122	Good	Creating a wish list
6_132	Good	Use of foods to remedy medical problems
6_008	Poor	The ad for a postcard designed for students
6_009	Poor	Details about Desktop Drive
6_018	Poor	Cost for developing a web site
6_019	Poor	The newsletter issue that discusses Internet attitudes and usage

Table 7.3: Site-specific information-seeking tasks (Health, Living, and Services).

Id	Age	Gen	Education	CmpExp	IExp	IUse	EngExp	EvalExp
1	18-25	F	College Grad.	Average	Average	>10	Expert	Average
2	18-25	M	Some College	Average	Average	3-5	Average	Average
3	18-25	F	Some College	Beginner	Beginner	1-2	Average	Beginner
4	18-25	M	Some College	Expert	Expert	>10	Expert	Expert
5	18-25	F	Some College	Expert	Expert	>10	Expert	Expert
6	18-25	M	Some College	Expert	Expert	>10	Expert	Beginner
7	18-25	M	Some College	Average	Average	>10	Average	Average
8	18-25	F	Some College	Expert	Expert	>10	Expert	Expert
9	18-25	M	Some College	Average	Average	>10	Average	Average
10	18-25	M	Some College	Average	Expert	>10	Expert	Average
11	18-25	F	Some College	Average	Average	1-2	Expert	Beginner
12	18-25	F	Some College	Average	Average	1-2	Expert	Average
13	18-25	F	≤High School	Average	Average	6-10	Expert	Average
14	18-25	M	Some College	Expert	Expert	>10	Expert	Expert
15	18-25	F	Some College	Average	Expert	>10	Expert	Average
16	26-35	M	Some College	Average	Average	1-2	Expert	Beginner
17	18-25	F	Some College	Average	Average	3-5	Expert	Average
18	18-25	F	Some College	Average	Average	>10	Expert	Average
19	18-25	F	Some College	Average	Average	6-10	Average	Beginner
20	18-25	F	≤High School	Average	Average	1-2	Average	Average
21	18-25	F	Some College	Average	Average	>10	Average	Beginner
22	26-35	M	Some College	Average	Expert	>10	Expert	Average
23	18-25	F	Some College	Average	Average	3-5	Expert	Beginner
24	18-25	F	Some College	Expert	Expert	>10	Expert	Expert
25	18-25	F	Some College	Average	Average	6-10	Average	Beginner
26	36-45	F	Some College	Average	Average	6-10	Expert	Average
27	18-25	F	Some College	Average	Average	6-10	Expert	Expert
28	18-25	F	Some College	Average	Expert	>10	Average	Average
29	18-25	M	≤High School	Expert	Expert	6-10	Expert	Expert
30	18-25	F	Some College	Average	Average	3-5	Average	Beginner

Table 7.4: Summary of participants' demographic information. Participants provided their age and gender (Gen) and described their proficiency with computers (CmpExp), the Internet (IExp), the English language (EngExp), and evaluating Web site quality (EvalExp). Participants also reported the number of hours they spend using the Internet weekly (IUse).

7.2.2 Participants

Study participants were recruited through undergraduate classes and sign-up sheets posted in campus buildings. Thirty participants, primarily UC Berkeley undergraduates, completed the study and were each compensated with $30. Participants answered demographic questions prior to starting the study. Specifically, participants provided their age, gender, as well as information about their education background, computer and Internet experience, the number of hours they use the Internet weekly, English experience, and Web site evaluation experience. Table 7.4 summarizes this information.

All but three of the participants were in the 18–25 age group, and all but four of the participants were undergraduates. Two thirds of the participants were females. Only one participant was a novice computer and Internet user; the remaining participants described themselves primarily as average users. Half of the participants reported using the Internet for over ten hours

a week. Half of the participants also reported having some experience evaluating the ease of use of Web sites. Finally, all of the participants were experienced with the English language.

Based on the demographic information, all of the participants appear to have been appropriate for this study.

7.2.3 Testing Session

The study consisted of a 150-minute session wherein participants interacted with 22 sites; the first two sites were for training. Participants were initially given an instruction sheet (see Figure 7.2) that provided an overview of the study procedure. After reviewing the instruction sheet, participants completed a statement of informed consent and moved to a computer station for completing the study. The study interface requested demographic information and assigned each participant to one of the 30 experiment designs. The interface subsequently guided participants through two types of tasks as discussed below.

- **Rate the site without completing information-seeking tasks.** Participants were instructed to initially explore the site for several minutes. Then, participants were asked to rate the site along a number of criteria and provide freeform comments about the site. This task type assessed the *perceived* usability of a site.

- **Rate the site after completing information-seeking tasks.** Participants were instructed to initially explore the site for several minutes. Then, participants were presented with three tasks and asked to locate the information as quickly as possible without using the site's search engine. The testing interface presented tasks in a randomized order and provided text fields for participants to enter the information that they found; there were also radio buttons for participants to record their progress on each task (didn't try, tried, and completed). The testing interface also recorded the total time spent on the three tasks. Finally, participants were asked to rate the site along a number of criteria and provide freeform comments about the site. This task type assessed the *actual* usability of the site, since participants attempted to use the site.

During the testing session, participants alternated between completing these two types of tasks on the 22 sites. The testing interface timed participants during the exploration and task completion phases. A message window appeared whenever participants spent more than five minutes on either phase.

For the rating phase, participants responded to 20 statements from the WAMMI [Kirakowski and Claridge 1998] questionnaire. WAMMI was the only validated usability scale for Web sites at the time of this study. Responses to WAMMI statements were aggregated into six scales: attractiveness, controllability, efficiency, helpfulness, learnability, and global usability (see Section 7.5). Jurek Kirakowski, Director of the Human Factors Research Group in Ireland, converted participant responses into WAMMI scales for this study. This conversion process also entails normalizing computed scales against a database of other sites that have been assessed with the questionnaire. The reported WAMMI scales are percentiles that reflect how the site's rating contrast to other sites that have been rated with the WAMMI questionnaire.

Participants were also asked about their confidence in responses given to WAMMI statements as well as their opinion about the appropriateness of tasks. Participants responded to all statements using a 5-point Likert scale, ranging from strongly agree (1) to strongly disagree (5); this order was required for consistency with the WAMMI questionnaire. Figure 7.3 depicts all of the statements used during the rating phase. The testing interface presented WAMMI statements

Web Site Usability Study Instructions

I. Overview of the Study
The purpose of this study is to get user feedback on the ease of use of a collection of web sites. Basically, you will explore the site and provide responses to a number of statements about the site's ease of use. There is no risk to you, and you will be compensated at a rate of $12/hour for your participation.

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

II. Overview of the Study Procedure
General goal:
Rate web sites. There will be two kinds of tasks. (a) Find some information on the web site. After you have completed all of the information-seeking tasks, then rate it. (b) Rate the web site after only exploring it.

Tasks
(a) **Rate web site after completing information-seeking tasks on it.**
In the first part of the task, you will be asked to just explore the site. Look around and get familiar with the content, layout, navigation, etc. of the site. **Please do not spend more than a few minutes on the site.**
In the second part of the task, you will be required to complete some information-seeking tasks on the site. Getting familiar with the site in the first part of the task will help you prepare for this part. **You are asked to locate the information as quickly as possible.** We will time you on the task; also try to respond as accurately as you can.
General instructions for the task

- Follow links, don't use search.
- Stay on the site. If by chance you choose a link that takes you out of the site, come back and try to stay on the site.
- Do not spend more than a few minutes exploring the site; you will be reminded.
- The first task is for training.

(b) **Rate web site without completing tasks on it.**
You will go to a particular web site and explore it. Then rate the site. **Do not spend more than a few minutes exploring the site; you will be reminded.** The first task is for training.

Ratings:
Ratings are for 21 criteria on a scale. For each statement you will provide a response that reflects your agreement with the statement (Strongly Agree – Strongly Disagree). You can make any general (free form) comments that you might have. Finally, please note your prior experience with the site before this study.

Take breaks whenever you want to. If you need a break, try to take one in between two sites, rather than in the middle of completing tasks on a site. Remember not to spend too much time on one site; there are 22 sites in the study.

Figure 7.2: Instructions given to study participants.

in a randomized order to increase the likelihood that participants actually read statements before responding.

Both the computed WAMMI scales and the original responses to WAMMI statements were analyzed. Likert scales for positive statements were inverted such that positive responses resulted in higher scores (i.e., higher is better); this adjusts responses to statements 2, 3, 5, 6, 7, 8, 10, 11, 15, 17, 20, 21, and 22 as depicted in Figure 7.3. Overall, participants reported that they were highly confident with their responses (statement 21) and felt that the tasks were useful (statement 22). Figure 7.4 depicts distributions of responses to these two statements.

Participants had the opportunity to provide freeform comments and to record the number of times (never, 1–3 times, and 3 or more times) they had used a site prior to this study. In all but three cases, participants had never used the sites in the study. Two participants reported using a site more than three times, and another participant reported using a site 1–3 times. These responses were eliminated from the analysis, since they were potentially biased based on prior use. In addition to subjective ratings, the testing interface recorded timing information for the exploration, task completion, and rating phases.

7.2.4 Testing Environment

Participants completed the study in one of seven group sessions in UC Berkeley's School of Information Management and Science's second-floor computer lab. Participants worked individually at a computer station during these sessions. Computer stations had PCs running Microsoft Windows NT 4.0 with 128 MB of RAM. Stations also had 17" monitors (1280 x 1024 pixels) and high speed, campus Ethernet connections. Participants used the Netscape Navigator 4.7 browser. The testing interface resized the browser window to 800 x 600 pixels (equivalent to a 15" monitor), and sites were not cached in order to provide a realistic experience. User surveys have shown that over 50% of Internet users access sites with 800 x 600 monitor resolution and 56K and slower connection speeds [DreamInk 2000].

7.3 Data Collection

Several sites became unavailable during the course of this study, and in some cases participants were unable to rate sites due to time constraints; responses were eliminated for these situations. Responses were also eliminated for training sites and the three cases where participants had used sites prior to the study, since prior use may suggest a bias. Finally, one case was eliminated wherein the participant did not attempt any tasks in the actual usability condition.

The final dataset consisted of 550 cases where each case included participant's responses to the statements in Figure 7.3, computed WAMMI scales (attractiveness, controllability, efficiency, helpfulness, learnability, and global usability), Webby scores (content, structure and navigation, visual design, functionality, interactivity, and overall experience), and the Webby factor for the site, participant's demographic information, objective measures (e.g., exploration and rating time), and participant's comments. There were 271 cases for the actual usability condition (i.e., with information-seeking tasks) and 279 cases for the perceived usability condition (i.e., without information-seeking tasks). Table 7.5 summarizes the distribution of cases. All of the sites were rated in both the actual and perceived usability conditions. An average of five participants rated each site in the two conditions; thus, a site was rated by an average of ten participants overall.

Each of the 550 cases contained 66 fields of information as described below.

Rating Criteria

1. This web site has some annoying features. (-, annoy)
2. I feel in control when I'm using this web site. (+, control)
3. Using this web site for the first time is easy. (+, easy)
4. This web site is too slow. (-, slow)
5. This web site helps me find what I am looking for. (+, find)
6. Everything on this web site is easy to understand. (+, clear)
7. The pages on this web site are very attractive. (+, pretty)
8. This web site seems logical to me. (+, logical)
9. It is difficult to tell if this web site has what I want. (-, nofind)
10. This web site has much that is of interest to me. (+, interest)
11. I can quickly find what I want on this web site. (+, qfind)
12. It is difficult to move around this web site. (-, nonav)
13. Remembering where I am on this web site is difficult. (-, noremind)
14. Using this web site is a waste of time. (-, waste)
15. I can easily contact the people I want to on this web site. (+, contact)
16. I don't like using this web site. (-, nolike)
17. I get what I expect when I click on things on this web site. (+, expect)
18. This web site needs more introductory explanations. (-, nointros)
19. Learning to find my way around this web site is a problem. (-, nolearn)
20. I feel efficient when I'm using this web site. (+, effic)
21. I feel very confident about my responses to the previous statements regarding this web site. (+, conf)
22. These tasks enabled me to get an overview of the site. (+, taskuse, for information-seeking tasks only)

Figure 7.3: Rating criteria used during the study. The first 20 statements are from the WAMMI questionnaire. Statements are noted as positive (+) or negative (-) and a short descriptor is provided for each.

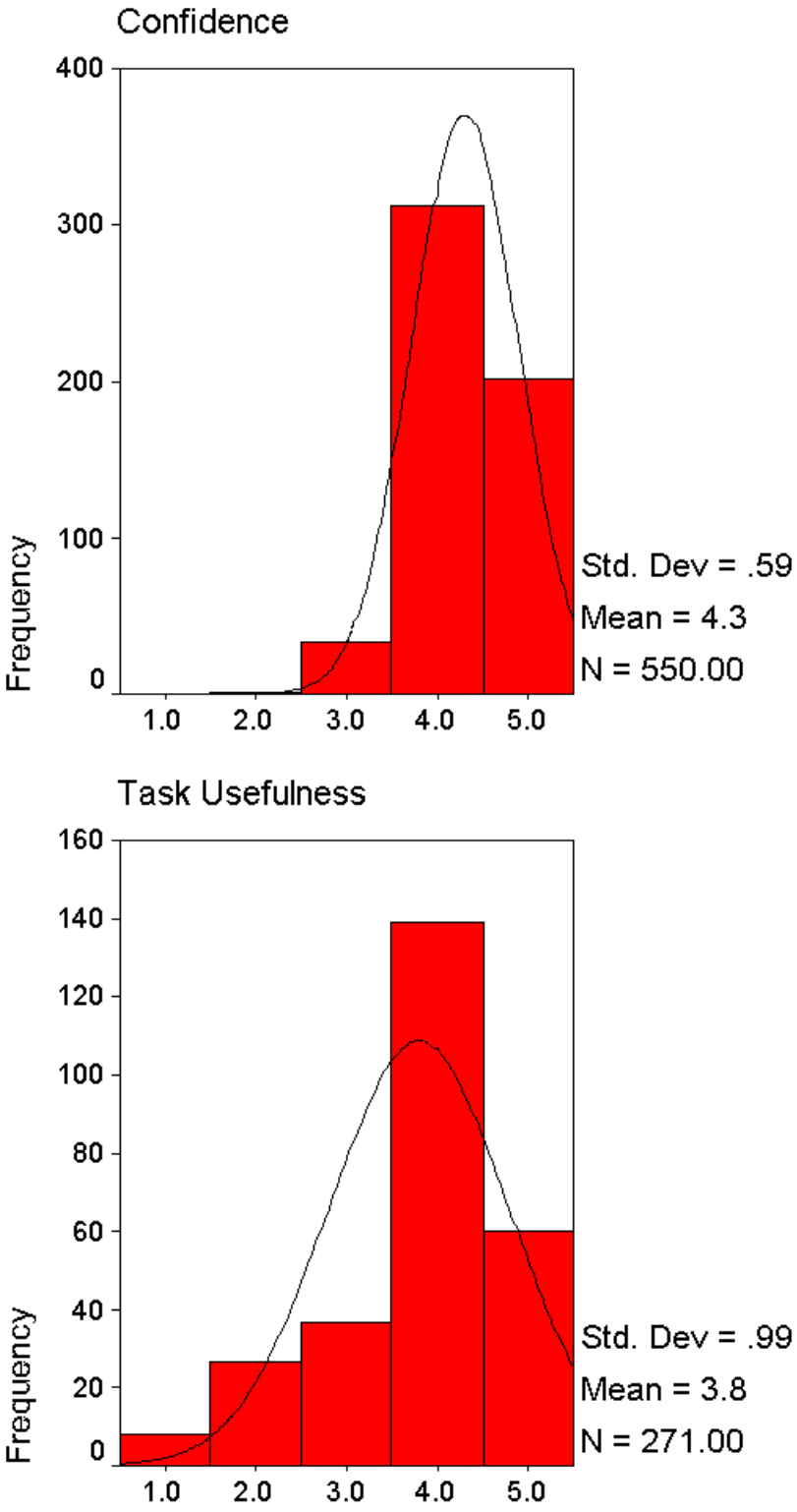

Figure 7.4: Study participants' confidence with responses to WAMMI statements (top graph) and reported usefulness of tasks in the actual usability condition (bottom graph). Responses range from low (1) to high (5).

Category	Good	Poor	Total
Actual Usability			
Community	25	25	50
Education	23	22	45
Finance	22	24	46
Health	24	25	49
Living	18	24	42
Services	20	19	39
Total	132	139	271
Perceived Usability			
Community	23	25	48
Education	25	25	50
Finance	24	25	49
Health	25	24	49
Living	19	25	44
Services	19	20	39
Total	135	144	279

Table 7.5: Number of cases (ratings) in the analyzed sample.

7.3.1 Basic Information

Site Identification: the site's category (Community, Education, Finance, Health, Living, or Services), URL, title, id, and whether the site belonged to the top or bottom 33% of reviewed sites (i.e., good or poor class).

Study Condition: participant id and experiment condition (actual or perceived usability).

Participant Demographics: participant's age, gender, educational background, computer experience, Internet experience, weekly Internet use, English experience, and Web site evaluation experience. Table 7.4 summarizes this information.

7.3.2 Objective Measures

Timing: site exploration time (exptime), time spent completing tasks in the actual usability condition (tasktime), and time spent rating the site in both conditions (ratetime). Total time spent on the site (exp+task) was also computed for analysis; this time is the same as site exploration time in the perceived usability condition.

Task Completion: whether the participant did not attempt to complete, attempted to complete, or completed each of the three tasks (t1suc, t2suc, and t3suc). The number of tasks that were not attempted (notry), not completed (nocomp), and completed (comp) were also tallied.

7.3.3 Subjective Measures

Responses to WAMMI Statements: subjective measures included the participant's responses to statements 1 through 20 of Figure 7.3 – annoy, control, easy, slow, find, clear, pretty, logical, nofind, interest, qfind, nonav, noremind, waste, contact, nolike, expect, nointros, nolearn, and effic – respectively. Positive statements were inverted as previously discussed.

Computed WAMMI Scales: attractiveness (wamattr), controllability (wamcon), efficiency (wameff), helpfulness (wamhelp), learnability (wamlearn), and global usability (wamglob).

Composite WAMMI Score: There was an attempt to compute a composite WAMMI score similarly to the computed Webby factor. However, there was little correlation among the six WAMMI scales. Although, there were large positive correlations between global usability and attractiveness, controllability, helpfulness, and learnability, there was a small negative correlation with efficiency. Hence, it was not possible to compute a single factor that could explain most of the variance in these scales. Instead, responses to the 20 statements were summed to represent a composite score (wamsum). Section 7.4 discusses this in more detail.

Webby Scores: content (content), structure and navigation (nav), visual design (vis), functionality (funct), interactivity (int), and overall experience (overall).

Composite Webby Score: the computed Webby factor (webbyfac); this measure was derived via principal components analysis over the six Webby scores.

Other Measures: confidence with responses to WAMMI statements (conf), usefulness of tasks in the actual usability condition (taskuse), and the number of times the participant had used the site prior to the study (prioruse).

Task Responses: answers and comments provided for each of the three information-seeking tasks (t1resp, t2resp, and t3resp) in the actual usability condition.

Comments: freeform comments on the site's ease of use.

7.3.4 Screening of Objective and Subjective Measures

The data was screened to replace extremely large and small outliers with the next smallest or largest values as appropriate for each of the objective and subjective measures; this is a standard statistical process used to remove potential errors in measurement from analysis data [Easton and McColl 1997; SPSS Inc. 1999]. Tests for normality and equal variances revealed that most of these measures did not follow a normal distribution, although most exhibited equal variances. Standard statistical analysis techniques assume these two conditions; hence, it was not possible to use parametric techniques with this dataset. Applying transformations to stabilize the data, such as square roots and reciprocals of square roots, were unsuccessful. Nonparametric analysis techniques were used during analysis, since they do not require data to satisfy the normality and equal variances conditions.

7.4 Developing a Composite WAMMI Score

When participants respond to multiple questions to provide subjective ratings, it is a common practice to summarize these responses with one factor. For example, the Questionnaire for User Interaction Satisfaction (QUIS) [Harper and Norman 1993] requires participants to rate an interface on 27 facets, the responses can then be summed to produce an overall measure of user satisfaction; it is also possible to aggregate subsets of responses into several interface factors, including system feedback and learning. For the second metrics study, principal components analysis was used to produce a composite Webby score – the Webby factor; this factor summarized judges'

Scale	wamattr	wamcon	wameff	wamhelp	wamlearn	wamglob
wamattr	1.00	**0.17**	-0.03	**0.16**	**0.11**	**0.49**
wamcon	**0.17**	1.00	**-0.35**	**0.45**	**0.17**	**0.55**
wameff	-0.03	**-0.35**	1.00	**-0.41**	**-0.18**	-0.07
wamhelp	**0.16**	**0.45**	**-0.41**	1.00	**0.11**	**0.64**
wamlearn	**0.11**	**0.17**	**-0.18**	**0.11**	1.00	**0.52**
wamglob	**0.49**	**0.55**	-0.07	**0.64**	**0.52**	1.00

Table 7.6: Correlations between the 6 WAMMI scales. Bold entries are significant.

ratings for the six criteria and made it possible to produce more accurate predictions than with the overall experience score [Ivory *et al.* 2001].

Similarly to the Webby factor, there was an attempt to compute a composite WAMMI factor based on the six scales. Table 7.6 shows correlations between pairs of WAMMI scales; Spearman correlation coefficients were computed, since this method is appropriate for nonparametric data. With the exception of the global usability scale (wamglob), there was only small to medium correlations between scales; correlations were surprisingly negative in some cases. The global usability scale had large positive correlations with all of the scales, except efficiency (wameff). Inconsistency among the scales is somewhat expected, since there were only five or fewer responses for each site in each of the two conditions. According to the developers of the WAMMI scales, scales typically require 20 or more responses to stabilize[3]. Given the instability of the scales, it was not possible to construct a single factor to summarize them. Several factor analysis approaches were used, such as principal components and maximum likelihood with and without rotation; this analysis could only produce two factors that explained 63% of the total variance in the scales.

A composite score was computed by summing responses to the 20 statements; this composite score, wamsum, ranges from 20 to 100 and naturally correlates with all of the statements. Figure 7.5 shows that most of the sums were towards the middle of the range.

7.5 Mapping Between WAMMI Scales and Webby Scores

As discussed in Section 7.3.1, subjective measures included the computed WAMMI scales (attractiveness, controllability, efficiency, helpfulness, learnability, and global usability) and the Webby scores (content, structure and navigation, visual design, functionality, interactivity, and overall experience). Kirakowski and Claridge [1998] claim that WAMMI scales were developed and validated through empirical studies, as is typically done with psychological scales. Although not verified, the Webby scores were developed based on consensus of the members of the International Academy of Digital Arts & Sciences. It is not clear how the academy members derived the scores or what instructions were given to judges to facilitate ratings.

Based solely on the descriptions of Webby scores and WAMMI scales, there appeared to be direct mappings between five criteria in the two rating schemes as depicted in Table 7.7. The learnability and interactivity criteria did not appear to be closely related. There are several key differences between the two ratings schemes:

- Webby scores are average ratings (typically for three judges), while WAMMI scales are computed from responses to individual statements and normalized against other assessed sites.

[3]Personal communication with Jurek Kirakowsi on December 2, 2000.

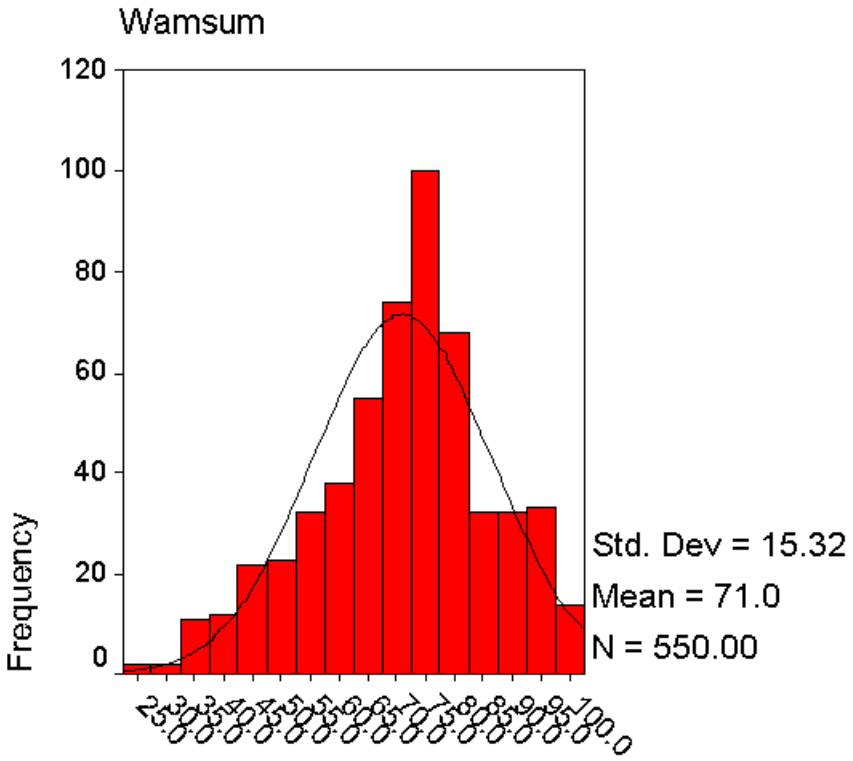

Figure 7.5: Wamsum scores (sums of responses to the 20 WAMMI statements) for the analyzed sample.

WAMMI Scales	**Webby Scores**
Helpfulness – "corresponds with the users' expectations about its content and structure..."	*Content* – "the information provided on the site..."
Controllability – "the users most probably feel they can navigate around it with ease..."	*Structure and Navigation* – "the organization of information on the site and the method in which you move through sections..."
Attractiveness – "An attractive site is visually pleasant..."	*Visual Design* – "appearance of the site..."
Efficiency – user "can quickly locate what is of interest to them and they feel that the web site responds..."	*Functionality* – "the site loads quickly, has live links, and any new technology used is functional and relevant..."
Global Usability – "a site must make it easy for users to access what they need or want..."	*Overall Experience* – "encompasses content, structure and navigation, visual design, functionality, and interactivity, but it also encompasses the intangibles that make one stay or leave..."
Learnability – "users feel they are able to start using the site with the minimum of introductions..."	–
–	*Interactivity* – "the way that a site allows a user to do something..."

Table 7.7: Description of the WAMMI usability scales and Webby scores in conjunction with mappings between these two rating schemes. Mappings are based solely on descriptions.

Criterion	Min.	Max.	Mean	Std. Dev.	Med.
Webby Scores					
content	1.3	10.0	6.4	2.5	7.0
nav	1.3	9.3	6.1	2.3	6.7
vis	1.3	9.0	5.5	2.1	6.0
funct	1.0	10.0	5.6	2.5	6.0
int	1.3	9.5	6.4	2.2	7.3
overall	1.0	10.0	5.9	2.5	5.3
WAMMI Scales					
wamattr	5.0	62.0	26.2	12.1	25.0
wamcon	2.0	65.0	15.8	14.5	11.0
wameff	4.0	61.0	29.6	13.1	27.0
wamhelp	2.0	83.0	27.8	20.9	19.0
wamlearn	4.0	71.0	28.8	15.4	24.5
wamglob	11.0	50.0	25.3	7.5	24.0

Table 7.8: Descriptive statistics for Webby scores and WAMMI scales.

Score	content	nav	vis	funct	int	overall
content	1.00	0.91	0.83	0.90	0.90	0.96
nav	0.91	1.00	0.90	0.90	0.93	0.93
vis	0.83	0.90	1.00	0.84	0.89	0.87
funct	0.90	0.90	0.84	1.00	0.93	0.91
int	0.90	0.93	0.89	0.93	1.00	0.93
overall	0.96	0.93	0.87	0.91	0.93	1.00

Table 7.9: Correlations between the six Webby scores. All of the entries are significant.

WAMMI scales are actually percentiles (e.g., a scale value of 70 means that the site scored better than 70% of Web sites and worse than 30% of sites).

- Webby scores range from 1 to 10, while WAMMI scales range from 1 to 100 (see Table 7.8); and
- All of the Webby scores are strongly correlated with each other (see Spearman correlation coefficients in Table 7.9), while WAMMI scales are not (see Table 7.6).

Despite differences between these rating schemes, they were used to directly compare judges' and participants' ratings for sites. To facilitate comparison, the WAMMI scales and Webby scores were transformed into Z scores (see Table 7.10). Each Z score is a standard deviation unit that indicates the relative position of each value within the distribution (i.e., $Z_i = \frac{x_i - \bar{x}}{\sigma}$, where x_i is the original value, $\bar{x}$ is the mean, and σ is the standard deviation). Ideally, one would start from non-normalized WAMMI scales and then compute the Z scores, but according to the WAMMI scale developers, there is no notion of non-normalized WAMMI scales. Composite WAMMI scores (wamsum) were also transformed into Z scores to compare these measures to the computed Webby factors (webbyfac). (Webby factors were already expressed as Z scores.) Z scores for the composite WAMMI scores were highly correlated with the non-normalized scales; the correlations were also significant.

Criterion	Min.	Max.	Med.
Webby Scores			
content	-2.1	1.5	0.2
nav	-2.1	1.4	0.2
vis	-2.0	1.6	0.2
funct	-1.8	1.5	0.1
int	-2.3	1.4	0.4
overall	-1.8	1.4	-0.2
WAMMI Scales			
wamattr	-1.8	3.0	-0.1
wamcon	-1.0	3.4	-0.3
wameff	-1.9	2.4	-0.2
wamhelp	-1.2	2.6	-0.4
wamlearn	-1.6	2.8	-0.3
wamglob	-1.9	3.3	-0.2

Table 7.10: Descriptive statistics for the Webby scores and WAMMI scales (Z scores). The means and standard deviations for each measure are zero and one, respectively.

Measure	Mean		Std. Dev.	
	Good	Poor	Good	Poor
exptime	**148.5**	**111.9**	93.8	83.1
ratetime	88.5	88.3	34.8	31.7

Table 7.11: Objective measures of perceived usability for good and poor sites. Bold entries represent significant differences in means.

7.6 Perceived Usability Results

The following sections summarize the relationship between participants' subjective and objective data for good and poor sites as well as the consistency between perceived usability ratings and Webby scores.

7.6.1 Perceived Usability of Good and Poor Sites (Objective and Subjective Measures)

Objective measures were analyzed to compare participants' usage of good and poor sites; how these usage patterns correlated with subjective ratings was then studied. Site exploration (exptime) and rating (ratetime) times were the only objective measures in the perceived usability condition. Table 7.11 reports results of comparing these times for sites in the good and poor category using the Mann-Whitney test. (The Mann-Whitney test [SPSS Inc. 1999] is a nonparametric alternative to t-tests for the equality of means, which is typically used for normally-distributed samples; it is related to the Wilcoxon test.) Participants explored good and poor sites for 149 and 112 seconds on average, respectively. Test results showed this difference to be significant (two-tailed p value less than 0.05); there was no significant difference for rating time.

The difference in exploration time suggests that participants spent more time exploring good sites, possibly because they were more usable or interesting. Sinha *et al.* [2001] conducted an empirical analysis of Webby scores for the 2000 Webby Awards dataset, which includes sites in this study; the authors found content to be the best predictor of Web site quality. This finding

provides some support for the hypothesis that participants may have considered good sites to be more interesting.

To test the hypothesis that participants may have considered good sites to be more usable, composite WAMMI scores (wamsum) for good and poor sites were compared. Specifically, the goal was to determine if participants rated good sites lower than poor sites, since they had spent more time on them; this result would have contradicted the hypothesis stated above. Mann-Whitney tests computed over composite ratings for good and poor sites revealed that good sites were rated higher (mean of 75.6 vs. 68.3); this difference was significant.

There were mixed results for the six WAMMI scales, most likely due to their instability; hence, results are not reported for them in this section.

7.6.2 Consistency of Perceived Usability Ratings and Webby Scores (Subjective Measures)

The previous section showed that good sites were rated more favorably on average, thus providing some evidence that there may be some consistency between judges' and participants' ratings of sites. This section explores rating consistency further by comparing mean ratings for the five Webby scores and WAMMI scales that appeared to be directly related based on their descriptions. Specifically, the analysis compares: Webby content (content) to WAMMI help (wamhelp); Webby structure and navigation (nav) to WAMMI controllability (wamcon); Webby visual design (vis) to WAMMI attractiveness (wamattr); Webby functionality to WAMMI efficiency (wameff); and Webby overall (overall) to WAMMI global usability (wamglob). The analysis also compares the Webby factor (webbyfac) to the composite WAMMI score (wamsum).

For this comparison, the WAMMI scales and Webby scores were transformed into Z scores as discussed in Section 7.5. Wilcoxon Signed Ranks tests [SPSS Inc. 1999] were then conducted on the Z scores to study mean differences between site ratings in the two schemes. (The Wilcoxon Signed Ranks test is equivalent to the paired t-test for related variables in normally-distributed samples.) If the Wilcoxon Signed Ranks test reports that a pair of ratings is significantly different, then the ratings are not consistent and vice versa. There was no difference between composite ratings – Webby factor and sum of responses to WAMMI statements; thus, perceived usability ratings were mostly consistent with judges' scores.

Judges' scores were also consistent with usability ratings for the individual scores, except for structure and navigation and controllability. The difference between the navigation and controllability measures may be due to incompatibility. Recall that mappings between Webby scores and WAMMI scales were determined strictly based on their descriptions (see Section 7.5). These two measures may actually be assessing different aspects. Furthermore, the controllability scale may be unstable due to the small number of responses per site.

7.7 Actual Usability Results

The following sections summarize the relationship between participants' subjective and objective data for good and poor sites as well as the consistency between usability ratings based on actual site usage and Webby scores.

Measure	Mean		Std. Dev.	
	Good	Poor	Good	Poor
exptime	**153.4**	**122.7**	96.1	90.5
tasktime	**196.8**	**157.8**	92.6	85.8
ratetime	89.0	92.7	30.6	33.9
t1suc	2.8	2.9	0.5	0.4
t2suc	2.8	2.9	0.5	0.4
t3suc	**2.4**	**2.6**	0.6	0.5
notry	**0.2**	**0.1**	0.5	0.2
nocomp	0.7	0.6	0.7	0.8
comp	**2.1**	**2.3**	0.9	0.9

Table 7.12: Objective measures of actual usability for good and poor sites. Bold entries represent significant differences in means.

7.7.1 Actual Usability of Good and Poor Sites (Objective and Subjective Measures)

Objective measures were analyzed to compare participants' actual usage of good and poor sites to complete information-seeking tasks; how these usage patterns correlated with subjective ratings was then studied. Objective measures in the actual usability condition include: site exploration time (exptime), task completion time (tasktime), rating time (ratetime), whether the participant did not attempt, attempted, or completed each of the three tasks (t1suc, t2suc, and t3suc), the number of tasks the participant did not attempt (notry), the number of tasks the participant did not complete (nocomp), and the number of completed tasks (comp).

Mann-Whitney tests on the perceived usability data showed that participants spent more time exploring good sites, possibly because they were more usable or interesting. Mann-Whitney tests on the actual usability data (see Table 7.12) revealed that exploration time (exptime), task completion time (tasktime), successful completion of the site-specific task (t3suc), the number of tasks not attempted (notry), and the number of tasks completed (comp) were all significantly different between good and poor sites in the full sample. Considering these measures in tandem suggests an interesting pattern – participants spent more time on good sites, but completed fewer tasks than they did on poor sites. The following significant differences supported this pattern:

- participants spent 151 seconds on average exploring good sites versus 125 seconds on poor sites;
- participants spent 198 seconds on average completing tasks on good sites versus 159 seconds on poor sites;
- t3suc measures (2.36 vs. 2.55) indicate that participants completed fewer site-specific tasks on good sites; and
- notry measures (0.18 vs. 0.05) indicate that a larger proportion of tasks (general and site-specific) were not even attempted on good sites.
- comp measures (2.1 vs. 2.3) indicate that fewer tasks (general and site-specific) were completed on good sites.

Based on this pattern, one may naturally expect that participants rated good sites lower than poor sites. However, this was not the case. Good sites had a mean wamsum score of 72.1

vs. 68.2 for poor sites, but this difference was not significant. There were mixed results for the six WAMMI scales, most likely due to their instability; hence, results are not reported for them in this section.

7.7.2 Consistency of Actual Usability Ratings and Webby Scores (Subjective Measures)

The previous section showed that good sites were rated slightly more favorably on average, although the difference was not significant. Results suggest that there may be some consistency between judges' and participants' ratings. To explore rating consistency further, the same comparison of WAMMI and Webby ratings was replicated (see Section 7.6.2). Overall, judges' scores were mostly consistent with participants' ratings. The Webby factor and composite WAMMI score were consistent as well as all other individual Webby scores, except for functionality and efficiency. There was a significant difference between the Webby functionality and the WAMMI efficiency measures, possibly due to the instability of the WAMMI scale.

7.8 Summary

This chapter presented results from a usability study of 57 Web sites wherein 30 participants rated sites in two conditions: after simply exploring sites (perceived usability); and after completing information-seeking tasks on sites (actual usability). The full data collection consisted of 550 cases. The analysis focused on answering the following questions.

- Are judges' scores consistent with perceived usability ratings?
- Are judges' scores consistent with actual usability ratings?

Analysis of objective and subjective data provided evidence that judges' scores are mostly consistent with both actual and perceived usability ratings. This suggests that profiles developed in Chapter 6 reflect usability to some degree. However, concrete conclusions about profiles reflecting usability cannot be drawn from this study due to the time difference between the judges' and users' evaluations. A follow up study needs to be conducted wherein non experts and experts rate identical sites. A better alternative is to develop the profiles from usability ratings in the first place; thus, eliminating the need for such a study. Unfortunately, the Webby Awards dataset was/is (at this time) the only large corpus of sites rated along dimensions that appear to be somewhat related to usability.