A Comparison of Questionnaires for Assessing Website Usability

Thomas S. Tullis and Jacqueline N. Stetson
Human Interface Design Department, Fidelity Center for Applied Technology
Fidelity Investments
82 Devonshire St., V4A
Boston, MA 02109
Contact: tom.tullis@fidelity.com

ABSTRACT:
Five questionnaires for assessing the usability of a website were compared in a study with 123 participants. The questionnaires studied were SUS, QUIS, CSUQ, a variant of Microsoft's Product Reaction Cards, and one that we have used in our Usability Lab for several years. Each participant performed two tasks on each of two websites: finance.yahoo.com and kiplinger.com. All five questionnaires revealed that one site was significantly preferred over the other. The data were analyzed to determine what the results would have been at different sample sizes from 6 to 14. At a sample size of 6, only 30-40% of the samples would have identified that one of the sites was significantly preferred. Most of the data reach an apparent asymptote at a sample size of 12, where two of the questionnaires (SUS and CSUQ) yielded the same conclusion as the full dataset at least 90% of the time.

Introduction

A variety of questionnaires have been used and reported in the literature for assessing the perceived usability of interactive systems, including QUIS [3], SUS [2], CSUQ [4], and Microsoft's Product Reaction Cards [1]. (See [5] for an overview.) In our Usability Lab, we have been using our own questionnaire for the past several years for assessing subjective reactions that participants in a usability test had to a web site. However, we had concerns about the reliability of our questionnaire (and others) given the relatively small number of participants in most typical usability tests. Consequently, we decided to conduct a study to determine the effectiveness of some of the standard questionnaires, plus our own, at various sample sizes. Our focus was specifically on websites.

Method

We decided to limit ourselves to our own questionnaire plus those in the published literature that we believed could be adapted to evaluating websites. The questionnaires we used were as follows (illustrated in Appendix A):

1. SUS (System Usability Scale)—This questionnaire, developed at Digital Equipment Corp., consists of ten questions. It was adapted by replacing the word "system" in every question with "website". Each question is a statement and a rating on a five-point scale of "Strongly Disagree" to "Strongly Agree".
2. QUIS (Questionnaire for User Interface Satisfaction)—The original questionnaire, developed at the University of Maryland, was composed of 27 questions. We dropped three that did not seem to be appropriate to websites (e.g., "Remembering names and use of commands"). The term "system" was replaced by "website", and the term "screen" was generally replaced by "web page". Each question is a rating on a ten-point scale with appropriate anchors at each end (e.g., "Overall Reaction to the Website: Terrible … Wonderful").
3. CSUQ (Computer System Usability Questionnaire)—This questionnaire, developed at IBM, is composed of 19 questions. The term "system" or "computer system" was

replaced by "website". Each question is a statement and a rating on a seven-point scale of "Strongly Disagree" to "Strongly Agree".

4. Words (adapted from Microsoft's Product Reaction Cards)—This questionnaire is based on the 118 words used by Microsoft on their Product Reaction Cards [1]. (We are grateful to Joey Benedek and Trish Miner of Microsoft for providing the complete list.) Each word was presented with a check-box and the user was asked to choose the words that best describe their interaction with the website. They were free to choose as many or as few words as they wished.

5. Our Questionnaire—This is one that we have been using for several years in usability tests of websites. It is composed of nine statements (e.g., "This website is visually appealing") to which the user responds on a seven-point scale from "Strongly Disagree" to "Strongly Agree". The points of the scale are numbered -3, -2, -1, 0, 1, 2, 3. Thus, there is an obvious neutral point at 0.

Note that other tools designed as commercial services for evaluating website usability (e.g., WAMMI [6], RelevantView [7], NetRaker [8], Vividence [9]) were not included in this study. Some of these tools use their own proprietary questionnaires and some allow for the construction of your own.

The entire study was conducted online via our company's Intranet. A total of 123 of our employees participated in the study. Each participant was randomly assigned to one of the five questionnaire conditions. Each was asked to perform two tasks on each of two well-known personal financial information sites: finance.Yahoo.com and Kiplinger.com. (In the rest of this paper they will simply be referred to as Site 1 and Site 2. No relationship between the site numbers and site names should be assumed.) The two tasks were as follows:

1. Find the highest price in the past year for a share of <company name>. (Note that a different company was used in each task.)
2. Find the mutual fund with the highest 3-year return.

The order of presentation of the two sites was randomized so that approximately half of the participants received Site 1 first and half received Site 2 first. After completing (or at least attempting) the two tasks on a site, the user was presented with the questionnaire for their randomly selected condition. Thus, each user completed the same questionnaire for the two sites. (Technically, "questionnaires" was a between-subjects variable and "sites" was a within-subjects variable.)

Data Analysis

For each participant, an overall score was calculated for each website by simply averaging all of the ratings on the questionnaire that was used. (All scales had been coded internally so that the "better" end corresponded to higher numbers.) Since the various questionnaires use different scales, these were converted to percentages by dividing each score by the maximum score possible on that scale. So, for example, a rating of 3 on SUS was converted to a percentage by dividing that by 5 (the maximum score for SUS), giving a percentage of 60%.

Special treatment was required for the "Words" condition since it did not involve rating scales. Before the study, we classified each of the words as being "Positive" (e.g., "Convenient") or "Negative" (e.g., "Unattractive"). (Note that they were not grouped or identified as such to the participants.) For each participant, an overall score was calculated by counting the total number of words that person selected and then dividing that number

into the number of those words that were "Positive". Thus, if someone selected 8 positive words and 10 words total, that yielded a score of 80%.

Results

The random assignment of participants to the questionnaire conditions yielded between 19 and 28 participants for each questionnaire. The frequency distributions of their ratings on each questionnaire for each site, converted to percentages as described above, are shown in Figures 1 through 5. Figure 6 shows the average scores for each site using each questionnaire.
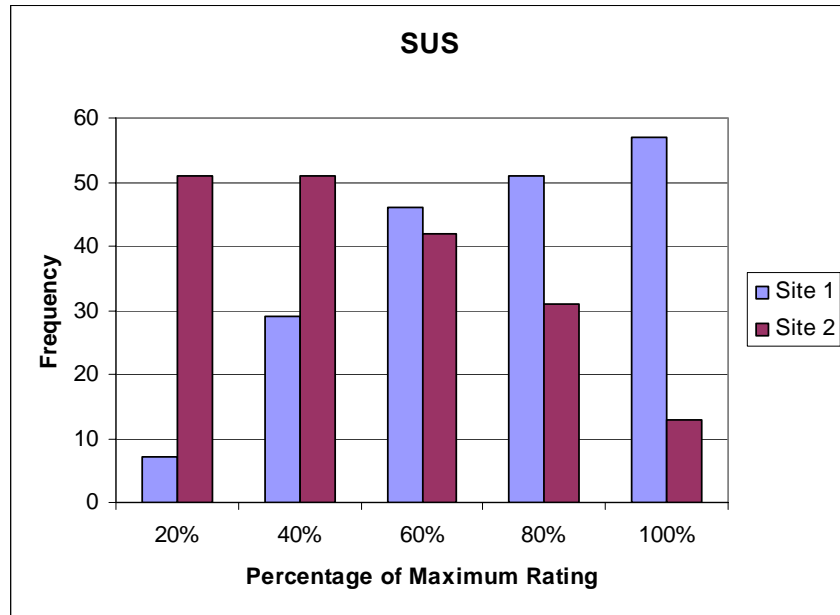
Figure 1. Results using SUS.
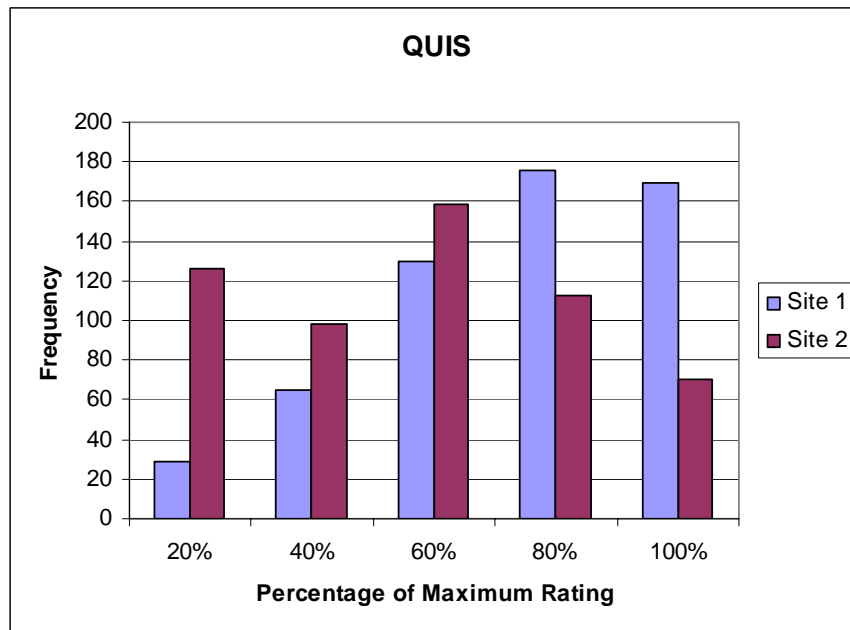
Figure 2. Results using QUIS.

A Comparison of Questionnaires for Assessing Website Usability
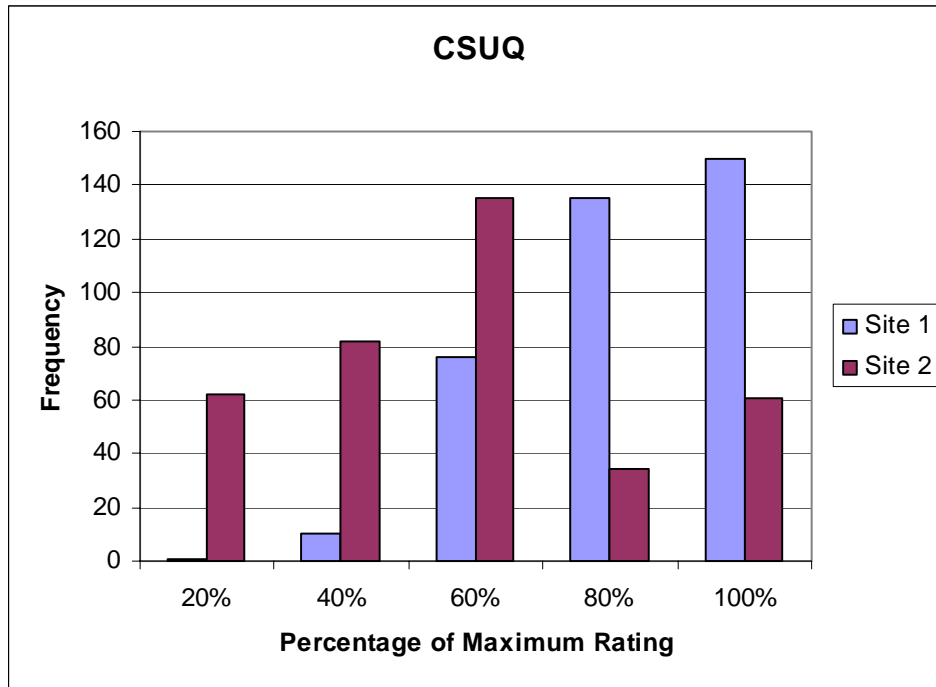
CSUQ

Figure 3. Results using CSUQ.

Survey 4: Words
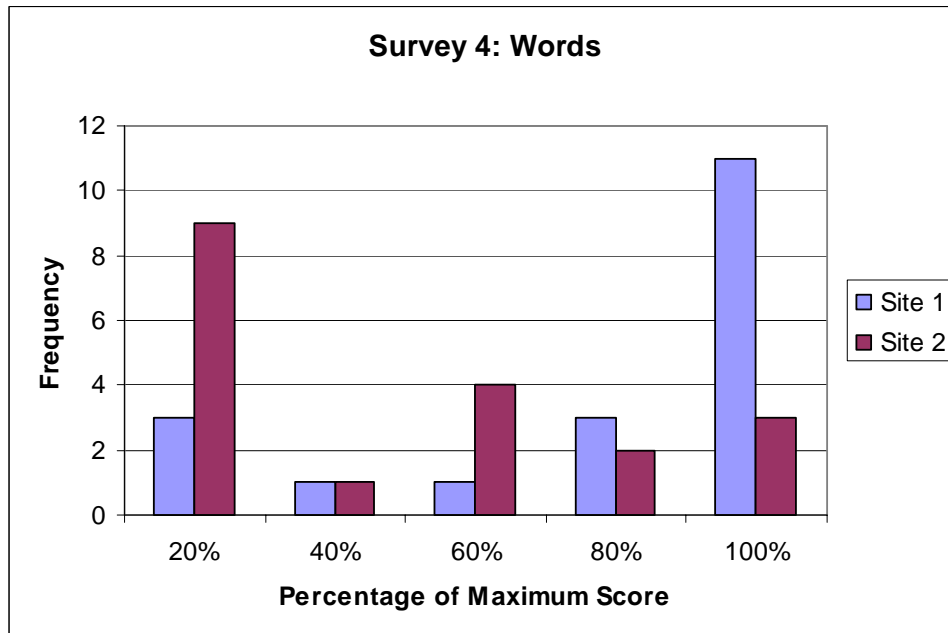
Figure 4. Results using Microsoft's Words

A Comparison of Questionnaires for Assessing Website Usability
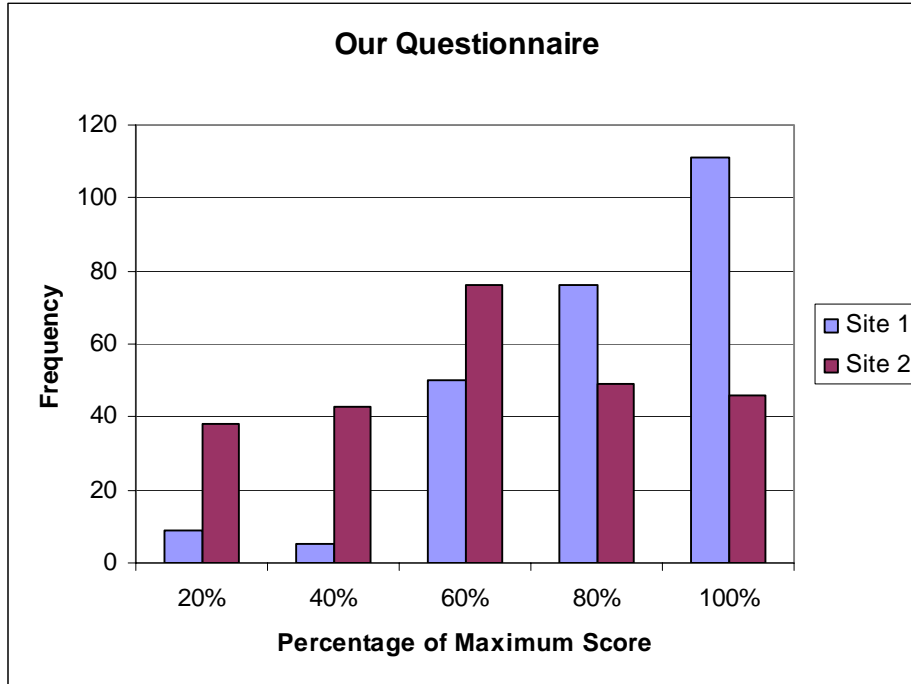
Our Questionnaire

Figure 5. Results using our questionnaire.

Comparison of Means
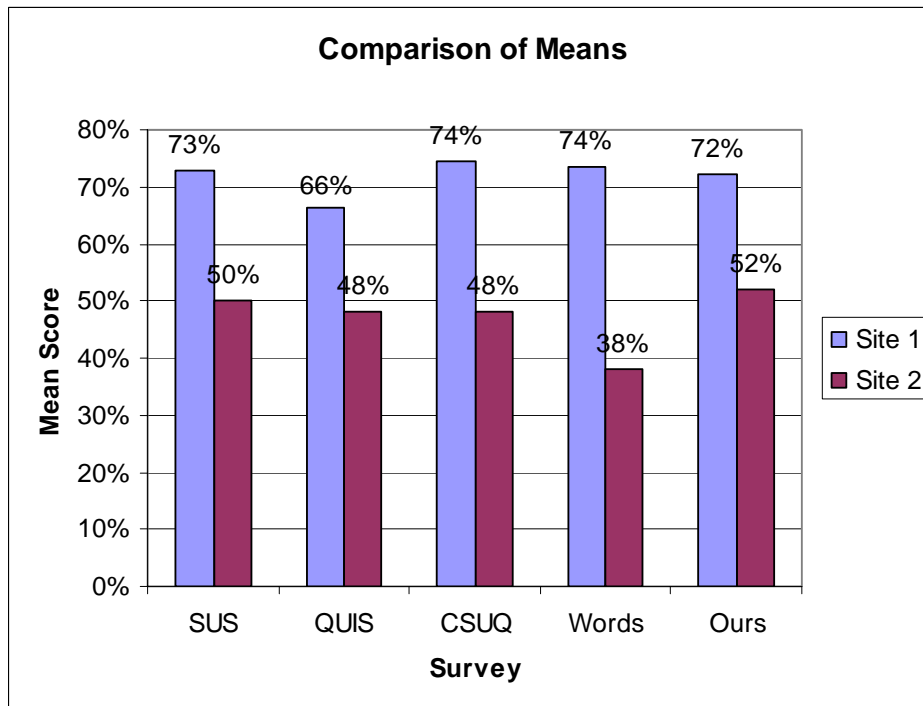
Figure 6. Comparison of mean scores for each site using each questionnaire.

All five questionnaires showed that Site 1 was significantly preferred over Site 2 (*p*<.01 via t-test for each). The largest mean difference (74% vs. 38%) was found using the Words questionnaire, but this was also the questionnaire that yielded the greatest variability in the responses. Both of these points are apparent from examination of Figure 4, where you can

A Comparison of Questionnaires for Assessing Website Usability

see that the modal values for the two sites are at the opposite ends of the scale, but there are some responses for both sites across the entire range.

The most interesting thing to look at now is what the results would have been using each questionnaire if the study had been done with a smaller number of participants. We chose to analyze randomly selected sub-samples of the data at size 6, 8, 10, 12, and 14. We felt these samples represented sizes commonly used in usability tests. This was an empirical sub-sampling in which 20 random sub-samples were taken from the full dataset at each of these different sample sizes, and a t-test was conducted to determine whether the results showed that Site 1 was significantly better than Site 2 (the conclusion from the full dataset). Figure 7 shows the results of this random sub-sampling.
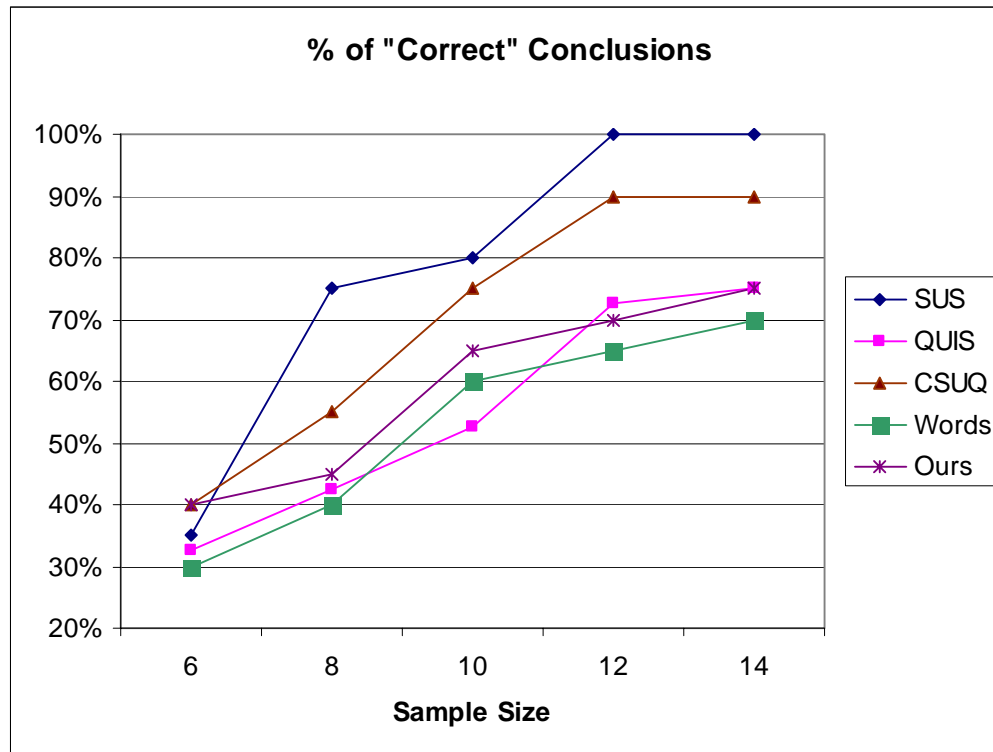
Figure 7. Data based on t-tests of random sub-samples of various sizes. Twenty sub-samples were taken at each sample size for each site and each questionnaire. What is plotted is the percentage of those 20 tests that yielded the same conclusion as the analysis of the full dataset (that Site 1 was significantly preferred over Site 2).

As one would expect, the accuracy of the analysis increases as the sample size gets larger. With a sample size of only 6, all of the questionnaires yield accuracy of only 30-40%, meaning that 60-70% of the time, at that sample size, you would fail to find a significant difference between the two sites. Interestingly, the accuracy of some of the questionnaires increases quicker than others. For example, SUS jumps up to about 75% accuracy at a sample size of 8, while the others stay down in the 40-55% range. It's also interesting to note that most of the questionnaires appear to reach an asymptote at a sample size of 12. The improvement by going to a sample size of 14 is small in most cases. Also, due to the different variances of the responses, some of the questionnaires reach a higher asymptote than others. For example, SUS and CSUQ reach asymptotes of 90-100% while the others are in the 70-75% range. Of course, the other questionnaires would have continued to yield improvement if larger samples had been tested.

Conclusions

First, some caveats need to be pointed out about the interpretation of these data. The primary one is that they really only directly apply to the analysis of the two sites that we studied. We selected two popular sites that provide financial information, finance.Yahoo.com and Kiplinger.com. We chose these sites because they provide similar kinds of information but in different ways. Had the two sites studied been even more similar to each other, it would have been more difficult for any of the questionnaires to yield a significant difference. Likewise, if they had been more different, it would have been easier for any of the questionnaires to yield a significant difference.

Another caveat is that the users' assessments of these sites were undoubtedly affected by the two tasks that we asked them to do on those sites. Again, we did not choose tasks that we thought would be particularly easier or more difficult on one site vs. the other. We chose tasks that we thought were typical of the tasks people might want to do on these kinds of sites.

It's also possible that the results could have been somewhat different if we had been able to collect data from more participants using each questionnaire. The minimum number of participants that we got for any one questionnaire was 19. Some researchers have argued that still larger numbers of participants are needed to get reliable data from some of these questionnaires. While that may be true, one of our goals was to study whether any of these questionnaires yield reliable results at the smaller sample sizes typically seen in usability tests.

Finally, this paper has only addressed the question of whether a given questionnaire was able to reliably distinguish between the ratings of one site vs. the other. In many usability tests, you have only one design that you are evaluating, not two or more that you are comparing. When evaluating only one design, possibly the most important information is related to the diagnostic value of the data you get from the questionnaire. In other words, how well does it help guide improvements to the design? That has not been analyzed in this study. Interestingly, on the surface at least, it appears that the Microsoft Words might provide the most diagnostic information, due to the potentially large number of descriptors involved.

Keeping all of those caveats in mind, it is interesting to note that one of the simplest questionnaires studied, SUS (with only 10 rating scales), yielded among the most reliable results across sample sizes. It is also interesting that SUS is the only questionnaire of those studied whose questions all address different aspects of the user's reaction to the website as a whole (e.g., "I found the website unnecessarily complex", "I felt very confident using the website") as opposed to asking the user to assess specific features of the website (e.g., visual appearance, organization of information, etc). These results also indicate that, for the conditions of this study, sample sizes of at least 12-14 participants are needed to get reasonably reliable results.

REFERENCES

1. Benedek, J., & Miner, T. (2002). Measuring desirability: New methods for evaluating desirability in a usability lab setting. Proceedings of UPA 2002 Conference, Orlando, FL, July 8-12, 2002.

2. Brooke, J. (1996). SUS: A Quick and Dirty Usability Scale. In: P.W. Jordan, B. Thomas, B.A. Weerdmeester & I.L. McClelland (Eds.), *Usability Evaluation in Industry*. London: Taylor & Francis. (Also see http://www.cee.hw.ac.uk/~ph/sus.html)

3. Chin, J. P., Diehl, V. A, & Norman, K. (1988). Development of an instrument measuring user satisfaction of the human-computer interface, Proceedings of ACM CHI '88 (Washington, DC), pp. 213-218. (Also see http://www.acm.org/~perlman/question.cgi?form=QUIS and http://www.lap.umd.edu/QUIS/index.html)

4. Lewis, J. (1995). IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. *International Journal of Human-Computer Interaction, 7 (1,)* 1995, 57-78. (Also see http://www.acm.org/~perlman/question.cgi?form=CSUQ)

5. Perlman, G. (Undated). Web-Based User Interface Evaluation with Questionnaires. Retrieved from http://www.acm.org/~perlman/question.html on Nov. 7, 2003.

6. WAMMI: http://www.wammi.com

7. RelevantView: http://www.relevantview.com/

8. NetRaker: http://www.netraker.com/

9. Vividence: http://www.vividence.com/

Appendix A: Screenshots of the Five Questionnaires Used

	Strongly Disagree				Strongly Agree
1. I think I would like to use this website frequently.	○	○	○	○	○
2. I found the website unnecessarily complex.	○	○	○	○	○
3. I thought the website was easy to use.	○	○	○	○	○
4. I think I would need Tech Support to be able to use this website.	○	○	○	○	○
5. I found the various functions in this website were well integrated.	○	○	○	○	○
6. I thought there was too much inconsistency in this website.	○	○	○	○	○
7. I would imagine that most people would learn to use this website very quickly.	○	○	○	○	○
8. I found the website very cumbersome to use.	○	○	○	○	○
9. I felt very confident using the website.	○	○	○	○	○
10. I need to learn a lot about this website before I could effectively use it.	○	○	○	○	○

SUS

A Comparison of Questionnaires for Assessing Website Usability

Overall Reaction to the Website		0	1	2	3	4	5	6	7	8	9		NA
1.	terrible	○	○	○	○	○	○	○	○	○	○	wonderful	○
2.	difficult	○	○	○	○	○	○	○	○	○	○	easy	○
3.	frustrating	○	○	○	○	○	○	○	○	○	○	satisfying	○
4.	dull	○	○	○	○	○	○	○	○	○	○	stimulating	○
5.	rigid	○	○	○	○	○	○	○	○	○	○	flexible	○

Web Page		0	1	2	3	4	5	6	7	8	9		NA
6. Reading characters on the page	hard	○	○	○	○	○	○	○	○	○	○	easy	○
7. Organization of information	confusing	○	○	○	○	○	○	○	○	○	○	very clear	○
8. Sequence of pages	confusing	○	○	○	○	○	○	○	○	○	○	very clear	○

Terminology and Website Information		0	1	2	3	4	5	6	7	8	9		NA
9. Use of terms throughout website	inconsistent	○	○	○	○	○	○	○	○	○	○	consistent	○
10. Terminology is intuitive	never	○	○	○	○	○	○	○	○	○	○	always	○
11. Position of messages on screen	inconsistent	○	○	○	○	○	○	○	○	○	○	consistent	○
12. Prompts for input	confusing	○	○	○	○	○	○	○	○	○	○	clear	○
13. Website informs about its progress	never	○	○	○	○	○	○	○	○	○	○	always	○
14. Error messages	unhelpful	○	○	○	○	○	○	○	○	○	○	helpful	○

Learning		0	1	2	3	4	5	6	7	8	9		NA
15. Learning to use the website	difficult	○	○	○	○	○	○	○	○	○	○	easy	○
16. Exploring new features by trial and error	difficult	○	○	○	○	○	○	○	○	○	○	easy	○
17. Performing tasks is straightforward	never	○	○	○	○	○	○	○	○	○	○	always	○
18. Help messages on the screen	unhelpful	○	○	○	○	○	○	○	○	○	○	helpful	○
19. Supplemental reference materials	confusing	○	○	○	○	○	○	○	○	○	○	clear	○

Website Capabilities		0	1	2	3	4	5	6	7	8	9		NA
20. Website speed	too slow	○	○	○	○	○	○	○	○	○	○	fast enough	○
21. Website reliability	unreliable	○	○	○	○	○	○	○	○	○	○	reliable	○
22. Sounds associated with this website	detracts value	○	○	○	○	○	○	○	○	○	○	adds value	○
23. Correcting your mistakes	difficult	○	○	○	○	○	○	○	○	○	○	easy	○
24. Designed for all levels of users	never	○	○	○	○	○	○	○	○	○	○	always	○

QUIS

A Comparison of Questionnaires for Assessing Website Usability

Overall Reaction to the Website		1	2	3	4	5	6	7		NA
1. Overall, I am satisfied with how easy it is to use this website	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
2. It was simple to use this website	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
3. I can effectively complete my work using this website	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
4. I am able to complete my work quickly using this website	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
5. I am able to efficiently complete my work using this website	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
6. I feel comfortable using this website	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
7. It was easy to learn to use this website	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
8. I believe I became productive quickly using this website	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
9. The website gives error messages that clearly tell me how to fix problems	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
10. Whenever I make a mistake using the website, I recover easily and quickly	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
11. The information (such as online help, on-page messages, and other documentation) provided with this website is clear	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
12. It is easy to find the information I need	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
13. The information provided by the website is easy to understand	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
14. The information is effective in helping me complete the tasks and scenarios	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
15. The organization of information on the website pages is clear	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
16. The interface of this website is pleasant	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
17. I like using the interface of this website	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
18. This website has all the functions and capabilities I expect it to have	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
19. Overall, I am satisfied with this website	strongly disagree	○	○	○	○	○	○	○	strongly agree	○

CSUQ

A Comparison of Questionnaires for Assessing Website Usability

☐ Convenient	☐ Familiar	☐ Slow	☐ Cutting Edge	☐ Friendly
☐ Busy	☐ Straight forward	☐ Personal	☐ Confusing	☐ Stressful
☐ Fun	☐ Boring	☐ Innovative	☐ Helpful	☐ Simplistic
☐ Empowering	☐ Usable	☐ Old	☐ Complex	☐ Irrelevant
☐ Dated	☐ Dull	☐ Advanced	☐ Patronizing	☐ Meaningful
☐ Clear	☐ Flexible	☐ Sophisticated	☐ Effective	☐ Difficult
☐ Accessible	☐ Time saving	☐ Business-like	☐ Too Technical	☐ Intuitive
☐ Organized	☐ Calm	☐ Disruptive	☐ Fast	☐ Consistent
☐ Controllable	☐ Annoying	☐ Easy to use	☐ Met Expectations	☐ Powerful
☐ Novel	☐ Comfortable	☐ Approachable	☐ Collaborative	☐ Distracting
☐ Frustrating	☐ Attractive	☐ Efficient	☐ Ordinary	☐ High Quality
☐ Valuable	☐ Gets in the way	☐ Fragile	☐ Connected	☐ Overbearing
☐ Sterile	☐ Exciting	☐ Desirable	☐ Overwhelming	☐ Secure
☐ Predictable	☐ Intimidating	☐ Appealing	☐ Clean	☐ Understandable
☐ Low Maintenance	☐ Comprehensive	☐ Unrefined	☐ Unconventional	☐ Effortless
☐ Inviting	☐ Entertaining	☐ Confident	☐ Optimistic	☐ Enthusiastic
☐ Customizable	☐ Time consuming	☐ Stable	☐ Unattractive	☐ Stimulating
☐ Not valuable	☐ Relevant	☐ Inconsistent	☐ Compelling	☐ Reliable
☐ Professional	☐ Compatible	☐ Incomprehensible	☐ Useful	☐ Undesirable
☐ Not secure	☐ Trustworthy	☐ Integrated	☐ Disconnected	☐ Energetic
☐ Engaging	☐ Motivating	☐ Poor quality	☐ Impressive	☐ Unpredictable
☐ Hard to use	☐ Ineffective	☐ Creative	☐ Uncontrollable	☐ Fresh
☐ Responsive	☐ Essential	☐ Satisfying	☐ Exceptional	
☐ Impersonal	☐ Inspiring	☐ Unapproachable	☐ Rigid	

Words (based on Microsoft's Product Reaction Cards)

	Strongly Disagree	-3	-2	-1	0	1	2	3	Strongly Agree
1.This website is visually appealing.		○	○	○	○	○	○	○	
2. It was easy to move from one page to another.		○	○	○	○	○	○	○	
3. The overall organization of the site is easy to understand.		○	○	○	○	○	○	○	
4. Individual pages are well designed.		○	○	○	○	○	○	○	
5. Terminology used in this website is clear.		○	○	○	○	○	○	○	
6. The content of the website met my expectations.		○	○	○	○	○	○	○	
7. I would be likely to use this website in the future.		○	○	○	○	○	○	○	
8. I was able to complete my tasks in a reasonable amount of time.		○	○	○	○	○	○	○	
9. Overall, the website is easy to use.		○	○	○	○	○	○	○	

Our Questionnaire