

Self-Reported Metrics

6

Perhaps the most obvious way to learn about the usability of something is to ask users to tell you about their experience with it. But exactly how to ask them so that you get good data is not so obvious. The questions you might ask could take on many forms, including various kinds of rating scales, lists of attributes that the participants could choose from, and open-ended questions like “List the top three things you liked the most about this application.” Some of the attributes you might ask about include overall satisfaction, ease of use, effectiveness of navigation, awareness of certain features, clarity of terminology, visual appeal, and many others. But the common feature of all of these is that you’re asking the participant for information, which is why we think *self-reported metrics* is the best term to use.

Two other terms sometimes used to describe this kind of data include *subjective data* and *preference data*. *Subjective* is used as a counterpart to *objective*, which is often used to describe the performance data from a usability study. But this implies that there’s a lack of objectivity to the data you’re collecting. Yes, it may be subjective to each participant who’s providing the input, but from the perspective of the usability specialist, it is completely objective. Similarly, *preference* is often used as a counterpart to *performance*. Although there’s nothing obviously wrong with that, we believe that preference implies a choice of one option over another, which is often not the case in usability studies.

6.1 IMPORTANCE OF SELF-REPORTED DATA

Self-reported data give you the most important information about users’ *perception* of the system and their interaction with it. At an emotional level, the data may even tell you something about how the users *feel* about the system. In many situations, these kinds of reactions are the main thing that you care about. Even if it takes users forever to perform something with a system, if the experience makes them happy, that may be the only thing that matters.

Your goal is to make the users think of your product first. For example, when deciding what travel-planning website to use for an upcoming vacation, users are more likely to think of the site that they liked the last time they used it. They're much less likely to remember how long the process was or that it took more mouse clicks than it should have. That's why users' subjective reactions to a website, product, or store may be the best predictor of their likelihood to return or make a purchase in the future.

6.2 COLLECTING SELF-REPORTED DATA

The most efficient way to capture self-reported data in a usability test is with some type of rating scale. Open-ended questions can also be very useful, but they are harder to analyze. Two of the classic rating scales are a Likert scale and a semantic differential scale.

6.2.1 Likert Scales

A typical item in a Likert scale is a statement to which the respondents rate their level of agreement. The statement may be positive (e.g., "The terminology used in this interface is clear") or negative (e.g., "I found the navigation options confusing"). Usually a 5-point scale of agreement like the following is used:

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

In the original version of the scale, Likert (1932) provided "anchor terms" for each point on the scale. Some people prefer to use a 7-point scale, but it gets a bit more difficult to come up with descriptive terms for each point as you get to higher numbers. This is one reason many researchers have dropped the intervening labels and just label the two ends (or anchor points). Many variations on Likert scales are still used today, but most Likert-scale purists would say that the two main characteristics of a Likert scale are (1) it expresses degree of agreement with a statement, and (2) it uses an odd number of response options, thus allowing a neutral response.

In designing the statements for Likert scales, you need to be careful how you word them. In general, you should avoid adverbs like *very*, *extremely*, or *absolutely* in the statements and use unmodified versions of adjectives. For example, the statement "This website is beautiful" may yield results that are quite different from "This website is absolutely beautiful," which may decrease the likelihood of strong agreement.

WHO WAS LIKERT?

Many people have heard of Likert scales when it comes to subjective rating scales, but not many know where the name came from or even how to pronounce it! It's pronounced "LICK-ert," not "LIKE-ert." This type of scale is named for Rensis Likert, who created it in 1932.

6.2.2 Semantic Differential Scales

The semantic differential technique involves presenting pairs of bipolar, or opposite, adjectives at either end of a series of scales, such as the following:

Weak	○ ○ ○ ○ ○ ○ ○	Strong
Beautiful	○ ○ ○ ○ ○ ○ ○	Ugly
Hot	○ ○ ○ ○ ○ ○ ○	Cold
Light	○ ○ ○ ○ ○ ○ ○	Dark

Like the Likert scale, a 5-point or 7-point scale is used. The difficult part about the semantic differential technique is coming up with words that are truly opposites. Sometimes a thesaurus can be helpful since it includes antonyms. But you need to be aware of the connotations of different pairings of words. For example, a pairing of "Friendly/Unfriendly" may have a somewhat different connotation and yield different results from "Friendly/Not Friendly" or "Friendly/Hostile."

OSGOOD'S SEMANTIC DIFFERENTIAL

The semantic differential technique was developed by Charles E. Osgood (Osgood et al., 1957), who designed it to measure the connotations of words or concepts. Using factor analysis of large sets of semantic differential data, he found three recurring attitudes that people used in assessing words and phrases: evaluation (e.g., "good/bad"), potency (e.g., "strong/weak"), and activity (e.g., "passive/active").

6.2.3 When to Collect Self-Reported Data

The two best times to collect self-reported data are at the end of each task (post-task ratings) and at the end of the entire session (post-study ratings). Post-study ratings tend to be the more common, but both have advantages. Quick ratings immediately after each task can help pinpoint tasks and parts of the interface that are particularly problematic. More in-depth ratings and open-ended questions at the end of the session can provide an effective overall evaluation after the participant has had a chance to interact with the product more fully.

6.2.4 How to Collect Self-Reported Data

Logistically, three techniques can be used to collect self-reported data in a usability test: answer questions or provide ratings orally, record responses on a paper form, or provide responses using some type of online tool. Each technique has its advantages and disadvantages. Having the participant provide responses orally is the easiest method from the participant's perspective, but, of course, it means that an observer needs to record the responses. This works best for a single, quick rating after each task.

Paper forms and online forms are suitable for both quick ratings and for longer surveys. Paper forms are generally easier to create than online forms, but they involve manual entry of the data, including the potential for errors in interpreting handwriting. Online forms are getting easier to create, as evidenced by the number of web-based questionnaire tools available, and participants are getting more accustomed to using them. One technique that works well is to have a laptop computer with the online questionnaire next to the participant's computer in the usability lab. The participant can then easily refer to the application or website while completing the online survey.

ONLINE SURVEY TOOLS

Many tools are available for creating and administering surveys via the web. Doing a search on "online survey tools" turns up a pretty extensive list. Some of them are SnapSurveys.com, SurveyGizmo.com, SurveyMonkey.com, SurveyShare.com, and Zoomerang.com.

Most of these tools support a variety of question types, including rating scales, check boxes, drop-down lists, grids, and open-ended questions. These tools generally have some type of free trial or other limited-functionality subscription that lets you try out the service for free. They provide mechanisms for authoring the surveys online, administering them, and analyzing the results. Most of the tools allow you to do some analyses of the data online, and some also provide the option of downloading data to Excel, although that may only be available with paid subscriptions.

6.2.5 Biases in Collecting Self-Reported Data

Some studies have shown that people who are asked directly for self-reported data, either in person or over the phone, provide more positive feedback than when asked through an anonymous web survey (e.g., Dillman et al., 2001). This is called the social desirability bias (Nancarrow & Brace, 2000), in which respondents tend to give answers they believe will make them look better in the eyes of others or not disappoint the evaluator. For example, people who are called on the phone and asked to evaluate their satisfaction with a product typically report much higher satisfaction than if they reported their satisfaction levels in a more anonymous way. Telephone respondents or participants in a usability lab essentially want to tell us

what they think we want to hear, and that is usually positive feedback about our product.

Therefore, we suggest collecting post-test data in such a way that the moderator or facilitator does not see the responses until after the participant has left. This might mean either turning away or leaving the room when the participant fills out the automated or paper survey. Making the survey itself anonymous may also elicit more honest reactions. Some usability researchers have suggested asking participants in a usability test to complete a post-test survey after they get back to their office or home and have received their incentive. This can be done by giving them a paper survey and a postage-paid envelope to mail it back or by e-mailing a pointer to an online survey. The main drawback of this approach is that you will typically have some dropoff in terms of who completes the survey.

6.2.6 General Guidelines for Rating Scales

When crafting your own rating scales to assess a specific attribute such as visual appeal, credibility, or responsiveness, the main thing to remember is that you will probably get more reliable data if you can think of a few different ways to ask participants to assess the attribute. In analyzing the results, you would average those responses together to arrive at the participant's overall reaction for that attribute. Likewise, the success of questionnaires that include both positive and negative statements to which participants respond would suggest the value of including both types of statements.

Finally, there's the issue of the number of scale values to use in any rating scales. This topic can be a source of heated debate among usability professionals. Most of the arguments center on the use of an even or odd number of points on the scale. An odd number of points has a center, or neutral, point, whereas an even number does not, thus forcing the user slightly toward one end or the other on the scale. This question is mainly an issue when you have a relatively small number of scale values, such as five or six. We believe that in most real-world situations a neutral reaction to something is perfectly valid and should be allowed on a rating scale. So in most cases we use rating scales with an odd number of points—usually five or seven.

6.2.7 Analyzing Self-Reported Data

One common technique for analyzing data from rating scales is to assign a numeric value to each of the scale positions and then compute the averages. For example, in the case of a 5-point Likert scale, you might assign a value of 1 to the “Strongly Disagree” end of the scale and a value of 5 to the “Strongly Agree” end. These averages can then be compared across different tasks, studies, user groups, and so on. This is common practice among most usability professionals as well as market researchers. Even though rating-scale data is not technically interval data, many professionals treat it as such. For example, we assume that the distance between a 1 and a 2 on a Likert scale is the same as the distance between a 2 and a 3 on the

same scale. This assumption is called *degrees of intervalness*. We also assume that a value *between* any two of the scale positions has meaning. The bottom line is that it is close enough to interval data that we can treat it as such.

Another common way to analyze rating-scale data is by looking at top-2- and bottom-2-boxes. A top-2-box score refers to someone choosing a 4 or 5 (on a 5-point scale), or a 6 or 7 (on a 7-point scale). The top-2 are those who agree with the statement (somewhat or strongly agree), and the bottom-2 are those who disagree with the statement (somewhat or strongly disagree). Keep in mind that when you convert to a top-2- or bottom-2-box, the data can no longer be considered interval. Therefore, you should just report the data as frequencies (i.e., the percentage of participants who are top-2-box). Sometimes it's helpful to focus just on top-2-box—for example, when you are only interested in those who really like something. However, sometimes it is useful to also look at participants who strongly disagree with something. Depending on your situation, you might want to present just top-2-box or both top-2- and bottom-2-box scores.

Summarizing responses from open-ended questions is always a challenge. We've never been able to come up with a magic solution to doing this quickly and easily. Part of the solution is to be relatively specific in your open-ended questions. For example, a question that asks participants to describe anything they found confusing about the interface is going to be easier to analyze than a general “comments” field. One technique that's sometimes helpful is to bring all the responses to an open-ended question into Word or Excel and then sort them in alphabetical order (having removed articles like *the*, *a*, and so on, from the beginning of each). This groups together some of the similar comments, but someone still needs to examine those comments and do a logical grouping that takes into account the various ways participants might have expressed similar reactions. Also, having more than one rater perform the groupings helps increase reliability.

6.3 POST-TASK RATINGS

The main goal of ratings associated with each task is to give you some insight into which tasks the participants thought were the most difficult. This can then point you toward parts of the system or aspects of the product that need improvement. One way to capture this information is to ask the participant to rate each task on one or more scales. The next few sections examine some of the specific techniques that have been used.

6.3.1 Ease of Use

Probably the most common self-reported metric is to ask users to rate how easy or how difficult each task was. This typically involves asking them to rate the task using a 5-point or 7-point scale. Some usability professionals prefer to use a traditional

Likert scale, such as “This task was easy to complete” (1 = Strongly Disagree, 3 = Neither Agree nor Disagree, 5 = Strongly Agree). Others prefer to use a semantic differential technique with anchor terms like “Easy/Difficult.” Either technique will provide you with a crude measure of perceived usability on a task level.

6.3.2 After-Scenario Questionnaire

Jim Lewis (1991) developed a set of three rating scales—the After-Scenario Questionnaire (ASQ)—designed to be used after the user completes a set of related tasks or a scenario:

1. “I am satisfied with the ease of completing the tasks in this scenario.”
2. “I am satisfied with the amount of time it took to complete the tasks in this scenario.”
3. “I am satisfied with the support information (online help, messages, documentation) when completing the tasks.”

Each of these statements is accompanied by a 7-point rating scale of “strongly disagree” to “strongly agree,” as shown in Figure 6.1. You can report the averages or top-2-/bottom-2- box score for each question or aggregated for each task. Note that each of these questions in the ASQ touches on three fundamental areas of usability: effectiveness (question 1), efficiency (question 2), and satisfaction (all three questions).

6.3.3 Expectation Measure

Albert and Dixon (2003) proposed a different approach to assessing subjective reactions after each task. Specifically, they argued that the most important thing about each task is how easy or difficult it was *in comparison to* how easy or difficult the participant *thought* it was going to be. So before participants actually did any of the tasks, Albert and Dixon asked them to rate how easy/difficult they *expected* each of the tasks to be, based simply on their understanding of the tasks and the product.

Participants expect some tasks to be easier than others. For example, getting a current quote on a stock should be easier than rebalancing an entire portfolio. Then, after performing each task, participants were asked to rate how easy/difficult the task *actually was*. The “before” rating is called the *expectation* rating, and the “after” rating is called the *experience* rating. They used the same 7-point rating scales (1 = Very Difficult, 7 = Very Easy) for both ratings. For each task you can then calculate an average *expectation rating* and an average *experience rating*. You can then display these two scores for each task as a scatterplot, as shown in Figure 6.2.

		1	2	3	4	5	6	7	
1. Overall, I am satisfied with the ease of completing the tasks in this scenario ☐	strongly disagree	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	strongly agree
2. Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario ☐	strongly disagree	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	strongly agree
3. Overall, I am satisfied with the support information (online-line help, messages, documentation) when completing the tasks ☐	strongly disagree	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	strongly agree

FIGURE 6.1

The ASQ developed by Lewis (1991). This implementation is Gary Perlman's web interface, which can be found at <http://www.acm.org/perlman/question.cgi?form=ASQ>.

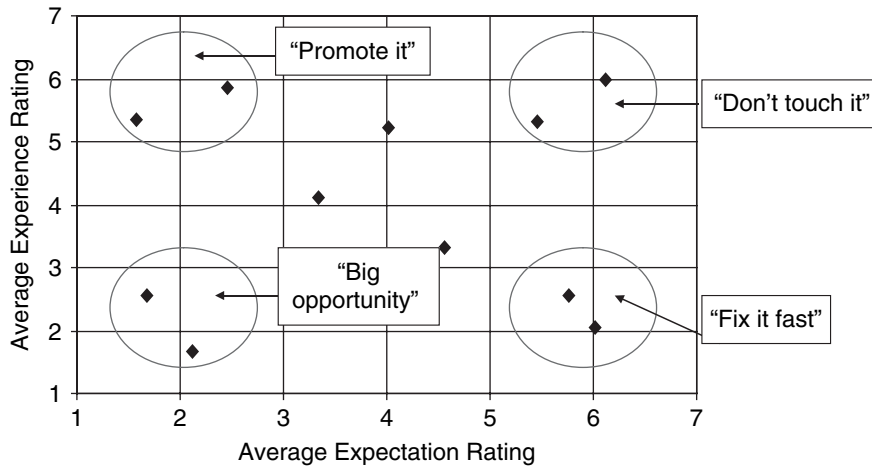


FIGURE 6.2

Comparison of the average expectation ratings and average experience ratings for a set of tasks in a usability test. Which quadrants the tasks fall into can help you prioritize which tasks to focus on improving. *Source:* Adapted from Albert and Dixon (2003); used with permission.

The four quadrants of the scatterplot provide some interesting insight into the tasks and where you should focus your attention when making improvements:

1. In the lower right are the tasks that the participants thought would be *easy* but actually turned out to be *difficult*. These probably represent the tasks that are the biggest dissatisfiers for the users—those that were the biggest disappointment. These are the tasks you should focus on first, which is why this is called the “Fix it fast” quadrant.
2. In the upper right are the tasks that the participants thought would be *easy* and actually *were* easy. These are working just fine. You don’t want to “break” them by making changes that would have a negative impact. That’s why this is called the “Don’t touch it” quadrant.
3. In the upper left are the tasks that the participants thought would be *difficult* and actually *were* *easy*. These are pleasant surprises, both for the users and the designers of the system! These could represent features of your site or system that may help distinguish you from the competition, which is why this is called the “Promote it” quadrant.
4. In the lower left are the tasks that the participants thought would be *difficult* and actually *were* difficult. There are no big surprises here, but there might be some important opportunities to make improvements. That’s why this is called the “Big opportunity” quadrant.

6.3.4 Usability Magnitude Estimation

A very different approach to task-based metrics has been proposed by Mick McGee (2004), who argued for a departure from the traditional rating-scale approach to self-reported measures. His method goes back to the magnitude estimation technique from classical psychophysics (Stevens, 1957). In traditional psychophysical studies, the experimenter would present a reference stimulus, such as a light source, and ask the subject to assign a value to some attribute of it, such as its brightness. Then, for example, a new light source would be shown, and the subject would be asked to assign a value to its brightness in comparison to the value from the reference light source. One of the keys to this method is that subjects are instructed to maintain a correspondence between the *ratios* of the numbers they assign and their perception of the magnitude. So, for example, a light that the subject perceives as being twice as bright as the reference light should get a value that's twice the reference value.

In adapting this method to usability studies, you may start by giving participants a reference “good design” and “bad design” for the same task. McGee (2004) used two examples illustrating good and bad versions of an interface for logging onto a website. In the good version, there were obvious fields for entering an ID and password and an explanation of what the site is about. This design was not presented as a *perfect* design, just a reasonably good one. The bad version had a long list of problems associated with it: poor contrast, confusing instructions, unlabeled input fields, confusing buttons, and the actual login on a separate page.

Next, you ask the participant to assign “usability values” to the reference good and bad designs. They can be any positive numbers. As illustrated in Figure 6.3, let's assume the participant gave a value of 20 to the reference bad design and a value of 200 to the reference good design. The participant is then asked to make judgments about the tasks she performs and the application's support for those tasks in comparison to these reference values. So, for example, if Task 1 was about three times better than the reference bad design, it would get a rating of 60. If Task 2 was half as good as

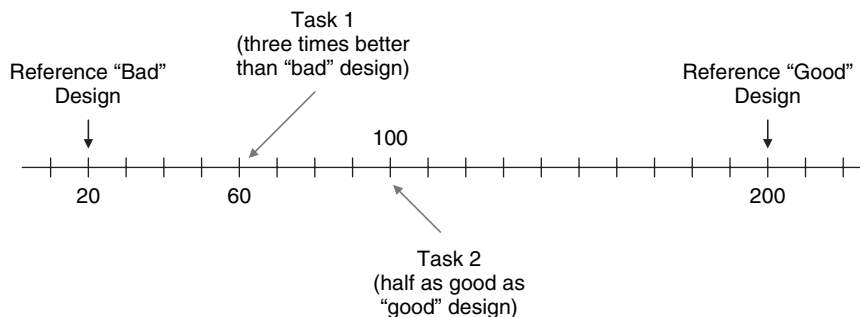


FIGURE 6.3

Example of a participant's "usability ruler" and how it might be used for assessing the usability of the system for various tasks.

the reference good design, it would get a rating of 100. In this way, each participant establishes his own “usability ruler” and uses it to express his perceptions.

Usability Magnitude Estimation can also be done without giving the participants reference good and bad designs. In this approach, participants simply start giving usability values for the tasks as they encounter them; in essence, they gradually build their own ruler as they encounter the tasks. In either approach, each participant’s data is transformed using geometric averaging (log transformations) to a consistent scale for comparison purposes (McGee, 2003).

6.3.5 Comparison of Post-Task Self-Reported Metrics

Tedesco and Tullis (2006) compared a variety of task-based self-reported metrics in an online usability study. Specifically, they tested the following five different methods for eliciting self-reported ratings after each task.

Condition 1: “Overall, this task was: Very Difficult ○ ○ ○ ○ Very Easy”

This was a very simple post-task rating scale that some usability teams commonly use.

Condition 2: “Please rate the usability of the site for this task: Very Difficult to Use ○ ○ ○ ○ Very Easy to Use”

Obviously, this is very similar to Condition 1 but with an emphasis on the usability of *the site* for the task. Perhaps only usability geeks detect the difference, but they wanted to find out!

Condition 3: “Overall, I am satisfied with the ease of completing this task: Strongly Disagree ○ ○ ○ ○ Strongly Agree”

“Overall, I’m satisfied with the amount of time it took to complete this task: Strongly Disagree ○ ○ ○ ○ Strongly Agree”

These are two of the three questions used in Lewis’s ASQ (1991). The third question in the ASQ asks about the support information such as online help, which was not relevant in this study, so it was not used.

Condition 4 (before doing all tasks): “How difficult or easy do you expect this task to be? Very Difficult ○ ○ ○ ○ Very Easy”

(after doing each task): “How difficult or easy did you find this task to be? Very Difficult ○ ○ ○ ○ Very Easy”

This is the expectation measure described by Albert and Dixon (2003).

Condition 5: “Please assign a number between 1 and 100 to represent how well the website supported you for this task. Remember: 1 would mean that the site was not at all supportive and completely unusable. A score of 100 would mean that the site was perfect and would require absolutely no improvement.”

This condition was originally based on Usability Magnitude Estimation but was significantly modified through iterations in the study planning. In pilot testing using a more traditional version of Usability Magnitude Estimation, they found that

participants had a very difficult time understanding the concepts and using the technique appropriately. As a result, they modified it to this simpler technique. This may mean that Usability Magnitude Estimation is better suited to use in a lab setting, or at least a moderated usability study, than in an online, unmoderated usability study.

The techniques were compared in an online study. The participants performed six tasks on a live application used to look up information about employees (phone number, location, manager, etc.). Each participant used only one of the five self-report techniques. A total of 1,131 employees participated in the online study, with at least 210 of them using each self-report technique.

The main goal of this study was to see whether these rating techniques are sensitive to differences in perceived difficulty of the tasks. But they also wanted to see how perceived difficulty corresponded to the task performance data. They collected task time and binary success data (i.e., whether the participants found the correct answer for each task and how long that took). As shown in Figure 6.4, there were significant differences in the performance data across the tasks. Task 2 appears to have been the most challenging, whereas Task 4 was the easiest.

As shown in Figure 6.5, a somewhat similar pattern of the tasks was reflected by the task ratings (averaged across all five techniques). In comparing task performance with the task ratings, correlations were significant for all five conditions ($p < 0.01$). Overall, Spearman rank correlation comparing the performance data and task ratings for the six tasks was significant: $R_s = 0.83$.

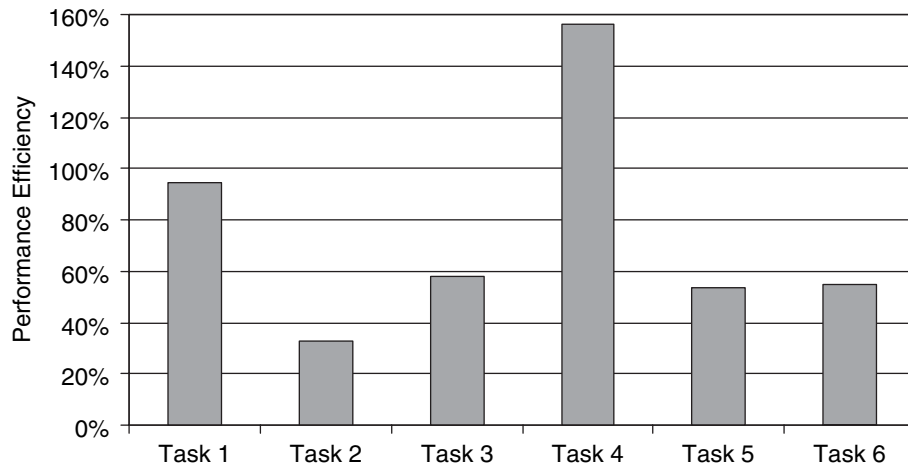
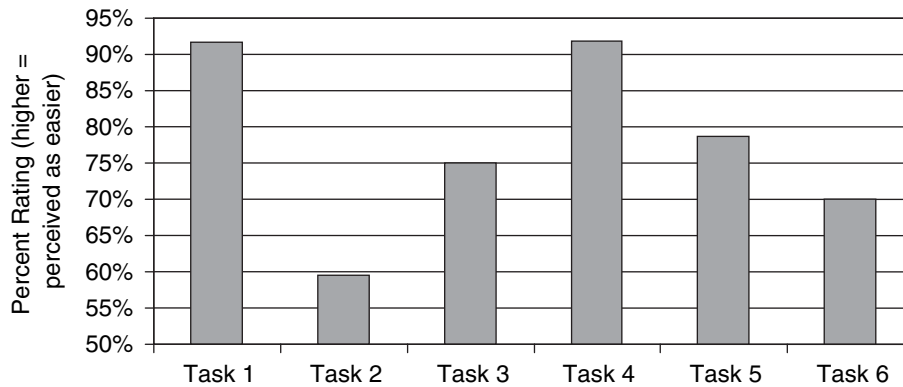


FIGURE 6.4

Performance data showing that participants had the most difficulty with Task 2 and the least difficulty with Task 4. *Source:* Adapted from Tedesco and Tullis (2006).

**FIGURE 6.5**

Average subjective ratings across all techniques. Ratings are expressed as a percentage of the maximum possible rating. Similar to the performance data, Task 2 yielded the worst ratings, while Task 4 yielded among the best. *Source:* Adapted from Tedesco and Tullis (2006).

Figure 6.6 shows the averages of the task ratings for each of the tasks, split out by condition. The key finding is that the pattern of the results was very similar regardless of which technique was used. This is not surprising, given the very large sample (total $n = 1,131$). In other words, at large sample sizes, all five of the techniques can effectively distinguish between the tasks.

But what about at the smaller sample sizes more typical of usability tests? To answer that question, they did a subsampling analysis looking at large numbers of random samples of different sizes taken from the full dataset. The results of this are shown in Figure 6.7, where the correlation between the data from the subsamples and the full dataset is shown for each subsample size.

The key finding was that one of the five conditions, Condition 1, resulted in better correlations starting at the smallest sample sizes and continuing. Even at a sample size of only seven, which is typical of many usability tests, its correlation with the full dataset averaged 0.91, which was significantly higher than any of the other conditions. So Condition 1, which was the simplest rating scale (“Overall, this task was Very Difficult ○ ○ ○ ○ Very Easy”), was also the most reliable at smaller sample sizes.

6.4 POST-SESSION RATINGS

One of the most common uses of self-reported metrics is as an overall measure of perceived usability that participants are asked to give after having completed their interactions with the product. This can be used as an overall “barometer” of the usability of the product, particularly if you establish a track record with the same

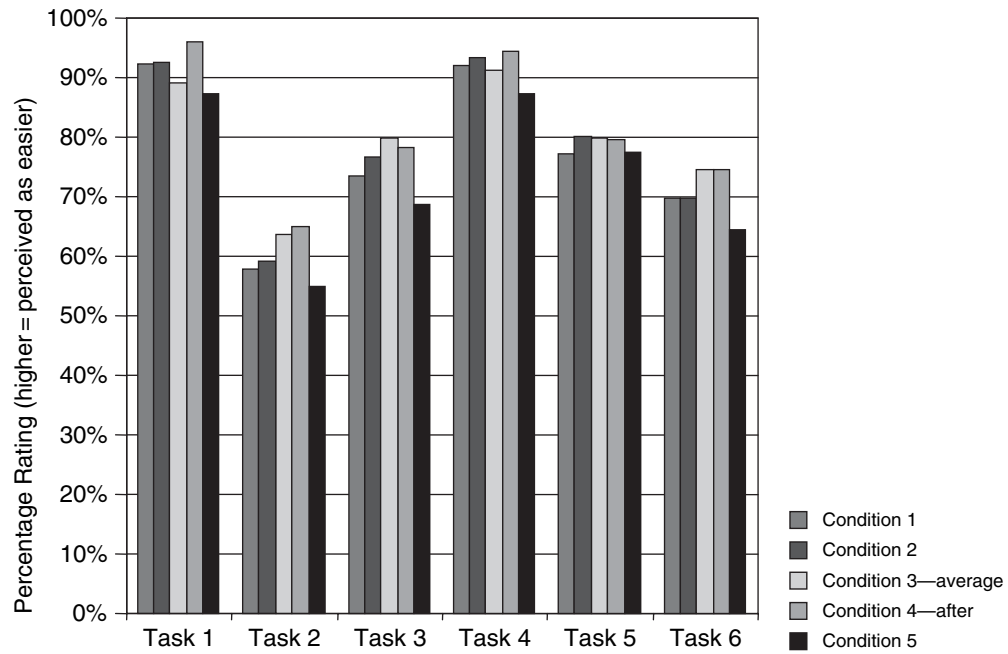
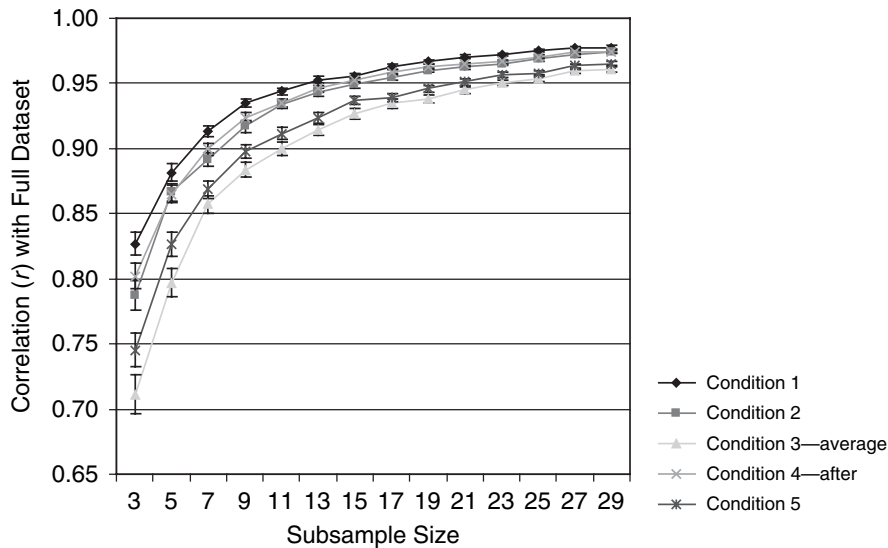


FIGURE 6.6

Average subjective ratings split by task and condition. All five conditions (self-report techniques) yielded essentially the same pattern of results for the six tasks. *Source:* Adapted from Tedesco and Tullis (2006).

**FIGURE 6.7**

Results of a subsampling analysis showing average correlations between ratings for the six tasks from subsamples of various sizes and the full dataset for each condition. Error bars represent the 95 percent confidence interval for the mean. *Source:* Adapted from Tedesco and Tullis (2006).

measurement technique over time. Similarly, these kinds of ratings can be used to compare multiple design alternatives in a single usability study or to compare your product, application, or website to the competition. Let's look at some of the post-session rating techniques that have been used.

6.4.1 Aggregating Individual Task Ratings

Perhaps the simplest way to look at post-session usability is to take an average of the self-reported data across all tasks. Of course, this assumes that you did in fact collect self-reported data after each task. If you did, then simply take an average of them. Keep in mind that these data are a little different from one snapshot at the end of the session. By looking at self-reported data across all tasks, you're really taking an average perception as it changes over time. Alternatively, when you collect the self-reported data just once at the end of the session, you are really measuring the participant's last impression of the experience.

This is the perception participants will leave with, which will likely influence any future decisions they make about your product. So if you want to measure perceived ease of use for the product based on individual task performance, then aggregate self-reported data from multiple tasks. However, if you're interested in

knowing the lasting usability perception, then we recommend using one of the following techniques that takes a single snapshot at the end of the session.

6.4.2 System Usability Scale

The System Usability Scale (SUS) was originally developed by John Brooke in 1986 while he was working at Digital Equipment Corporation (Brooke, 1996). As shown in Figure 6.8, it consists of ten statements to which participants rate their level of agreement. Half the statements are positively worded and half negatively worded. A 5-point scale of agreement is used for each. A technique for combining the ten ratings into an overall score (on a scale of 0 to 100) is also given. No attempt is made to assess different attributes of the system (e.g., usability, usefulness, etc.). In fact, the intent is that you should not look at the ratings for the ten statements individually but only look at the combined rating.

	Strongly disagree					Strongly agree	
1. I think that I would like to use this system frequently.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>		4
	1	2	3	4	5		
2. I found the system unnecessarily complex.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>		1
	1	2	3	4	5		
3. I thought the system was easy to use.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		1
	1	2	3	4	5		
4. I think I would need the support of a technical person to be able to use this system.	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		4
	1	2	3	4	5		
5. I found the various functions in this system were well integrated.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		1
	1	2	3	4	5		
6. I thought this system was too inconsistent.	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		2
	1	2	3	4	5		
7. I would imagine that most people would learn to use this system very quickly.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		1
	1	2	3	4	5		
8. I found the system very cumbersome to use.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>		1
	1	2	3	4	5		
9. I felt very confident using the system.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>		4
	1	2	3	4	5		
10. I needed to learn a lot of things before I could get going with this system.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		3
	1	2	3	4	5		
Total = 22							SUS Score = 22 × 2.5 = 55

FIGURE 6.8

SUS, developed by John Brooke at Digital Equipment Corporation, and an example of scoring it. *Source:* From Brooke (1996).

It's convenient to think of SUS scores as percentages, since they are on a scale of 0 to 100, with 100 representing a perfect score.

CALCULATING A SUS SCORE

To calculate a SUS score, first sum the score contributions from each item. Each item's score contribution will range from 0 to 4. For items 1, 3, 5, 7, and 9, the score contribution is the scale position minus 1. For items 2, 4, 6, 8, and 10, the contribution is 5 minus the scale position. Multiply the sum of the scores by 2.5 to obtain the overall SUS score. Consider the sample data in Figure 6.8. The sum of the values, using these rules, is 22. Multiply that by 2.5 to get the overall SUS score of 55 percent. SUS has been made freely available for use in usability studies, for both research purposes and industry use. The only prerequisite for its use is that any published report acknowledge the source of the measure.

6.4.3 Computer System Usability Questionnaire

Jim Lewis (1995), who developed the ASQ technique for post-task ratings, also developed the Computer System Usability Questionnaire (CSUQ) to do an overall assessment of a system at the end of a usability study. The CSUQ is very similar to Lewis's Post-Study System Usability Questionnaire (PSSUQ), with only minor changes in wording. PSSUQ was originally designed to be administered in person, whereas CSUQ was designed to be administered by mail or online. As shown in Figure 6.9, CSUQ consists of 19 statements to which the user rates agreement on a 7-point scale of "Strongly Disagree" to "Strongly Agree," plus N/A. Unlike SUS, all of the statements in CSUQ are worded positively. Factor analyses of a large number of CSUQ and PSSUQ responses have shown that the results may be viewed in four main categories: System Usefulness, Information Quality, Interface Quality, and Overall Satisfaction.

6.4.4 Questionnaire for User Interface Satisfaction

The Questionnaire for User Interface Satisfaction (QUIS) was developed by a team in the Human-Computer Interaction Laboratory (HCIL) at the University of Maryland (Chin, Diehl, & Norman, 1988). As shown in Figure 6.10, QUIS consists of 27 rating scales divided into five categories: Overall Reaction, Screen, Terminology/System Information, Learning, and System Capabilities. The ratings are on 10-point scales whose anchors change depending on the statement. The first six scales (assessing Overall Reaction) are polar opposites with no statements (e.g., Terrible/Wonderful, Difficult/Easy, Frustrating/Satisfying). QUIS can be licensed from the University of Maryland's Office of Technology Commercialization (<http://www.lap.umd.edu/QUIS/index.html>); it is also available in printed and web versions in multiple languages.

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

FIGURE 6.9

The CSUQ. *Source:* Adapted from the work of Lewis (1995); used with permission.

OVERALL REACTION TO THE SOFTWARE		0	1	2	3	4	5	6	7	8	9	NA
1. <input type="checkbox"/>	terrible	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	wonderful	<input type="radio"/>
2. <input type="checkbox"/>	difficult	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	easy	<input type="radio"/>
3. <input type="checkbox"/>	frustrating	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	satisfying	<input type="radio"/>
4. <input type="checkbox"/>	inadequate power	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	adequate power	<input type="radio"/>
5. <input type="checkbox"/>	dull	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	stimulating	<input type="radio"/>
6. <input type="checkbox"/>	rigid	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	flexible	<input type="radio"/>
SCREEN		0	1	2	3	4	5	6	7	8	9	NA
7. Reading characters on the screen <input type="checkbox"/>	hard	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	easy	<input type="radio"/>
8. Highlighting simplifies task <input type="checkbox"/>	not at all	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	very much	<input type="radio"/>
9. Organization of information <input type="checkbox"/>	confusing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	very clear	<input type="radio"/>
10. Sequence of screens <input type="checkbox"/>	confusing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	very clear	<input type="radio"/>
TERMINOLOGY AND SYSTEM INFORMATION		0	1	2	3	4	5	6	7	8	9	NA
11. Use of terms throughout system <input type="checkbox"/>	inconsistent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	consistent	<input type="radio"/>
12. Terminology related to task <input type="checkbox"/>	never	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	always	<input type="radio"/>
13. Position of messages on screen <input type="checkbox"/>	inconsistent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	consistent	<input type="radio"/>
14. Prompts for input <input type="checkbox"/>	confusing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	clear	<input type="radio"/>
15. Computer informs about its progress <input type="checkbox"/>	never	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	always	<input type="radio"/>
16. Error messages <input type="checkbox"/>	unhelpful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	helpful	<input type="radio"/>
LEARNING		0	1	2	3	4	5	6	7	8	9	NA
17. Learning to operate the system <input type="checkbox"/>	difficult	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	easy	<input type="radio"/>
18. Exploring new features by trial and error <input type="checkbox"/>	difficult	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	easy	<input type="radio"/>
19. Remembering names and use of commands <input type="checkbox"/>	difficult	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	easy	<input type="radio"/>
20. Performing tasks is straightforward <input type="checkbox"/>	never	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	always	<input type="radio"/>
21. Help messages on the screen <input type="checkbox"/>	unhelpful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	helpful	<input type="radio"/>
22. Supplemental reference materials <input type="checkbox"/>	confusing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	clear	<input type="radio"/>
SYSTEM CAPABILITIES		0	1	2	3	4	5	6	7	8	9	NA
23. System speed <input type="checkbox"/>	too slow	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	fast enough	<input type="radio"/>
24. System reliability <input type="checkbox"/>	unreliable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	reliable	<input type="radio"/>
25. System tends to be <input type="checkbox"/>	noisy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	quiet	<input type="radio"/>
26. Correcting your mistakes <input type="checkbox"/>	difficult	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	easy	<input type="radio"/>
27. Designed for all levels of users <input type="checkbox"/>	never	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	always	<input type="radio"/>
		0	1	2	3	4	5	6	7	8	9	NA

FIGURE 6.10

Questionnaire for User Interface Satisfaction. *Source:* Developed by the HCIL at the University of Maryland. Commercial use requires a license from the Office of Technology Commercialization at the University of Maryland. Used with permission.

GARY PERLMAN'S ONLINE QUESTIONNAIRES

Several of the questionnaires shown in this chapter, as well as a few others, are available for use online through a web interface created by Gary Perlman (<http://www.acm.org/perlman/question.html>). The questionnaires include QUIS, ASQ, and CSUQ. Options are provided for specifying which questionnaire to use, an e-mail address to submit results, and the name of the system being evaluated. These can be specified as parameters associated with the URL for the online questionnaire. So, for example, to specify the following:

- Name of System: MyPage
- Questionnaire: CSUQ
- Send Results to: me@gmail.com

the URL would be *<http://www.acm.org/perlman/question.cgi?system=MyPage&form=CSUQ&email=me@gmail.com>*.

By default, all rating scales also provide a mechanism for the user to enter comments. Once the user clicks on the Submit button, the data is e-mailed to the address specified, formatted in a name=value format with one name and value per line.

6.4.5 Usefulness, Satisfaction, and Ease of Use Questionnaire

Arnie Lund (2001) proposed the Usefulness, Satisfaction, and Ease of Use (USE) questionnaire, shown in Figure 6.11, which consists of 30 rating scales divided into four categories: Usefulness, Satisfaction, Ease of Use, and Ease of Learning. Each is a positive statement (e.g., “I would recommend it to a friend”), to which the user rates level of agreement on a 7-point Likert scale. In analyzing a large number of responses using this questionnaire, he found that 21 of the 30 scales (identified in Figure 6.11, on page 144) yielded the highest weights for each of the categories, indicating that they contributed most to the results.

6.4.6 Product Reaction Cards

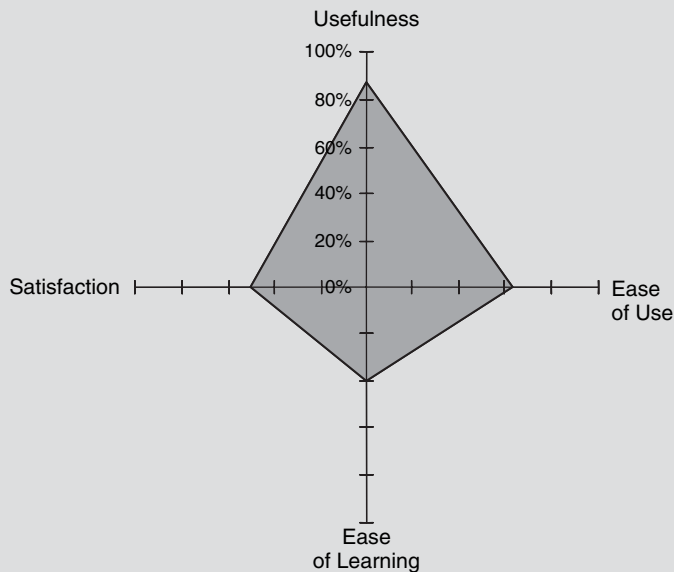
A very different approach to capturing post-session subjective reactions to a product was presented by Joey Benedek and Trish Miner (2002) from Microsoft. As illustrated in Figure 6.12 (on page 145), they presented a set of 118 cards containing adjectives (e.g., Fresh, Slow, Sophisticated, Inviting, Entertaining, Incomprehensible). Some of the words are positive and some are negative. The participants would then simply choose the cards they felt described the system. After selecting the cards, participants were asked to pick the top five cards and explain why they chose each. This technique is intended to be more qualitative in that its main purpose is to elicit commentary. But it can also be used in a somewhat quantitative way by counting the number of positive and negative terms chosen by each participant.

VISUALIZING DATA USING RADAR CHARTS

Some of the techniques for capturing self-reported data yield values on several dimensions. For example, the USE questionnaire can yield values for Usefulness, Satisfaction, Ease of Use, and Ease of Learning. Similarly, CSUQ can yield values for System Usefulness, Information Quality, Interface Quality, and Overall Satisfaction. One technique that can be useful for visualizing the results in a situation like this is a radar chart. Assume you have the following summary values from a study with the USE questionnaire:

- Usefulness = 90%
- Satisfaction = 50%
- Ease of Use = 45%
- Ease of Learning = 40%

Plotting these values as a radar chart would give you the chart shown here.



To create these charts, choose “Radar” as the “Chart Type” in Excel. “Filled” radar charts, like the example here, usually work best. The advantage these charts provide is that they help the viewer easily detect patterns as represented by different shapes. For example, a tall, skinny radar chart like the one shown here reflects the fact that users thought the product being evaluated was useful but not particularly easy to use, easy to learn, or satisfying.

Usefulness

- It helps me be more effective.
- It helps me be more productive.
- It is useful.
- It gives me more control over the activities in my life.
- It makes the things I want to accomplish easier to get done.
- It saves me time when I use it.
- *It meets my needs.*
- It does everything I would expect it to do.

Ease of Use

- It is easy to use.
- It is simple to use.
- It is user friendly.
- It requires the fewest steps possible to accomplish what I want to do with it.
- *It is flexible.*
- *Using it is effortless.*
- *I can use it without written instructions.*
- *I don't notice any inconsistencies as I use it.*
- *Both occasional and regular users would like it.*
- *I can recover from mistakes quickly and easily.*
- *I can use it successfully every time.*

Ease of Learning

- I learned to use it quickly.
- I easily remember how to use it.
- It is easy to learn to use it.
- *I quickly became skillful with it.*

Satisfaction

- I am satisfied with it.
- I would recommend it to a friend.
- It is fun to use.
- It works the way I want it to work.
- It is wonderful.
- I feel I need to have it.
- It is pleasant to use.

Users rate agreement with these statements on a 7-point Likert scale, ranging from strongly disagree to strongly agree. Statements in *italics* were found to weight less heavily than the others.

FIGURE 6.11

The USE questionnaire. *Source:* From the work of Lund (2001); used with permission.

6.4.7 Comparison of Post-Session Self-Reported Metrics

Tullis and Stetson (2004) reported a study that compared a variety of post-session questionnaires for measuring user reactions to websites in an online usability study. They studied the following questionnaires, adapted in the manner indicated for the evaluation of sites.

SUS: The word *system* in every question was replaced with *website*.

QUIS: Three of the original rating scales that did not seem to be appropriate to websites were dropped (e.g., “Remembering names and use of commands”). The term *system* was replaced with *website*, and the term *screen* was generally replaced by *web page*.

CSUQ: The term *system* or *computer system* was replaced by *website*.

Microsoft’s Product Reaction Cards: Each word was presented with a check box, and participants were 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.

The Complete Set of 118 Product Reaction Cards				
Accessible	Creative	Fast	Meaningful	Slow
Advanced	Customizable	Flexible	Motivating	Sophisticated
Annoying	Cutting edge	Fragile	Not secure	Stable
Appealing	Dated	Fresh	Not valuable	Sterile
Approachable	Desirable	Friendly	Novel	Stimulating
Attractive	Difficult	Frustrating	Old	Straightforward
Boring	Disconnected	Fun	Optimistic	Stressful
Business-like	Disruptive	Gets in the way	Ordinary	Time-consuming
Busy	Distracting	Hard to use	Organized	Time-saving
Calm	Dull	Helpful	Overbearing	Too technical
Clean	Easy to use	High quality	Overwhelming	Trustworthy
Clear	Effective	Impersonal	Patronizing	Unapproachable
Collaborative	Efficient	Impressive	Personal	Unattractive
Comfortable	Effortless	Incomprehensible	Poor quality	Uncontrollable
Compatible	Empowering	Inconsistent	Powerful	Unconventional
Compelling	Energetic	Ineffective	Predictable	Understandable
Complex	Engaging	Innovative	Professional	Undesirable
Comprehensive	Entertaining	Inspiring	Relevant	Unpredictable
Confident	Enthusiastic	Integrated	Reliable	Unrefined
Confusing	Essential	Intimidating	Responsive	Usable
Connected	Exceptional	Intuitive	Rigid	Useful
Consistent	Exciting	Inviting	Satisfying	Valuable
Controllable	Expected	Irrelevant	Secure	
Convenient	Familiar	Low maintenance	Simplistic	

FIGURE 6.12

The complete set of reaction cards developed by Joey Benedek and Trish Miner at Microsoft. *Source:* From Microsoft—"Permission is granted to use this Tool for personal, academic and commercial purposes. If you wish to use this Tool, or the results obtained from the use of this Tool for personal or academic purposes or in your commercial application, you are required to include the following attribution: Developed by and © 2002 Microsoft Corporation. All rights reserved."

Their Questionnaire: They had been using this questionnaire for several years in usability tests of websites. It was composed of nine positive statements (e.g., "This website is visually appealing"), to which the site's user responds on a 7-point Likert scale from "Strongly Disagree" to "Strongly Agree."

They used these questionnaires to evaluate two web portals in an online usability study. There were a total of 123 participants in the study, with each participant using one of the questionnaires to evaluate both websites.

Participants performed two tasks on each website before completing the questionnaire for that site. When the study authors analyzed the data from all the participants, they found that all five of the questionnaires revealed that Site 1 got significantly better ratings than Site 2. The data were then analyzed to determine what the results would have been at different sample sizes from 6 to 14, as shown in Figure 6.13. At a sample size of 6, only 30 to 40 percent of the samples would have identified that Site 1 was significantly preferred. But at a sample size of 8, which is relatively common in many lab-based usability tests, they found that SUS would have identified Site 1 as the preferred site 75 percent of the time—a significantly higher percentage than any of the other questionnaires.

It's interesting to speculate why SUS appears to yield more consistent ratings at relatively small sample sizes. One reason may be its use of both positive and negative statements with which participants must rate their level of agreement. This seems to keep participants more alert. Another reason may be that it doesn't try to break down the assessment into more detailed components (e.g., ease of learning, ease of navigation, etc.). All ten of the rating scales in SUS are simply asking for an assessment of the site as a whole, just in slightly different ways.

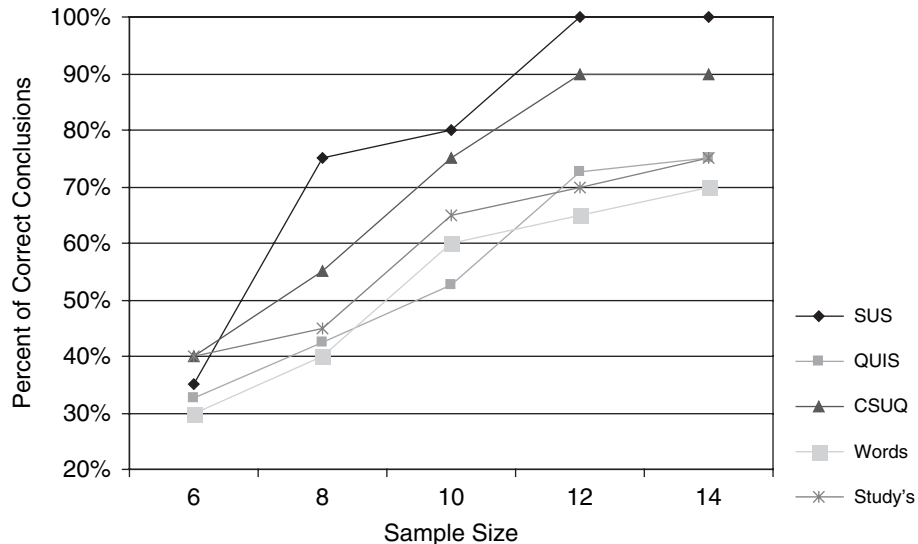


FIGURE 6.13

Data illustrating the accuracy of the results from random subsamples ranging from size 6 to size 14. This graph shows what percentage of the random samples yielded the same answer as the full dataset at the different sample sizes. *Source:* Adapted from Tullis and Stetson (2004).

6.5 USING SUS TO COMPARE DESIGNS

A number of usability studies that involved comparing different designs for accomplishing similar tasks have used the SUS questionnaire as one of the techniques for making the comparison (typically in addition to performance data).

6.5.1 Comparison of “Senior-Friendly” Websites

Traci Hart (2004) of the Software Usability Research Laboratory at Wichita State University conducted a usability study comparing three different websites designed for older adults: SeniorNet, SeniorResource, and Seniors-Place. After attempting tasks on each website, the participants rated each of them using the SUS questionnaire. The results are shown in Figure 6.14. The average SUS score for the SeniorResource site was 80 percent, which was significantly better than the average scores for SeniorNet and Seniors-Place, both of which averaged 63 percent.

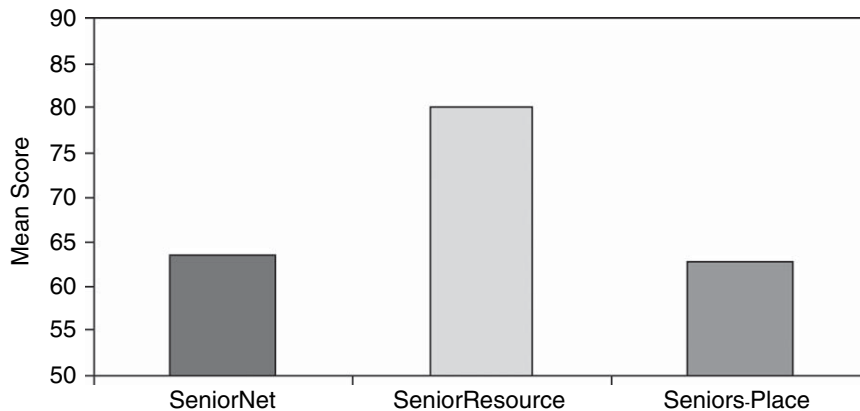


FIGURE 6.14

Data showing the average SUS scores from a study of three websites designed for older adults. The average SUS rating for SeniorResource was significantly higher than the average rating for either of the other sites. Participants were 21 adults over the age of 50, who performed tasks on all three sites and then rated all three. *Source:* Adapted from Hart (2004).

6.5.2 Comparison of Windows ME and Windows XP

The American Institutes for Research (2001) conducted a usability study comparing Microsoft’s Windows ME and Windows XP. They recruited 36 participants

whose expertise with Windows ranged from novice to intermediate. They attempted tasks using both versions of Windows and then completed the SUS questionnaire for both. They found that the average SUS score for Windows XP (74 percent) was significantly higher than the average for Windows ME (56 percent) ($p < 0.0001$).

6.5.3 Comparison of Paper Ballots

Sarah Everett, Michael Byrne, and Kristen Greene (2006), from Rice University, conducted a usability study comparing three different types of paper ballots: bubble, arrow, and open response. These ballots, which are illustrated in Figure 6.15,

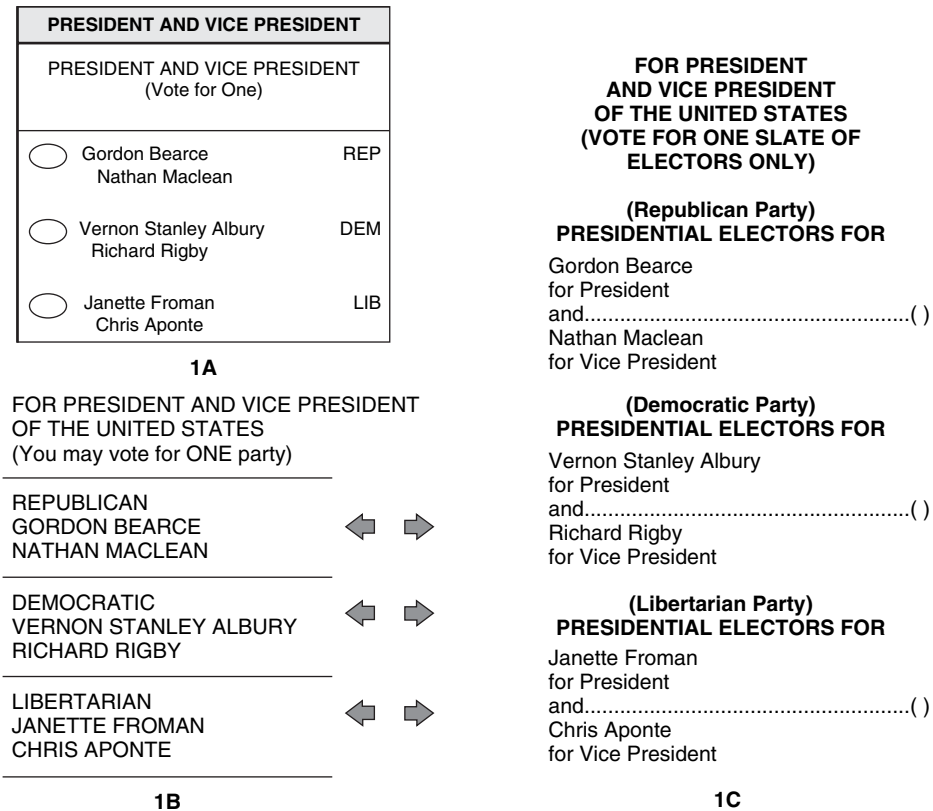
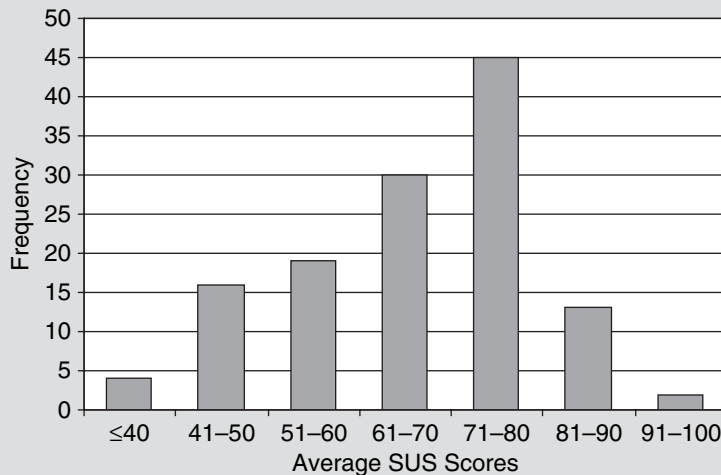


FIGURE 6.15
Three sample ballots. 1A: Bubble ballot; 1B: Arrow ballot; and 1C: Open-response ballot.
Source: From Everett et al. (2006); reprinted with permission from *Human Factors* 45(4).
Copyright © 2003 by the Human Factors and Ergonomics Society. All rights reserved.

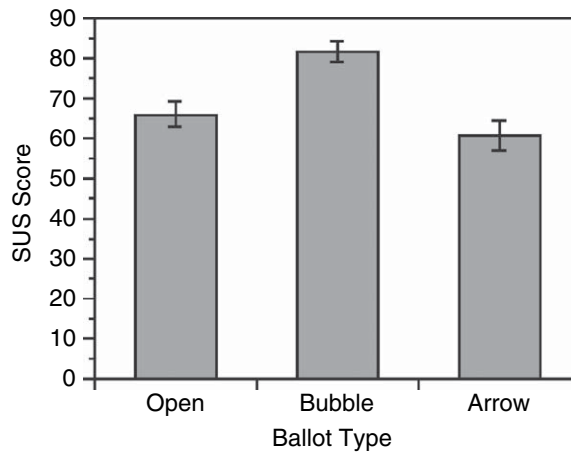
were based on actual ballots used in the 2004 U.S. elections. After using each of the ballots in a simulated election, the 42 participants used the SUS questionnaire to rate each one. The results are shown in Figure 6.16. They found that the bubble ballot (1A) received significantly higher SUS ratings than either of the other two ($p < 0.001$).

WHAT IS A GOOD (OR BAD) SUS SCORE?

After reporting an average of the post-session ratings from a particular study, usability professionals very often hear the question “So, is that good or bad?” It’s hard to answer that without having some points for comparison. So we did a review of a large number of published usability studies (using the ACM Digital Library and web search tools) and found 50 studies that reported average SUS scores across a total of 129 different conditions. The studies covered a wide range of subjects, including websites, applications, computer hardware, mobile devices, and voice systems. They were conducted in various parts of the world, including the United States, Germany, Switzerland, the United Kingdom, and New Zealand. The distribution of the average SUS scores from these 129 conditions is shown here.



Each of these studies had at least five participants; the maximum number was 81. The average SUS score from the 129 conditions was 66 percent and the median was 69 percent. The 25th percentile was 57 percent and the 75th percentile was 77 percent. (More details can be found on our website at www.MeasuringUserExperience.com.) These numbers would tend to suggest that you can think of an average SUS score under about 60 percent as relatively poor, while one over about 80 percent could be considered pretty good. But don’t forget to consider the confidence interval for any average SUS score that you calculate.

**FIGURE 6.16**

Average SUS ratings for three sample ballots. Error bars represent one standard error of the mean. *Source:* From Everett et al. (2006); reprinted with permission from *Human Factors* 45(4). Copyright © 2003 by the Human Factors and Ergonomics Society. All rights reserved.

6.6 ONLINE SERVICES

More and more companies are appreciating the value of getting feedback from the users of their websites. The currently in-vogue term for this process is listening to the “*Voice of the Customer*,” or VoC. This is essentially the same process as in post-session self-reported metrics. The main difference is that VoC studies are typically done on live websites. The common approach is that a randomly selected percentage of live-site users are offered a pop-up survey asking for their feedback at a specific point in their interaction with the site—usually on logout, exiting the site, or completing a transaction. Another approach is to provide a standard mechanism for getting this feedback at various places in the site. The following sections present some of these online services. This list is not intended to be exhaustive, but it is at least representative.

6.6.1 Website Analysis and Measurement Inventory

The Website Analysis and Measurement Inventory (WAMMI—www.wammi.com) is an online service that grew out of an earlier tool called Software Usability Measurement Inventory (SUMI), both of which were developed by the Human Factors Research Group (HFRG) of University College Cork in Ireland. Although SUMI is designed for evaluation of software applications, WAMMI is designed for evaluation of websites. Note that this same team is now developing a questionnaire called Measuring the Usability of Multi-Media Systems (MUMMS).

As shown in Figure 6.17, WAMMI is composed of 20 statements with associated 5-point Likert scales of agreement. Like SUS, some of the statements are positive and some are negative. WAMMI is available in most European languages. The primary advantage that a service like WAMMI has over creating your own questionnaire and associated rating scales is that WAMMI has already been used in the evaluation of hundreds of websites worldwide. When used on your site, the results are delivered in the form of a comparison against their reference database built from tests of these hundreds of sites.

Results from a WAMMI analysis, as illustrated in Figure 6.18, are divided into five areas: Attractiveness, Controllability, Efficiency, Helpfulness, and Learnability, plus an overall usability score. Each of these scores is standardized (from comparison to their reference database), so a score of 50 is average and 100 is perfect.

6.6.2 American Customer Satisfaction Index

The American Customer Satisfaction Index (ACSI—www.TheACSI.org) was developed at the Stephen M. Ross Business School of The University of Michigan. It covers a wide range of industries, including retail, automotive, and manufacturing. ForeSee Results (www.ForeSeeResults.com) applies the methodology of the ACSI to measure customer satisfaction with the online experience and produce industry-specific indices. The ACSI has become particularly popular for analyzing U.S. government websites. For example, 92 websites were included in the 2nd Quarter 2006 analysis of e-government websites (ForeSee Results, 2006). Similarly, ForeSee Results' annual Top 40 Online Retail Satisfaction Index assesses such popular sites as Amazon.com, Netflix, L.L. Bean, J.C. Penney, Dell, and CompUSA.

The ForeSee Results ACSI-based questionnaire for websites is composed of a core set of 14 to 20 questions (example shown in Figure 6.19 on page 154) customized to the function of the website (e.g., information, e-commerce, etc.). Each model question asks for a rating on a 10-point scale of different attributes of the web experience such as the quality of information, freshness of information, clarity of site organization, overall satisfaction with the site, and future behaviors (e.g., likelihood to return to the site or recommend the site). In addition, custom questions are added to the survey to profile site visitors in terms of visit intent, visit frequency, and other specific information that helps organizations profile visitors in terms meaningful to their business or mission.

As shown in Figure 6.20 (see page 155), which is an example of an informational website, the results for the website are divided into six quality-related elements that drive satisfaction—Content, Functionality, Look & Feel, Navigation, Search, and Site Performance—and overall satisfaction. The screenshot shows not only the ratings for satisfaction elements, but the relative impact of improving satisfaction for each element on increasing satisfaction overall. In addition, ratings are provided for two Future Behaviors—Likelihood to Return and Likelihood to Recommend the site to Others. All scores are on a 100-point index scale. This cause-and-effect


Statement 1-10 of 20Strongly
AgreeStrongly
Disagree

This web site has much that is of interest to me.



It is difficult to move around this web site.



I can quickly find what I want on this web site.



This web site seems logical to me.



This web site needs more introductory explanations.



The pages on this web site are very attractive.



I feel in control when I'm using this web site.



This web site is too slow.



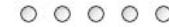
This web site helps me find what I am looking for.



Learning to find my way around this web site is a problem.

**Statement 11-20 of 20**Strongly
AgreeStrongly
Disagree

I don't like using this web site.



I can easily contact the people I want to on this web site.



I feel efficient when I'm using this web site.



It is difficult to tell if this web site has what I want.



Using this web site for the first time is easy.



This web site has some annoying features.



Remembering where I am on this web site is difficult.



Using this web site is a waste of time.



I get what I expect when I click on things on this web site.



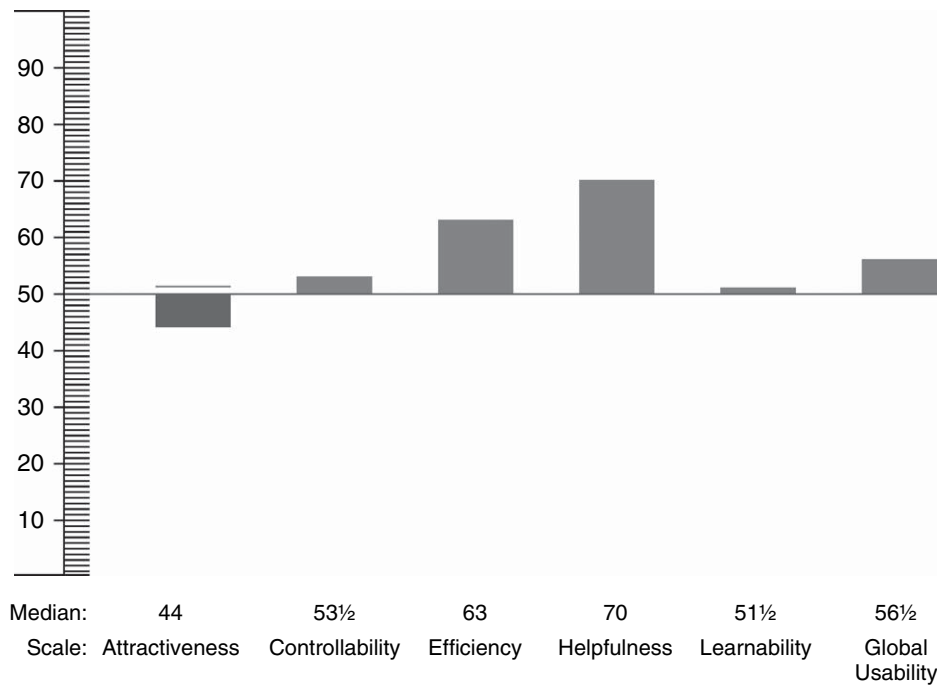
Everything on this web site is easy to understand.



Copyright © 2005 WAMMI

FIGURE 6.17

The 20 rating scales used by the WAMMI online service. *Source:* Reprinted with permission from Dr. J. Kirakowski, Human Factors Research Group, University College, Cork, Ireland.

**FIGURE 6.18**

Sample data from the WAMMI online service showing average scores in each of five categories, plus an overall usability score. Reprinted with permission from Dr. J. Kirakowski, Human Factors Research Group, University College, Cork, Ireland.


modeling shows the quantitative relationships between what drives satisfaction, overall satisfaction, and behaviors resulting from satisfaction that have a financial impact on the organization.

Finally, they also make assessments of the impact that each of the quality scores has on overall satisfaction. This allows you to view the results in four quadrants, as shown in Figure 6.21, plotting the quality scores on the vertical axis and the impact on overall satisfaction on the horizontal axis. The scores in the lower right quadrant (high impact, low score) indicate the areas where you should focus your improvements to get the maximum return on satisfaction and on investment.

6.6.3 OpinionLab

A somewhat different approach is taken by OpinionLab (www.OpinionLab.com), which provides for page-level feedback from users. In some ways, this can be thought of as a page-level analog of the task-level feedback discussed earlier. As

Customer Satisfaction Survey



Thank you for visiting our site. You have been randomly selected to take part in this survey to let us know what we are doing well and where we need to do better. Please take a minute or two to give us your opinions. The feedback you provide will help us enhance our site and serve you better in the future. All responses are strictly confidential.

1: Please rate the quality of information on this site.											
1=Poor	1	2	3	4	5	6	7	8	9	10=Excellent	Don't Know
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2: Please rate the freshness of content on this site.											
1=Poor	1	2	3	4	5	6	7	8	9	10=Excellent	Don't Know
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3: Please rate the convenience of the services on this site.											
1=Poor	1	2	3	4	5	6	7	8	9	10=Excellent	Don't Know
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4: Please rate the ability to accomplish what you wanted to on this site.											
1=Poor	1	2	3	4	5	6	7	8	9	10=Excellent	Don't Know
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
5: Please rate the clarity of site organization .											
1=Poor	1	2	3	4	5	6	7	8	9	10=Excellent	Don't Know
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
6: Please rate the clean layout of this site.											
1=Poor	1	2	3	4	5	6	7	8	9	10=Excellent	Don't Know
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
7: Please rate the ability to find information you want on this site.											
1=Poor	1	2	3	4	5	6	7	8	9	10=Excellent	Don't Know
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
8: Please rate the clarity of site map/directory .											
1=Poor	1	2	3	4	5	6	7	8	9	10=Excellent	Don't Know
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
9: Please rate the reliability of site performance on this site.											
1=Poor	1	2	3	4	5	6	7	8	9	10=Excellent	Don't Know
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
10: What is your overall satisfaction with this site?											
1=Poor	1	2	3	4	5	6	7	8	9	10=Excellent	
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
11: How well does this site meet your expectations ?											
1=Poor	1	2	3	4	5	6	7	8	9	10=Excellent	
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
12: How does this site compare to your idea of an ideal website ?											
1=Poor	1	2	3	4	5	6	7	8	9	10=Excellent	
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
13: How likely are you to return to this site ?											
1=Not Very Likely	1	2	3	4	5	6	7	8	9	10=Very Likely	
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
14: How likely are you to recommend this site to someone else ?											
1=Not Very Likely	1	2	3	4	5	6	7	8	9	10=Very Likely	
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
15: How frequently do you visit this site?											
	Please Select <input type="button" value="v"/>										
16: What would like to see improved on our site? (optional)											
	<input type="text"/>										

Thank you for taking the time to complete this survey. We value your input as we strive to continuously improve our site to serve you better.

FIGURE 6.19

Typical questions in an ACSI-based survey for a website. Used with permission.

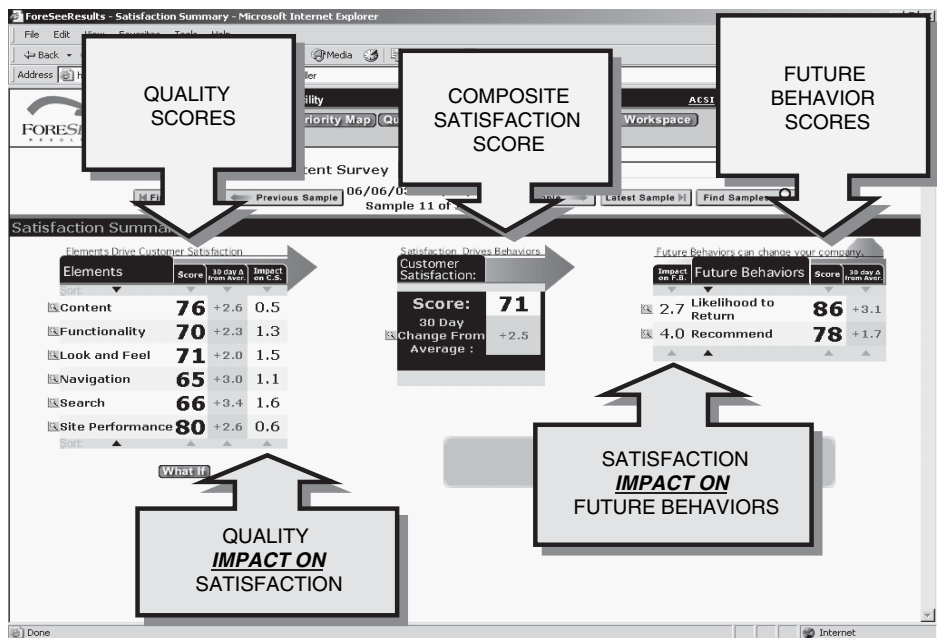


FIGURE 6.20 Sample results from a ForeSee Results analysis for a website: scores for six quality areas (*left*) along with values estimating the impact that each score has on overall customer satisfaction (*center*); scores for two "future behavior" areas (*right*) along with values estimating the satisfaction impact on those areas. *Source:* From ForeSee Results. Used with permission.

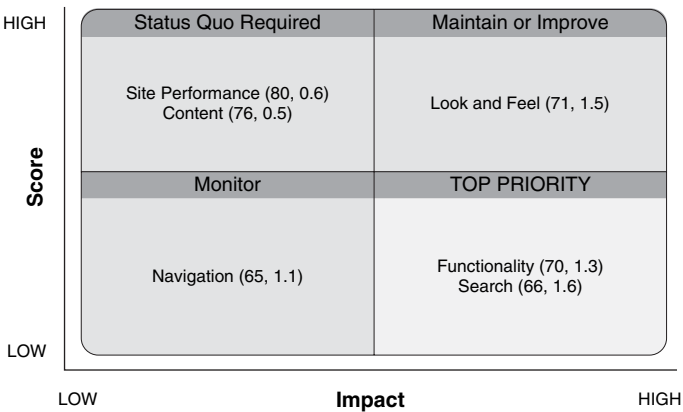


FIGURE 6.21 Sample results from a ForeSee Results analysis for a website. High and low scores for the six quality areas are represented on the vertical axis, and high and low impact scores are shown on the horizontal axis. The quality areas that fall in the lower right quadrant (Functionality and Search) should be top priorities for improvement. *Source:* From ForeSee Results. Used with permission.

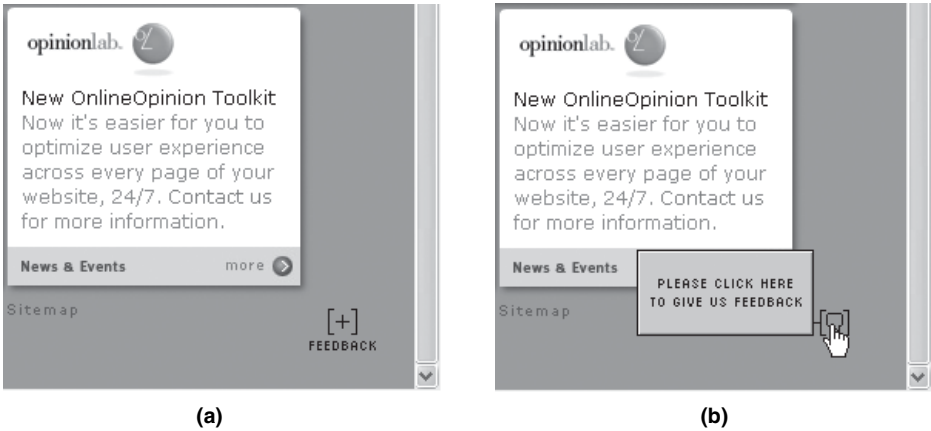


FIGURE 6.22

Web page (a) containing OpinionLab's feedback mechanism (*lower right*). This animated icon stays in that position while the user scrolls the page. Moving the mouse pointer over the icon reveals another version (b).



FIGURE 6.23

Examples of OpinionLab mechanisms for capturing feedback about a web page. The version on the left (a) allows the user to give the page a quick overall rating. The version on the right (b) allows for more detailed feedback on a few different scales.

shown in Figure 6.22, a common way for OpinionLab to allow for this page-level feedback is through a floating icon that always stays at the bottom right corner of the page regardless of the scroll position.

Clicking on that icon then leads to one of the methods shown in Figure 6.23 for capturing the feedback. The OpinionLab scales use five points that are marked simply as: --, -, +-, +, and ++. OpinionLab provides a variety of techniques for visualizing the data for a website, including the one shown in Figure 6.24, which

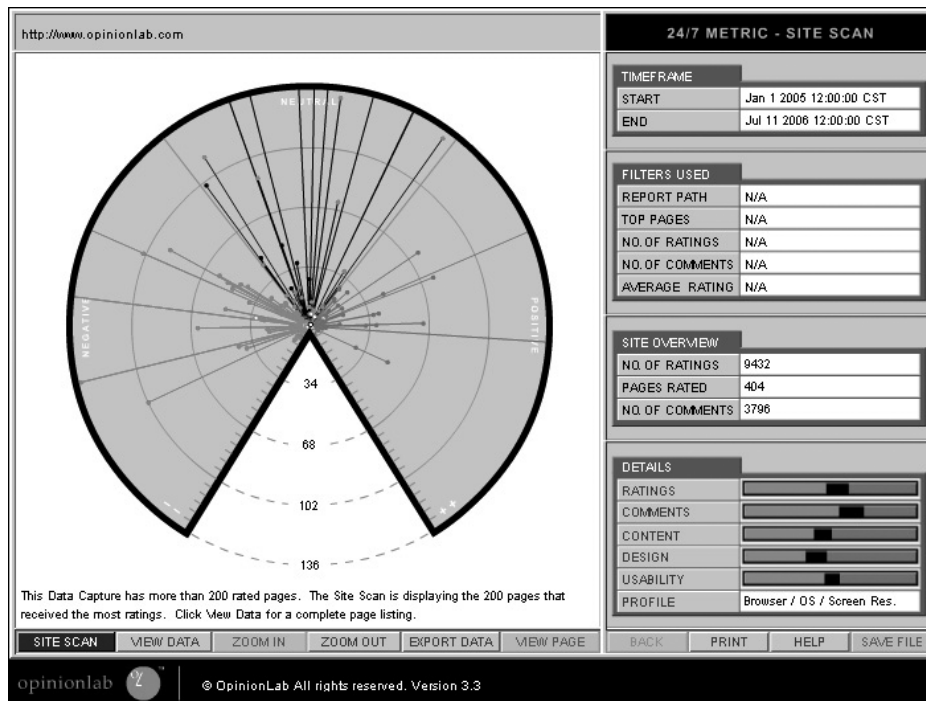


FIGURE 6.24

OpinionLab provides a variety of techniques for visualizing website data. On the left of the visualization shown here, the most-rated 200 pages are represented graphically. The pages receiving the most negative ratings are at the left, those with neutral ratings at the top, and those with the most positive ratings at the right.

allows you to easily spot the pages that are getting the most negative feedback and those that are getting the most positive feedback.

6.6.4 Issues with Live-Site Surveys

The following are some of the issues you will need to address when you use live-site surveys.

Number of questions. The fewer questions you have, the higher your response rate is likely to be. That's one reason that companies like OpinionLab keep the number of questions to a minimum. You need to strike a balance between getting the information you need and “scaring off” potential respondents. With every question you consider adding, ask yourself if you absolutely must have the information. Some researchers believe that 20 is about the maximum number of questions you should ask in this type of survey.

Self-selection of respondents. Because respondents make a decision about whether to complete the survey, they are self-selecting. You should at least ask yourself if this biases the responses in any way. Some researchers argue that people who are unhappy with the website are more likely to respond than those who are happy (or at least satisfied). If your main purpose is to uncover areas of the site to improve, that may not be a problem.

Number of respondents. Many of these services work on the basis of a percentage of visitors to offer the survey to. Depending on the amount of traffic your site gets, this percentage could be quite small and still generate a large number of responses. You should closely monitor responses to see if you need to increase or decrease the percentage.

Nonduplication of respondents. Most of these services provide a mechanism for noting (typically via a browser cookie) when the survey has already been offered to someone. As long as the user doesn't clear her cookies and is using the same computer, the survey won't be presented to her again for a specified time period. This prevents duplicate responses from an individual and also prevents annoying those users who don't want to respond.

6.7 OTHER TYPES OF SELF-REPORTED METRICS

Many of the self-report techniques described so far have sought to assess users reactions to products or websites as a whole or to tasks performed using them. But depending on a usability study's objectives, you might want to assess users reactions to specific product *attributes* overall or specific product *elements*.

6.7.1 Assessing Specific Attributes

Here are some of the attributes of a product or website that you might be interested in assessing:

- Visual appeal
- Perceived efficiency
- Usefulness
- Enjoyment
- Credibility
- Appropriateness of terminology
- Ease of navigation
- Responsiveness

Covering in detail the ways you might assess all the specific attributes you are interested in is beyond the scope of this book. Instead, we will describe a few examples of usability studies that have focused on assessing specific attributes.

Gitte Lindgaard and her associates at Carleton University in Ottawa, Ontario, were interested in learning how quickly users form an impression of the visual appeal of a web page (Lindgaard et al., 2006). They flashed images of web pages for either 50 msec or 500 msec to the participants in their study. Each web page was rated on an overall scale of visual appeal and on the following bipolar scales: Interesting/Boring, Good Design/Bad Design, Good Color/Bad Color, Good Layout/Bad Layout, and Imaginative/Unimaginative. They found that the ratings on all five of these scales correlated very strongly with visual appeal ($r^2 = 0.86$ to 0.92). They also found that the results were consistent across the participants at both the 50-msec and 500-msec exposure levels, indicating that even at 50 msec (or 1/20th of a second), users can form a consistent impression about the visual appeal of a web page.

Several years ago, we conducted an online study of ten different websites to learn more about what makes a website *engaging*. We defined an engaging website as one that (1) stimulates your interest and curiosity, (2) makes you want to explore the site further, and (3) makes you want to revisit the site. After exploring each site, the participants responded to a single rating worded as “This website is: Not At All Engaging . . . Highly Engaging” using a 5-point scale. The two sites that received the highest ratings on this scale are shown in Figure 6.25.

One of the techniques often used in analyzing the data from subjective rating scales is to focus on the responses that fall into the extremes of the scale: the top one or two or bottom one or two values. As mentioned earlier, these are often referred to as “top-2-box” or “bottom-2-box” scores. We recently used this technique in an online study assessing participants’ reactions to various load times for an intranet homepage. We artificially manipulated the load time over a range of 1 to 11 seconds. Different load times were presented in a random order, and the participants were never told what the load time was. After experiencing each load time, the participants were asked to rate that load time on a 5-point scale of “Completely Unacceptable” to “Completely Acceptable.” In analyzing the data, we focused on the “Unacceptable” ratings (1 or 2) and the “Acceptable” ratings (4 or 5). These are plotted in Figure 6.26 as a function of the load time. Looking at the data this way makes it clear that a “crossover” from acceptable to unacceptable happened between three and five seconds.

B. J. Fogg and his associates at the Stanford Persuasive Technology Lab conducted a series of studies to learn more about what makes a website *credible* (Fogg et al., 2001). For example, they used a 51-item questionnaire to assess how believable a website is. Each item was a statement about some aspect of the site, such as “This site makes it hard to distinguish ads from content,” and an associated 7-point scale from “Much less believable” to “Much more believable,” on which the users rated the impact of that aspect on how believable the site is. They found that data from the 51 items fell into seven scales, which they labeled as Real-World Feel, Ease of Use, Expertise, Trustworthiness, Tailoring, Commercial Implications, and Amateurism. For example, one of the 51 items that weighted strongly in the “Real-World Feel” scale was “The site lists the organization’s physical address.”

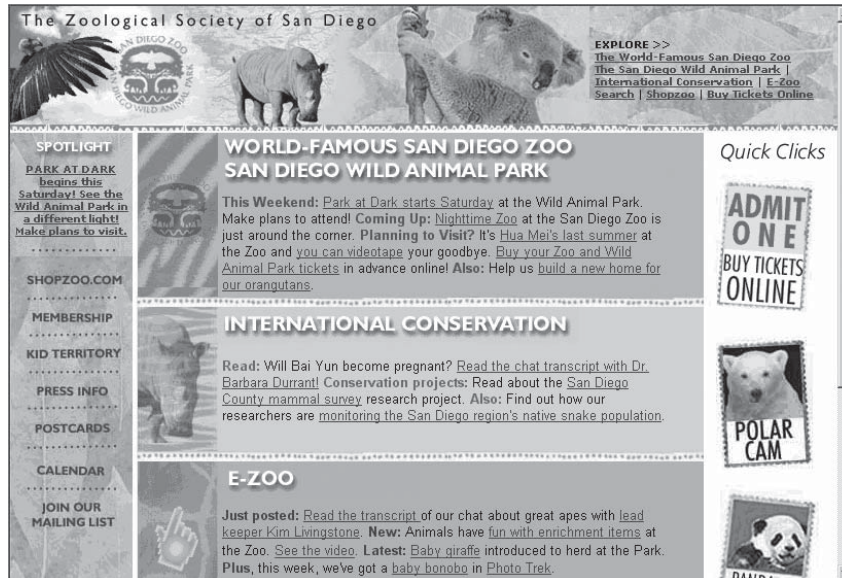
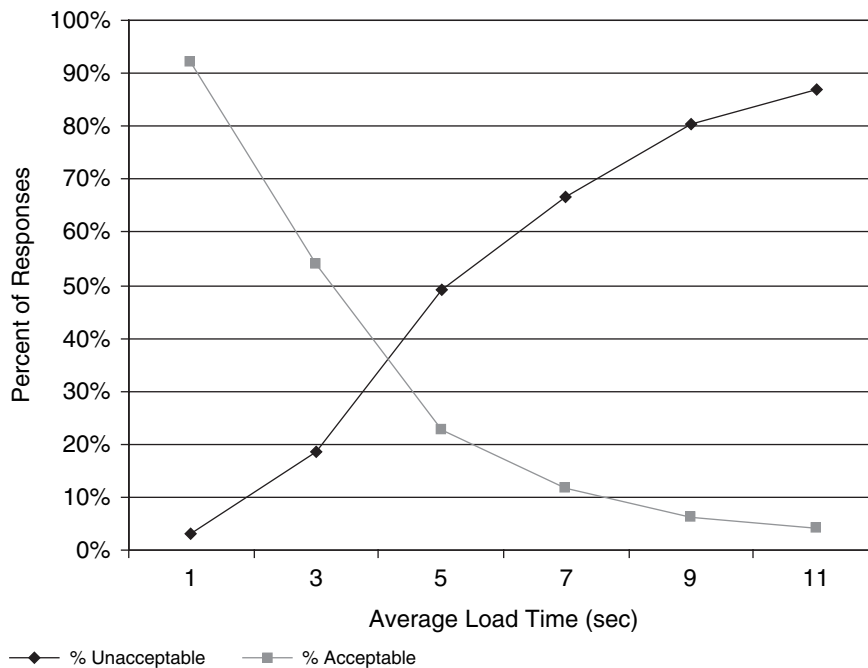


FIGURE 6.25

Screenshots of two websites that were rated as the most engaging of the ten sites that were examined.

**FIGURE 6.26**

Data from participants who rated the acceptability of various load times for an intranet homepage presented in a random order. The ratings were on a 5-point scale, and the data from the online study shown here are for the bottom two (Unacceptable) and top two (Acceptable) values only.

SHOULD YOU NUMBER SCALE VALUES?

One of the issues that comes up in designing rating scales is whether to show a numeric value for each scale position. The examples we've shown in this chapter have included both numbered and unnumbered scales. Our sense is that with scales of no more than five or seven values, adding numeric values for each position is probably not necessary. But as you increase the number of scale values, numbers might become more useful in helping the user keep track of where she or he is on the scale.

6.7.2 Assessing Specific Elements

In addition to assessing specific *aspects* of a product or website, you might be interested in assessing specific *elements* of it, such as instructions, FAQs, or online help; the homepage; the search function; or the site map. The techniques for

assessing subjective reactions to specific elements are basically the same as for assessing specific aspects. You simply ask the participant to focus on the specific element and then present some appropriate rating scales.

The Nielsen Norman Group (Stover, Coyne, & Nielsen, 2002) conducted a study that focused specifically on the site maps of ten different websites. After interacting with a site, the participants completed a questionnaire that included six statements related to the site map:

- The site map is easy to find.
- The information on the site map is helpful.
- The site map is easy to use.
- The site map made it easy to find the information I was looking for.
- The site map made it easy to understand the structure of the website.
- The site map made it clear what content is available on the website.

Each statement was accompanied by a 7-point Likert scale of “Strongly Disagree” to “Strongly Agree.” They then averaged the ratings from the six scales to get an overall rating of the site map for each of the ten sites. This is an example of getting more reliable ratings of a feature of a website by asking for several different ratings of the feature and then averaging them together.

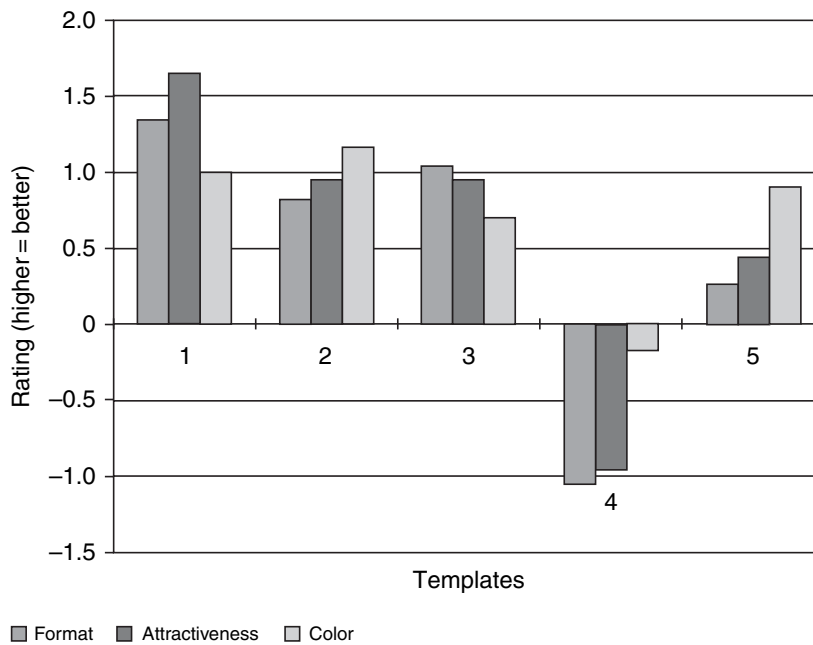
Tullis (1998) conducted a study that focused on candidate homepage designs for a website. (In fact, the designs were really just templates containing “placeholder” text.) One of the techniques he used for comparing the designs was to ask participants in the study to rate the designs on three rating scales: page format, attractiveness, and use of color. Each was rated on a 5-point scale (–2, –1, 0, 1, 2) of “Poor” to “Excellent.” The results for the five designs are shown in Figure 6.27. The design that received the best ratings was Template 1, and the design that received the worst ratings was Template 4.

This study also illustrates another common technique in studies that involve a comparison of alternatives. The participants were asked to rank-order the five templates from their most preferred to least preferred. These data can be analyzed by looking at the average rank for each alternative or at the percentages of high or low ranks. In this study, 48 percent of the participants ranked Template 1 as their first choice, while 57 percent ranked Template 4 as their last choice.

6.7.3 Open-Ended Questions

Most questionnaires in usability studies include some open-ended questions in addition to the various kinds of rating scales that we’ve discussed in this chapter. In fact, one common technique is to allow the participant to add comments related to any of the individual rating scales. Although the utility of these comments to the calculation of any metrics is probably limited, they can be very helpful in identifying ways to improve the product.

Another flavor of open-ended question commonly used in usability studies is to ask the participants to list three to five things they like the *most* about the product

**FIGURE 6.27**

Data in which five different designs for a website's homepage were each rated on three scales: format, attractiveness, and use of color. *Source:* Adapted from Tullis (1998).

and three to five things they like the *least*. These can be translated into metrics by counting the number of instances of essentially the same thing being listed and then reporting those frequencies.

6.7.4 Awareness and Comprehension

A technique that somewhat blurs the distinction between self-reported data and performance data involves asking the users some questions about what they saw or remember from interacting with the application or website after they have performed some tasks with it, and not being allowed to refer back to it. One flavor of this is a check for awareness of various features of a website. For example, consider the homepage shown in Figure 6.28. First, the participant would be given a chance to explore the site a little and complete a few very general tasks like reading the latest news about the International Space Station, finding out how to get images from the Hubble Space Telescope, and learning more about the continuing impact of Hurricane Katrina. Then, with the site no longer available to the participant, a questionnaire is given that lists a variety of specific pieces of content that the site may or may not have had.

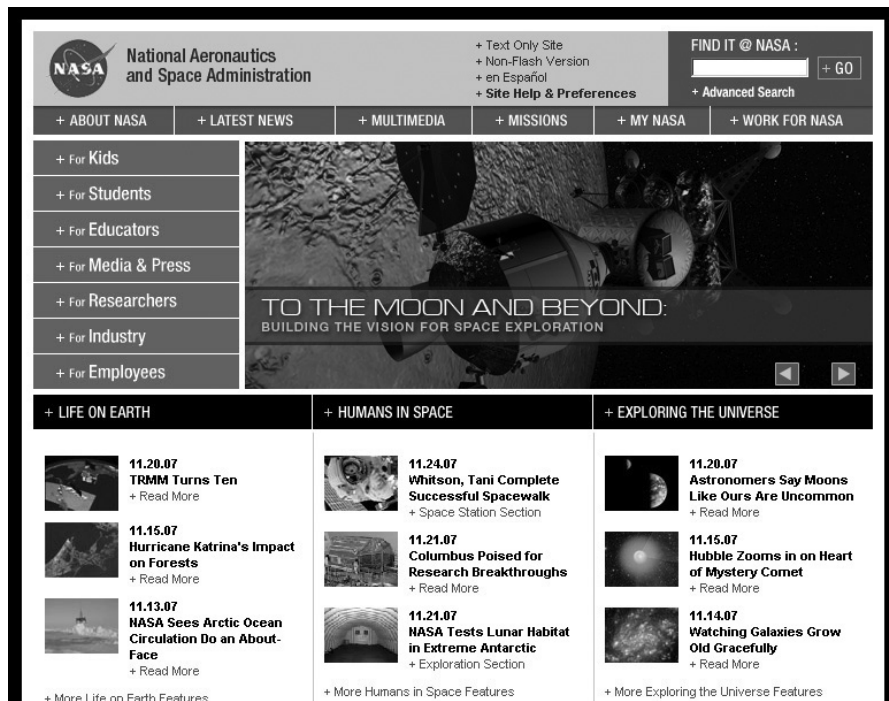


FIGURE 6.28

This NASA homepage illustrates one technique for assessing how "attention-grabbing" various elements of a web page are. After letting participants interact with it, you ask them to identify from a list of content items which ones were actually on the site.

These would generally be content *not* directly related to the specific tasks that the participant was asked to perform. You're interested in whether some of these other pieces of content "stood out." The participant then indicates on the questionnaire which of the pieces of content he or she remembers seeing on the site. For example, two of the items on the questionnaire might be "NASA's testing of a possible lunar habitat" and "studies of Arctic Ocean circulation," both of which are links on the homepage. One of the challenges in designing such a questionnaire is that it must include logical "distracter" items as well—items that were not on the website (or page, if you limit the study to one page) but that look like they could have been.

A closely related technique involves testing for the participants' learning and comprehension related to some of the website's content. After interacting with a site, they are given a quiz to test their comprehension of some of the information on the site. If the information is something that some of the participants might have already known prior to using the site, it would be necessary to administer a pretest to determine what they already know and then compare their results from

the post-test to that. When the participants are not overtly directed to the information during their interaction with the site, this is usually called an “incidental learning” technique.

6.7.5 Awareness and Usefulness Gaps

One type of analysis that can be very valuable is to look at the difference between participants’ *awareness* of a specific piece of information or functionality and the perceived *usefulness* of that same piece of information or functionality once they are made aware of it. For example, if a vast majority of participants are unaware of some specific functionality, but once they notice it they find it very useful, you should promote or highlight that functionality in some way.

To analyze awareness–usefulness gaps, you must have both an awareness and a usefulness metric. We typically ask participants about awareness as a yes/no question—for example, “Were you aware of this functionality prior to this study? (yes or no).” Then we ask: “On a 1 to 5 scale, how useful is this functionality to you? (1 = Not at all useful; 5 = Very useful).” This assumes that they have had a couple of minutes to explore the functionality. Next, you will need to convert the rating-scale data into a top-2-box score so that you have an apples-to-apples comparison. Simply plot the percent of participants who are aware of the functionality next to the percent of those who found the functionality useful (percent top-2-box). The difference between the two bars is called the *awareness–usefulness gap* (see Figure 6.29).

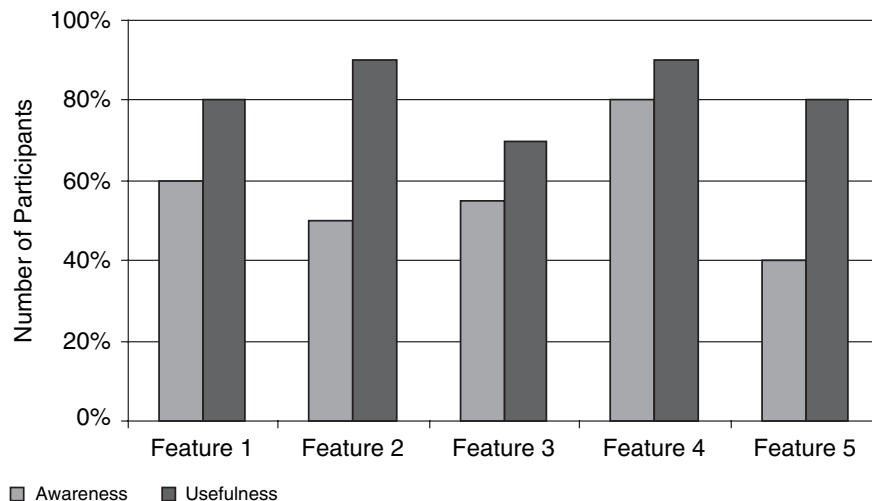


FIGURE 6.29

Data from a study looking at awareness–usefulness gaps. Items with the greatest difference between the awareness and usefulness ratings, such as Features 2 and 5, are those you should consider making more obvious in the interface.

6.8 SUMMARY

Many different techniques are available for getting usability metrics from self-reported data. The following are some of the key points to remember.

1. Consider getting self-reported data both at a task level and at the end of the usability session. Task-level data can help you identify areas that need improvement. Session-level data can help you get a sense of overall usability.
2. When testing in a lab, consider using one of the standard questionnaires for assessing subjective reactions to a system. The System Usability Scale has been shown to be robust even with relatively small numbers of participants (e.g., 8–10).
3. When testing a live website, consider using one of the online services for measuring user satisfaction. The major advantage they provide is the ability to show you how the results for your website compare to a large number of sites in their reference database.
4. Be creative in the use of other techniques in addition to simple rating scales. When possible, ask for ratings on a given topic in several different ways and average the results to get more consistent data. Carefully construct any new rating scales. Make appropriate use of open-ended questions, and consider techniques like checking for awareness or comprehension after interacting with the product.