Measuring
the User Experience
Collecting, Analyzing, and Presenting Usability Metrics

Second Edition

Tom Tullis
Bill Albert

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK
OXFORD • PARIS • SAN DIEGO • SAN FRANCISCO
SINGAPORE • SYDNEY • TOKYO
Morgan Kaufmann is an imprint of Elsevier

CHAPTER 6

Self-Reported Metrics

CONTENTS

Measuring the User Experience. DOI: http://dx.doi.org/10.1016/B978-0-12-415781-1.00006-6

Perhaps the most obvious way to learn about the usability of something is to ask the participants to tell you about their experience with it. But exactly how to ask participants 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 choose from, and open-ended questions such as "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, trust in a company that sponsors a website, enjoyment in playing a game, and many others. But the common feature of all of these is you're asking the participant for information, which is why we think *self-reported* best describes these metrics. And as we will see, one critical type of self-reported data is the verbatim comments made by participants while using a product.

THE EVOLUTION OF USABILITY AND USER EXPERIENCE

One of the historical precedents for the usability field was human factors, or ergonomics, which itself grew primarily out of World War II and a desire to improve airplane cockpits to minimize pilot error. With this ancestry, it's not surprising that much of the early focus of usability was on performance data (e.g., speed and accuracy). But that has been changing, quite significantly we think. Part of the reason for the widespread adoption of the term "user experience," or UX, is the focus that it provides on the entire range of experience that the user has with a product. Even the Usability Professionals Association changed its name in 2012 to the User Experience Professionals Association. All of this reflects the importance of the kind of metrics discussed in this chapter, which try to encompass such states as delight, joy, trust, fun, challenge, anger, frustration, and many more. An interesting analysis was done by Bargas-Avila and Hornbæk (2011) of 66 empirical studies in the UX literature from 2005 to 2009 showing how the studies reflect some of these shifts. They found, for example, that emotions, enjoyment, and aesthetics were the most frequently assessed UX dimensions in the recent studies.

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 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 user experience professional, 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 UX 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 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 RATING SCALES

One of the most common ways to capture self-reported data in a UX study is with some type of rating scales. Two of the classic approaches to rating scales are the Likert scale and the semantic differential scale.

6.2.1 Likert Scales

A typical item in a Likert scale is a statement to which 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 five-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, such as Agree, and did not use numbers. Some people prefer to use a seven-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) and perhaps the middle, or neutral, point. 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 a degree of agreement with a statement and (2) it uses an odd number of response options, thus allowing a neutral response. By convention, the "Strongly Agree" end of a Likert scale is generally shown on the right when presented horizontally.

In designing the *statements* for Likert scales, you need to be careful how you word them. You should avoid adverbs such as *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, 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	o	o	o	o	o	o	o	Strong
Ugly	o	o	o	o	o	o	o	Beautiful
Cool	o	o	o	o	o	o	o	Warm
Amateur	o	o	o	o	o	o	o	Professional

Like the Likert scale, a five- or seven-point scale is commonly 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 (such as "good/bad"), potency (such as "strong/weak"), and activity (such as "passive/active").

6.2.3 When to Collect Self-Reported Data

During a usability study, you might collect self-reported data in the form of verbatim comments from a think-aloud protocol while the participants are interacting with the product. Two additional times when you might want to probe more explicitly for self-reported data are immediately after each task (post-task ratings) and at the end of the entire session (poststudy ratings). Poststudy 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 Ratings

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, and may introduce some bias as participants sometimes feel uncomfortable verbally stating poor ratings. This works best for a single, quick rating after each task.

Paper forms and online forms are suitable both for quick ratings and for longer surveys. Paper forms may be easier to create than online, but they involve manual entry of 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 or perhaps tablet computer with the online questionnaire next to the participant's computer in the usability lab. The participant can then refer to the application or website easily 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 include Google Docs' Forms, Qualtrics.com, 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.

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., 2008). 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. For example, people who are called on the phone and asked to evaluate their satisfaction with a product typically report 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 user's responses until after the participant has left. This might mean either turning away or leaving the room when the user fills out the automated or paper survey. Making the survey itself anonymous may also elicit more honest reactions. Some UX researchers have suggested asking participants in a usability test to complete a post-test survey after they get back to their office or home. 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 drop-off in terms of who completes the survey. Another drawback is that it increases the amount of time between users' interaction with the product and their evaluation via the survey, which could have unpredictable results.

6.2.6 General Guidelines for Rating Scales

Crafting good rating scales and questions is difficult; it's both an art and a science. So before you go off on your own, look at existing sets of questions, such as those in this chapter, to see if you can't use those instead. But if you decide that you need to create your own, here are some general points to consider:

- **Multiple scales to help "triangulate."** When creating 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.
- **Odd or even number of values?** The number of values to use in rating scales can be a source of heated debate among UX professionals. Many 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. We believe that in most real-world

situations a neutral reaction is a perfectly valid reaction and should be allowed on a rating scale. So in most cases we use rating scales with an odd number of points. However, there's some indication that not including a midpoint may minimize the effect of the social desirability bias in a face-to-face administration of rating scales (e.g., Garland, 1991).

- **Total number of points.** The other issue, of course, is the actual number of points to use on the rating scales. Some people seem to believe "more is always better," but we don't really agree with that. The survey literature suggests that any more than nine points rarely provides useful additional information (e.g., Cox, 1980; Friedman & Friedman, 1986). In practice, we almost always use five or seven points.

IS FIVE POINTS ENOUGH FOR A RATING SCALE?

Kraig Finstad (2010) did an interesting study comparing five- and seven-point versions of the same set of rating scales [the System Usability Scale (SUS), discussed later in this chapter]. The ratings were administered orally. He counted the number of times that the participant answered with an "interpolation," such as 3.5, 3½, or "between 3 and 4." In other words, the participant wanted to pick a value *between* two of the values given on the scale. He found that participants using the five-point version of the scale were significantly more likely to use interpolations than those using the seven-point version. In fact, about 3% of the individual ratings on the five-point scale were interpolations, while *none* of the ratings on the seven-point scale were. This would suggest that verbal (and perhaps paper-based) rating scales, where the participant could be tempted to use interpolations, might yield more accurate results with seven-point scales.

SHOULD YOU NUMBER SCALE VALUES?

One of the issues that comes up in designing rating scales is whether to show the user a numeric value for each scale position. Our sense is that with scales of no more than five or seven values, adding numbers for each position is 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. But don't use something like −3, −2, −1, 0, +1, +2, +3. Studies have shown that people tend to avoid using zero or negative values (e.g., Sangster et al., 2001; Schwartz et al., 1991).

6.2.7 Analyzing Rating-Scale Data

The most 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 five-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 UX professionals as well as market researchers. Even though rating-scale data are not technically interval data, many professionals treat it as interval. For example, we assume 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.

When analyzing data from rating scales, it's always important to look at the actual frequency distribution of the responses. Because of the relatively small number of response options (e.g., 5–9) for each rating scale, it's even more important to look at the distribution than it is for truly continuous data such as task times. You might see important information in the distribution of responses that you would totally miss if you just looked at the average. For example, let's assume you asked 20 users to rate their agreement with the statement "This website is easy to use" on a 1 to 7 scale, and the resulting average rating was 4 (right in the middle). You might conclude that the users were basically just lukewarm about the site's ease of use. But then you look at the distribution of the ratings and you see that 10 users rated it a "1" and 10 rated it a "7". So, in fact, no one was lukewarm. They either thought it was great or they hated it. You might then want to do some segmentation analysis to see if the people who hated it have anything in common (e.g., they had never used the site before) vs the people who loved it (e.g., long-time users of the site).

WHAT NUMBER SHOULD RATING SCALES START WITH?

Regardless of whether you show numbers for each scale value to the *user*, you will normally use numbers *internally* for analysis. But what number should the scales start with, zero or one? It generally doesn't matter, as long as you report what the scale is whenever showing mean ratings (e.g., a mean of 3.2 on a scale of 1 to 5). But there are some cases where it's convenient to start the scale at zero, particularly if you want to express the ratings as percentages of the best possible rating. On a scale of 1 to 5, a rating of 5 would correspond to 100%, but a rating of 1 does not correspond to 20%, as some might think (e.g., calculating the percentage by multiplying the rating by 20, which is wrong). On a scale of 1 to 5, 1 is the lowest possible rating, so it should correspond to 0%. Consequently, we find it's easier to keep our sanity by internally numbering rating scales starting at zero, so that a rating of 0 corresponds to 0%.

Another way to analyze rating-scale data is by looking at top-box or top-2-box scores. Assume you're using a rating scale of 1 to 5, with 5 meaning "Strongly Agree." The sample data in Figure 6.1 illustrate the calculation of top-box and top-2-box scores. A top-box score would be the percentage of participants who gave a rating of 5. Similarly, a top-2-box score would be the

percentage of participants who gave a rating of 4 or 5. (Top-2-box scores are used more commonly with larger scales, such as 7 or 9 points.) The theory behind this method of analysis is that it lets you focus on how many participants gave very positive ratings. (Note that the analysis can also be done as a bottom-box or bottom-2-box analysis, focusing on the other extreme.) Keep in mind that when you convert to a top-box or top-2-box score, the data can no longer be considered interval. Therefore, you should just report the data as frequencies (e.g., the percentage of users who gave a top-box rating). Also keep in mind that you lose information by calculating a top-box or top-2-box score. Lower ratings are ignored by this analysis.

C2 f_x =IF(B2>4,1,0)

	A	B	C	D
1	**Participant**	**Rating (1-5)**	**Top Box?**	**Top 2 Box?**
2	P1	4	0	1
3	P2	5	1	1
4	P3	3	0	0
5	P4	4	0	1
6	P5	2	0	0
7	P6	3	0	0
8	P7	5	1	1
9	P8	4	0	1
10	P9	3	0	0
11	P10	5	1	1
12	**Averages**	**3.8**	**30%**	**60%**

Figure 6.1 Example of the calculation of top-box and top-2-box scores from ratings in Excel. The "=IF" function in Excel is used to check whether an individual rating is greater than 4 (for Top Box) or greater than 3 (for Top 2 Box). If it is, a value of "1" is given. If not, a value of "0" is given. Averaging these 1's and 0's together gives you the percentage of Top-Box or Top-2-Box scores.

From a practical standpoint, what difference does it make when you analyze rating scales using means vs top-box or top-2-box scores? To illustrate the difference, we looked at data from an online study conducted on the eve of the 2008 U.S. presidential election (Tullis, 2008). There were two leading candidates, Barack Obama and John McCain, both of whom had websites about their candidacy. Participants were asked to perform the same four tasks on one of the sites (which they were assigned to randomly). After each task, they were asked to rate how easy it was on a scale of 1 to 5, with 1 = Very Difficult and 5 = Very Easy. A total of 25 participants performed tasks on the Obama site and 19 on the McCain site. We then analyzed the task ease ratings by calculating the means, top-box scores, and top-2-box scores. The results are shown in Figure 6.2.

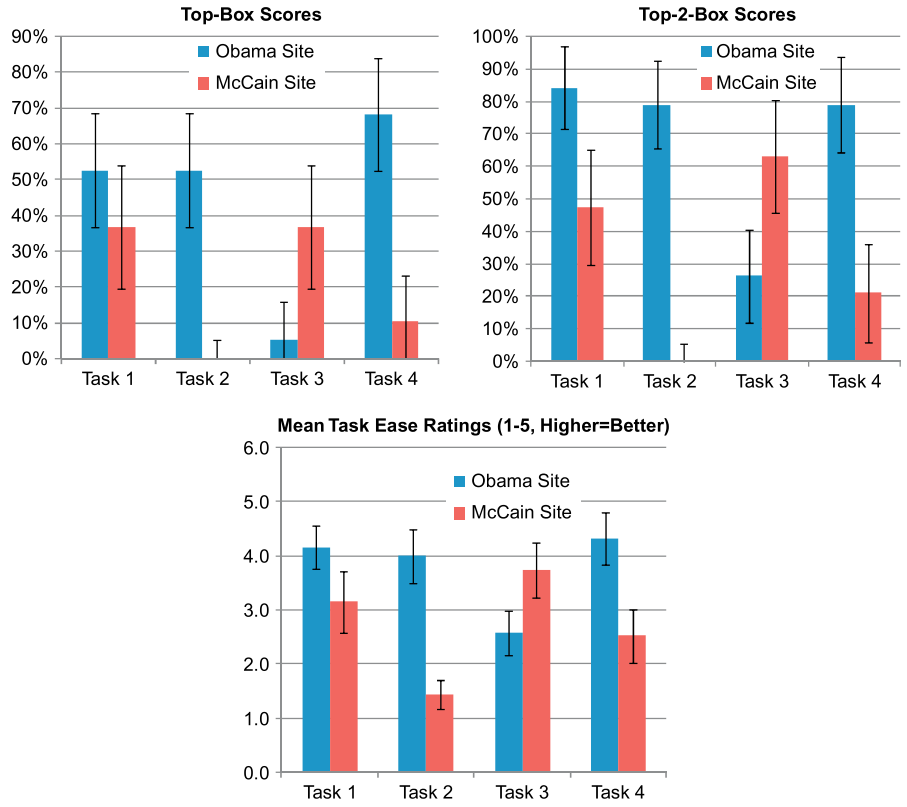


Figure 6.2 Three different analyses of task ease ratings from a study of the Obama and McCain websites (Tullis, 2008): mean ratings, top-2-box scores, and top-box scores. Note how similar patterns are revealed in all three analysis methods, but the apparent disparity between the two sites differs. In each chart, error bars represent a 90% confidence interval.

All three charts seem to indicate that the Obama site got a higher rating than the McCain site for three tasks (Tasks 1, 2, and 4), while the McCain site got a higher rating than the Obama site for one task (Task 3). However, the apparent disparity between the two sites differs depending on the analysis method. There tends to be a greater difference between the two sites with the top-box and top-2-box scores compared to the means. (And no, that's not an error in the top-box and top-2-box charts for Task 2. *None* of the participants gave that task a top-box or top-2-box rating for the McCain site.) But also note that the error bars tend to be larger with the top-box and top-2-box scores compared to the means.

Should you analyze rating scales using means or top-box scores? In practice, we generally use means because they take all data into account (not ignoring some ratings as in top-box or top-2-box analyses). But because some companies or senior executives are more familiar with top-box scores (often from market

research), we use top-box scores in some situations. (It's always important to understand who you're presenting your results to.)

HOW DO YOU CALCULATE CONFIDENCE INTERVALS FOR TOP-BOX SCORES?

If you're calculating means of ratings, then you can calculate confidence intervals in the same way you do for any other continuous data: using the "=CONFIDENCE" function in Excel. But if you're calculating top-box or top-2-box scores, it's not so simple. When you calculate a top-box or top-2-box value for each rating, you're turning it into binary data: each rating is either a top-box value (or top-2-box value) or it's not. This is obvious from Figure 6.1, where each of the top-box (or top-2-box) values is either a "0" or a "1". This should ring some mental bells: it's like the task success data that we examined in Chapter 4. When dealing with binary data, confidence intervals need to be calculated using the Adjusted Wald Method. See Chapter 4 for details.

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. For example, the data shown in Figure 6.2 show that users of the Obama site rated Task 3 as the most difficult, while users of the McCain site rated Task 2 as the most difficult.

6.3.1 Ease of Use

Probably the most common rating scale involves simply asking users to rate how easy or how difficult each task was. This typically involves asking them to rate the task using a five- or seven-point scale. Some UX 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 such as "Easy/Difficult." Either technique will provide you with a measure of perceived usability on a task level. Sauro and Dumas (2009) tested a single seven-point rating scale, which they coined the "Single Ease Question":

Overall, this task was?

Very Difficult o o o o o o o Very Easy

They compared it to several other post-task ratings and found it to be among the most effective.

6.3.2 After-Scenario Questionnaire (ASQ)

Jim Lewis (1991) developed a set of three rating scales—the After-Scenario Questionnaire—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 seven-point rating scale of "Strongly Disagree" to "Strongly Agree." Note that these questions in the ASQ touch upon three fundamental areas of usability: effectiveness (question 1), efficiency (question 2), and satisfaction (all three).

6.3.3 Expectation Measure

Albert and Dixon (2003) proposed a different approach to assessing users' subjective reactions to 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 user *thought* it was going to be. So before the users actually did any of the tasks, they asked them to rate how easy/difficult they *expect* each of the tasks to be, based simply on their understanding of the tasks and the type of product. Users expect some tasks to be easier than others. For example, getting the current quote on a stock should be easier than rebalancing an entire portfolio. Then, after performing each task, the users 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 seven-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 visualize these two scores for each task as a scatterplot, as shown in Figure 6.3.

Figure 6.3 Comparison of 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. 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 users 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 users 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 users 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 users 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 Opportunities" quadrant.

6.3.4 A 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 many 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 we wanted to find out!
- *Condition 3:* "Overall, I am satisfied with the ease of completing this task: Strongly Disagree ….. Strongly Agree"and "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 (1991) ASQ. The third question in the ASQ asks about 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 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 loosely based on a method called Usability Magnitude Estimation (McGee, 2004) in which test participants are asked to create their own "usability scale."

These 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 1131 people participated in the online study, with at least 210 participants using each self-report technique.

The main goal of this study was to see if these rating techniques are sensitive to detecting differences in perceived difficulty of the tasks. But we also wanted to see how the perceived difficulty of the tasks corresponded to the task performance data. We collected task time and binary success data (i.e., whether users 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.

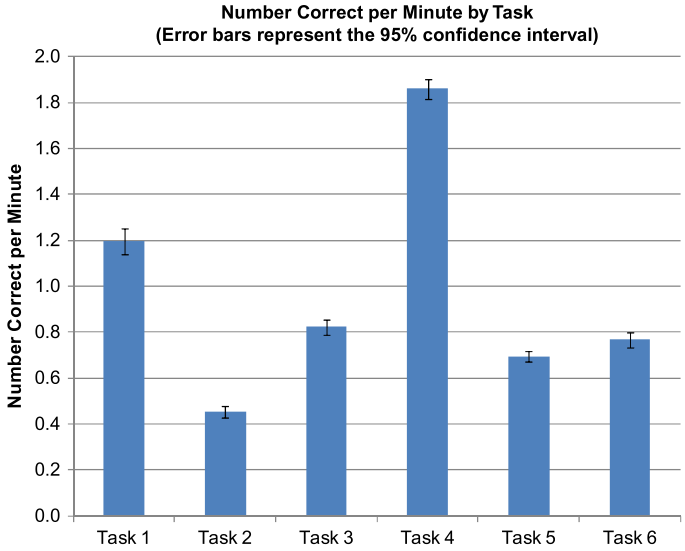


Figure 6.4 Performance data showing that users had the most difficulty with Task 2 and the least difficulty with Task 4. Adapted from Tedesco and Tullis (2006); used with permission.

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 task ratings, correlations were significant for all five conditions ($p < 0.01$). Overall, Spearman rank correlation comparing performance data and task ratings for the six tasks was significant: $Rs = 0.83$.

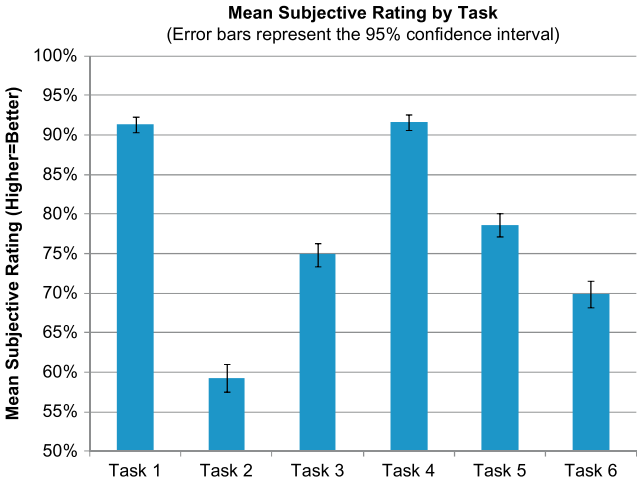


Figure 6.5 Average subjective ratings across all techniques. Ratings are all expressed as a percentage of the maximum possible rating. Similar to the performance data, Task 2 yielded the worst ratings, whereas Task 4 yielded among the best. Adapted from Tedesco and Tullis (2006); used with permission.

Figure 6.6 shows the averages of 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* of 1131). In other words, at large sample sizes, all five of the techniques can effectively distinguish between the tasks.

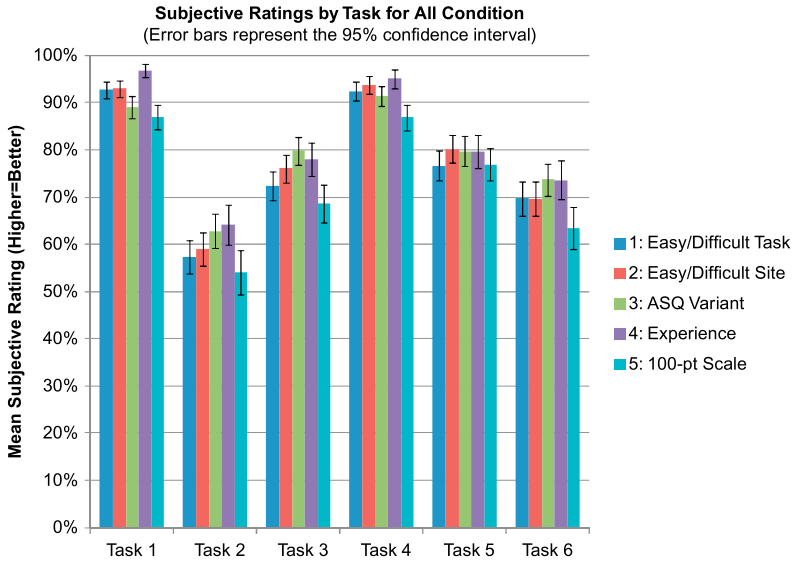


Figure 6.6 Average subjective ratings split by task and condition. All five conditions (self-report techniques) yielded essentially the same pattern of results for the six tasks. Adapted from Tedesco and Tullis (2006); used with permission.

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

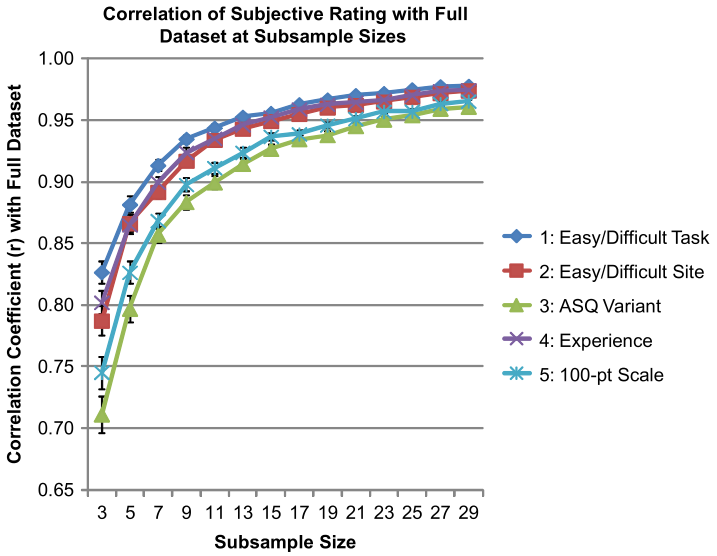


Figure 6.7 Results of a subsampling analysis showing average correlations between ratings for the six tasks from subsamples of various sizes and the full data set for each condition. Error bars represent the 95% confidence interval for the mean. Adapted from Tedesco and Tullis (2006); used with permission.

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 data set 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.

RATINGS DURING A TASK?

At least one study (Teague et al., 2001) indicated that you might get a more accurate measure of the user's experience with a task by asking for ratings *during* the conduct of the task. They found that participants' ratings of ease of use were significantly higher after the task was competed than during the task. It could be that task success changes participants' perception of how difficult the task was to complete.

6.4 POSTSESSION 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. These can be used as an overall "barometer" of the usability of the product, particularly if you establish a track record with the same 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 postsession rating techniques that have been used.

6.4.1 Aggregating Individual Task Ratings

Perhaps the simplest way to look at overall perceived usability is to take an average of the individual task-based ratings. Of course, this assumes that you did in fact collect ratings (e.g., ease of use) after each task. If you did, then simply take an average of them. Or, if some tasks are more important than others, take a weighted average. Keep in mind that these data are 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 self-reported data just once at the end of the session, you are really measuring the participant's last impression of the experience.

This last impression is the perception they 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

One of the most widely used tools for assessing the perceived usability of a system or product is the System Usability Scale. It 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 10 statements to which users rate their level of agreement. Half the statements are worded positively and half are worded negatively. A five-point scale of agreement is used for each. A technique for combining the 10 ratings into an overall score (on a scale of 0 to 100) is also given. It's convenient to think of SUS scores as percentages, as they are on a scale of 0 to 100, with 100 representing a perfect score.

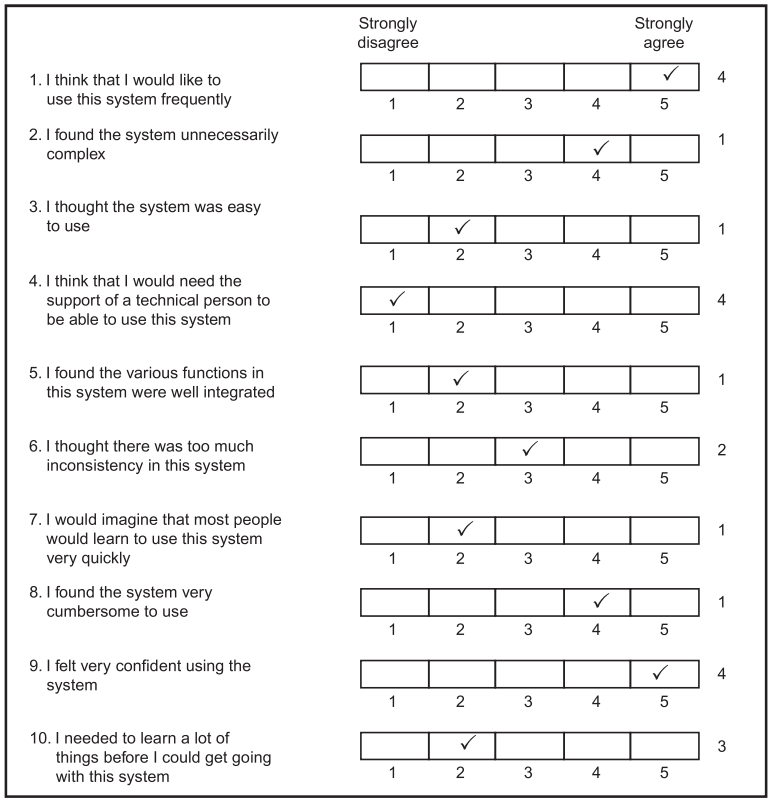


Figure 6.8 The System Usability Scale, developed by John Brooke at Digital Equipment Corporation and an example of scoring it.

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 or, better yet, download our spreadsheet for calculating SUS scores from www.MeasuringUX.com.

The SUS has been made freely available for use in usability studies, both for research purposes and for industry use. The only prerequisite for its use is that any published report should acknowledge the source of the measure. Because it has been so widely used, quite a few studies in the usability literature have

reported SUS scores for many different products and systems, including desktop applications, websites, voice-response systems, and various consumer products. Tullis (2008) and Bangor, Kortum, and Miller (2009) both reported analyses of SUS scores from a wide variety of studies. Tullis (2008a) reported data from 129 different uses of SUS, while Bangor and colleagues (2009) reported data from 206. Frequency distributions of the two sets of data are remarkably similar, as shown in Figure 6.9, with a median study score of 69 for Tullis data and 71 for Bangor et al. data. Bangor and colleagues suggested the following interpretation of SUS scores based on their data:

- <50: Not acceptable
- 50–70: Marginal
- >70: Acceptable

FACTORS IN SUS

Although SUS was originally designed to assess perceived usability as a single attribute, Lewis and Sauro (2009) found that there are actually two factors in SUS. Eight of the questions reflect a usability factor and two reflect a learnability factor. It's easy to compute both from raw SUS ratings.

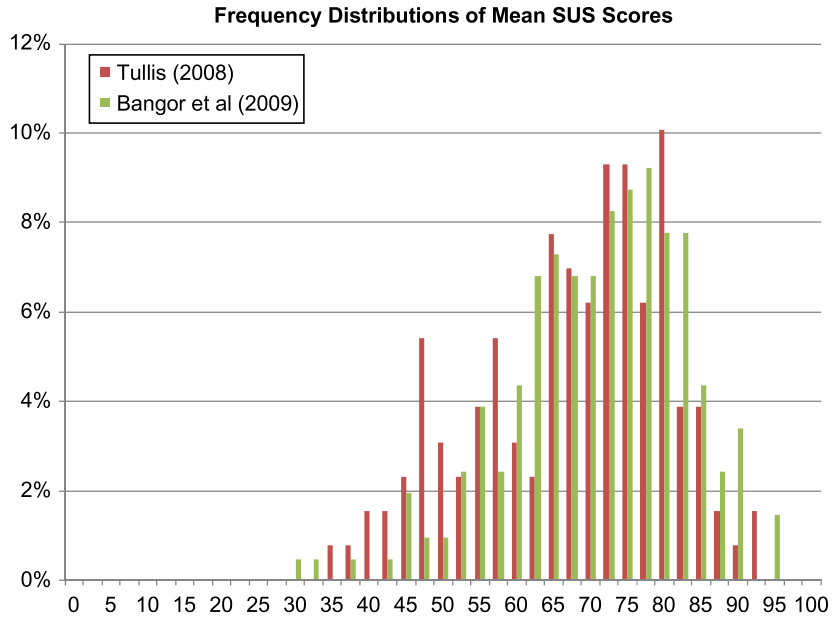


Figure 6.9 Frequency distributions of mean SUS scores reported by Tullis (2008a) and by Bangor et al. (2009). Tullis data are based on a total of 129 study conditions, and Bangor et al. data are based on 206.

DO YOU NEED BOTH POSITIVE AND NEGATIVE STATEMENTS IN SUS?

As shown in Figure 6.8, half of the statements in SUS are positive and half are negative. While some argue that this approach keeps participants "on their toes," others have argued that it also seems to confuse some participants, perhaps causing erroneous responses. Sauro and Lewis (2011) conducted a study in which they compared the traditional version of SUS to an all-positive version. They found no significant difference between mean SUS scores for traditional and all-positive versions. But in a review of 27 SUS data sets, they found evidence that 11% of the studies had some miscoding of SUS data and 13% of the individual SUS questionnaires contained mistakes by users. They suggest using an all-positive version of SUS to avoid some of those possible errors. If you want to use the all-positive version, see Sauro and Lewis (2011) for an example.

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. The PSSUQ was originally designed to be administered in person, whereas CSUQ was designed to be administered by mail or online. CSUQ consists of the following 19 statements to which the user rates agreement on a seven-point scale of "Strongly Disagree" to "Strongly Agree," plus N/A:

1. Overall, I am satisfied with how easy it is to use this system.
2. It was simple to use this system.
3. I could effectively complete the tasks and scenarios using this system.
4. I was able to complete the tasks and scenarios quickly using this system.
5. I was able to efficiently complete the tasks and scenarios using this system.
6. I felt comfortable using this system.
7. It was easy to learn to use this system.
8. I believe I could become productive quickly using this system.
9. The system gave error messages that clearly told me how to fix problems.
10. Whenever I made a mistake using the system, I could recover easily and quickly.
11. The information (such as online help, on-screen messages, and other documentation) provided with this system was clear.
12. It was easy to find the information I needed.
13. The information provided for the system was easy to understand.
14. The information was effective in helping me complete the tasks and scenarios.
15. The organization of information on the system screens was clear.
16. The interface of this system was pleasant.
17. I liked using the interface of this system.

18. This system has all the functions and capabilities I expect it to have.
19. Overall, I am satisfied with this system.

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

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

Figure 6.10 QUIS, 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.

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 6 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) and is available in printed and web versions in multiple languages.

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, data are 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 seven-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) yielded the highest weights for each of the categories, indicating that they contributed most to the results.

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 seven-point Likert scale, ranging from strongly disagree to strongly agree. Statements in *italics* were found to weight less heavily than the others.

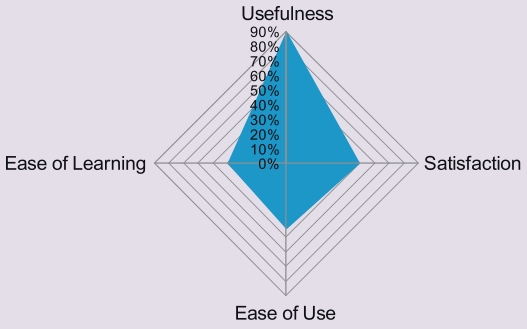
Figure 6.11 The USE questionnaire. From the work of Lund (2001); used with permission.

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 had 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" (under "Other Charts") 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.

6.4.6 Product Reaction Cards

A very different approach to capturing post-test subjective reactions to a product was presented by Joey Benedek and Trish Miner (2002) from Microsoft. As illustrated in Figure 6.12, they presented a set of 118 cards, each containing adjectives (e.g., Fresh, Slow, Sophisticated, Inviting, Entertaining, Incomprehensible). Some of the words are positive and some are negative. The users would then simply choose the cards they felt described the system. After selecting the cards, they 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 from the users. But it can also be used in a quantitative way by counting the number of times each word is chosen by participants. Results can also be visualized using a word cloud (e.g., using Wordle.net). Case study 10.5 provides good examples of word clouds from the product reaction cards.

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	Straight Forward
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 Complete set of 118 product reaction cards developed by Joey Benedek and Trish Miner at Microsoft. 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."

6.4.7 A Comparison of Postsession Self-Reported Metrics

Tullis and Stetson (2004) conducted a study in which we compared a variety of postsession questionnaires for measuring user reactions to websites in an online usability study. We studied the following questionnaires, adapted in the manner indicated for the evaluation of websites.

- *SUS.* It was adapted by replacing the word *system* in every question 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 the user was asked to choose the words that best describe their interaction with the website. They were free to choose as many or as few words as they wished.
- *Our Questionnaire.* We 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 user responds on a seven-point Likert scale from "Strongly Disagree" to "Strongly Agree."

We 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 we analyzed the data from all the participants, we found that all five of the questionnaires revealed that Site 1 got significantly better ratings than Site 2. 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% of the samples would have identified that Site 1 was preferred significantly. But at a sample size of 8, which is relatively common in many lab-based usability tests, we found that SUS would have identified Site 1 as the preferred site 75% of the time—a significantly higher percentage than any of the other questionnaires.

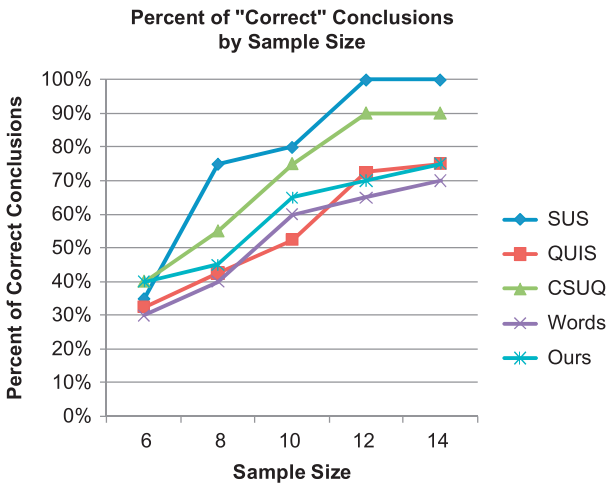


Figure 6.13 Data illustrating the accuracy of results from random subsamples ranging from size 6 to 14. This graph shows what percentage of the random samples yielded the same answer as the full data set at the different sample sizes. Adapted from Tullis and Stetson (2004); used with permission.

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 users must rate their level of agreement. This may keep participants more alert. Another possible 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). All 10 of the rating scales in SUS are simply asking for an assessment of the site as a whole, just in slightly different ways.

6.4.8 Net Promoter Score

One self-reported metric that has gained rapidly in popularity, especially among senior executives, is the Net Promoter Score (NPS). It's intended to be a measure of customer loyalty, and was originated by Fred Reichheld in his 2003 article in the *Harvard Business Review*: "One Number You Need to Grow" (Reichheld, 2003). The power of NPS seems to derive from its simplicity, as it uses only one question: "How likely is it that you would recommend [this company, product, website, etc] to a friend or colleague?" The respondent answers using an 11-point scale of 0 (Not at all likely) to 10 (Extremely likely). The respondents are then divided into three categories:

- Detractors: Those who gave ratings of 0–6
- Passives: Those who gave ratings of 7 or 8
- Promoters: Those who gave ratings of 9 or 10

Note that the categorization into Detractors, Passives, and Promoters is nowhere near symmetrical. By design, the bar is set pretty high to be a Promoter, while it's very easy to be a Detractor. To calculate the NPS, you subtract the percentage of Detractors (ratings of 0–6) from the percentage of Promoters (ratings of 9 or 10). Passives are ignored in the calculation. In theory, NPSs can range from −100 to +100.

The NPS is not without its own detractors. One criticism is that the reduction of scores from an 11-point scale to just three categories (Detractors, Passives, Promoters) results in a loss of statistical power and precision. This is similar to the loss of precision when using the "Top Box" or "Top-2-Box" method of analysis discussed earlier in this chapter. But you lose even more precision when you take the *difference* between two percentages (Promoters minus Detractors), which is similar to subtracting "Bottom Box" scores from "Top Box" scores. Each percentage (% Promoters and % Detractors) has its own confidence interval (or margin of error) associated with it. The confidence interval associated with the *difference* between the two percentages is essentially the combination of the two individual confidence intervals. You would typically need a sample size two to four times larger to get an NPS margin of error equivalent to the margin of error for a traditional Top-2-Box score. Case Study 10.1 provides an excellent example of how NPS can be used to improve the user experience.

DOES PERCEIVED USABILITY PREDICT CUSTOMER LOYALTY?

Jeff Sauro (2010) wanted to know whether usability, as measured by SUS, tended to predict customer loyalty, as measured by NPS. He analyzed data from 146 users asked to complete both the SUS questions and the NPS question for a variety of products, including websites and financial applications. The result was a correlation of $r = 0.61$, which is highly significant ($p < 0.001$). He found that Promoters had an average SUS score of 82, while Detractors had an average SUS score of 67.

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).

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, participants rated each of them using the SUS questionnaire. The average SUS score for the SeniorResource site was 80%, which was significantly better than the average scores for SeniorNet and Seniors-Place, both of which averaged 63%.

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%) was significantly higher than the average for Windows ME (56%)($p < 0.0001$).

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 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. They found that the bubble ballot received significantly higher SUS ratings than either of the other two ($p < 0.001$).

There's also some evidence that participants who have more experience with a product tend to give it higher SUS ratings than those with less experience. In testing two different applications (one web based and one desktop based), McLellan, Muddimer, and Peres (2012) found that the SUS scores from users who had more extensive experience with a product tended to be about 15% higher compared to users with either no or limited experience with the product.

6.6 ONLINE SERVICES

More and more companies are learning 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 studies. This is essentially the same process as in postsession 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 get 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

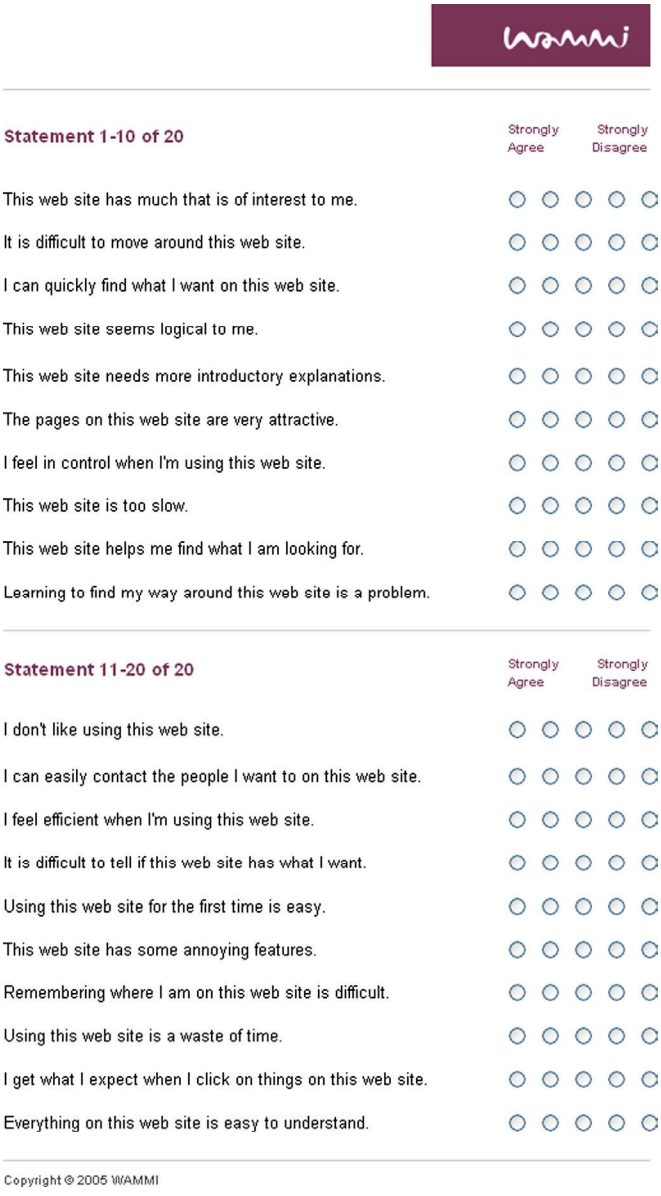


Figure 6.14 The 20 rating scales used by the WAMMI online service.

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 in the Human Factors Research Group of University College Cork in Ireland. Although SUMI is designed for evaluation of software applications, WAMMI is designed for the evaluation of websites.

As shown in Figure 6.14, WAMMI is composed of 20 statements with associated five-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, 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 6.15, 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. The analysis of websites using the ACSI methodology is done by Foresee Results (www.ForeseeResults.com). The ACSI has become particularly popular for analyzing U.S. government websites. For example, 100 U.S. government websites were included in their fourth quarter 2012 analyses of e-government websites (ForeSee Results, 2012). Similarly, their annual Top 100 Online Retail Satisfaction Index assesses such popular sites as Amazon, NetFlix, L.L. Bean, J.C. Penney, Avon, and QVC.

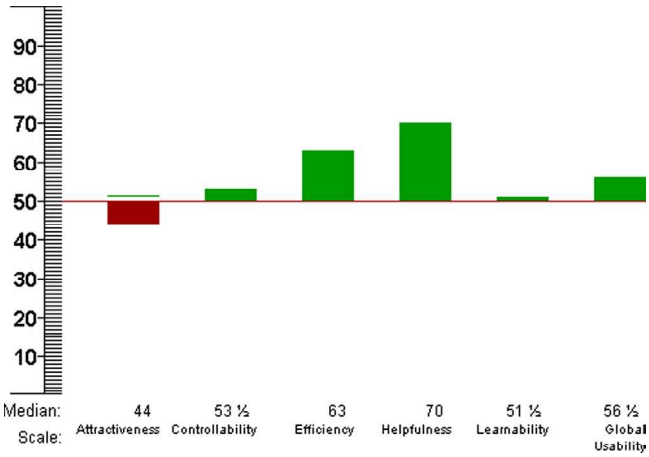


Figure 6.15 Sample data from the WAMMI online service showing average scores in each of five categories, plus an overall usability score.

The ACSI questionnaire for websites is composed of a core set of 14 questions, as shown in Figure 6.16. Each asks for a rating on a 10-point scale of different attributes, such as the quality of information, freshness of content, clarity of site organization, overall satisfaction, and likelihood to return. Specific implementations of the ACSI commonly add additional questions or rating scales.

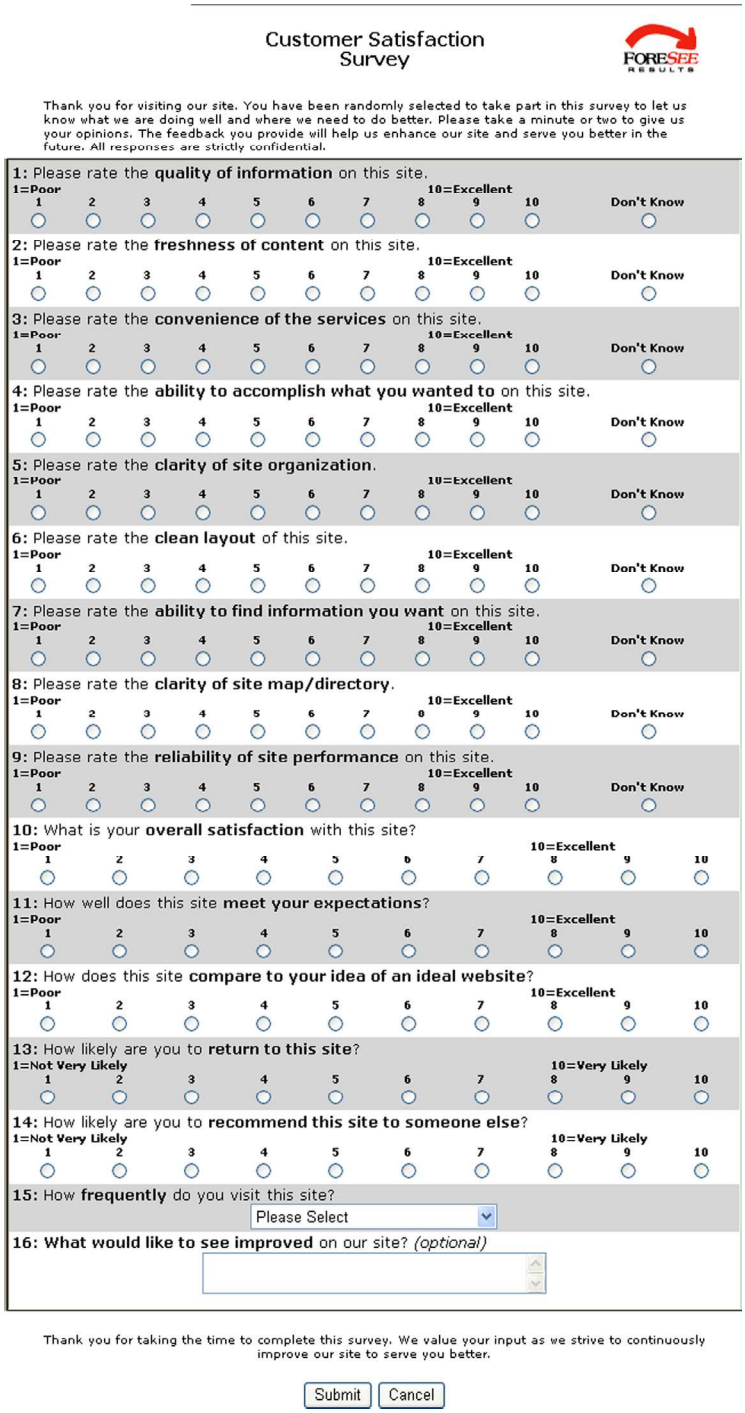
As shown in Figure 6.17, the ACSI results for a website are divided into six quality categories: Content, Functionality, Look & Feel, Navigation, Search, and Site Performance, plus an overall satisfaction score. In addition, they provide average ratings for two "Future Behavior" scores: Likelihood to Return and Recommend to Others. All of the scores are a 100-point scale.

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 6.18, plotting the quality scores on the vertical axis and the impact on overall satisfaction on the horizontal axis. Scores in the lower right quadrant (high impact, low score) indicate the areas where you should focus your improvements.

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 shown in Figure 6.19, 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 6.20 for capturing the feedback. Their scales use five points that are marked simply as − −,−,+−, +, and ++. OpinionLab provides a variety of techniques for



Customer Satisfaction Survey

FORESEE RESULTS

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 10 Don't Know

2: Please rate the **freshness of content** on this site.
1=Poor 1 2 3 4 5 6 7 8 9 10=Excellent 10 Don't Know

3: Please rate the **convenience of the services** on this site.
1=Poor 1 2 3 4 5 6 7 8 9 10=Excellent 10 Don't Know

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 10 Don't Know

5: Please rate the **clarity of site organization**.
1=Poor 1 2 3 4 5 6 7 8 9 10=Excellent 10 Don't Know

6: Please rate the **clean layout** of this site.
1=Poor 1 2 3 4 5 6 7 8 9 10=Excellent 10 Don't Know

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 10 Don't Know

8: Please rate the **clarity of site map/directory**.
1=Poor 1 2 3 4 5 6 7 8 9 10=Excellent 10 Don't Know

9: Please rate the **reliability of site performance** on this site.
1=Poor 1 2 3 4 5 6 7 8 9 10=Excellent 10 Don't Know

10: What is your **overall satisfaction** with this site?
1=Poor 1 2 3 4 5 6 7 8 9 10=Excellent 10

11: How well does this site **meet your expectations**?
1=Poor 1 2 3 4 5 6 7 8 9 10=Excellent 10

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 10

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 10

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 10

15: How **frequently** do you visit this site?
Please Select

16: What would like to see improved on our site? *(optional)*

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.

Submit Cancel

Figure 6.16 Typical questions in an ACSI survey for a website.

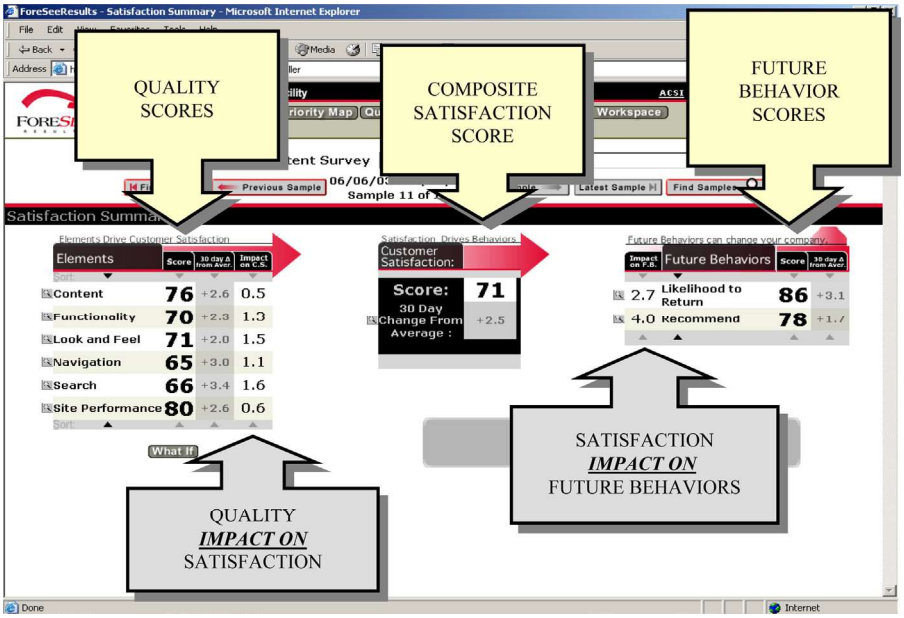


Figure 6.17 Sample results from an ACSI analysis for a website. Scores for six quality areas are shown on the left, along with values estimating the impact that each of those scores has on overall customer satisfaction, which is shown in the center. Scores for two "future behavior" areas are shown on the right, along with values estimating the satisfaction impact on those areas.

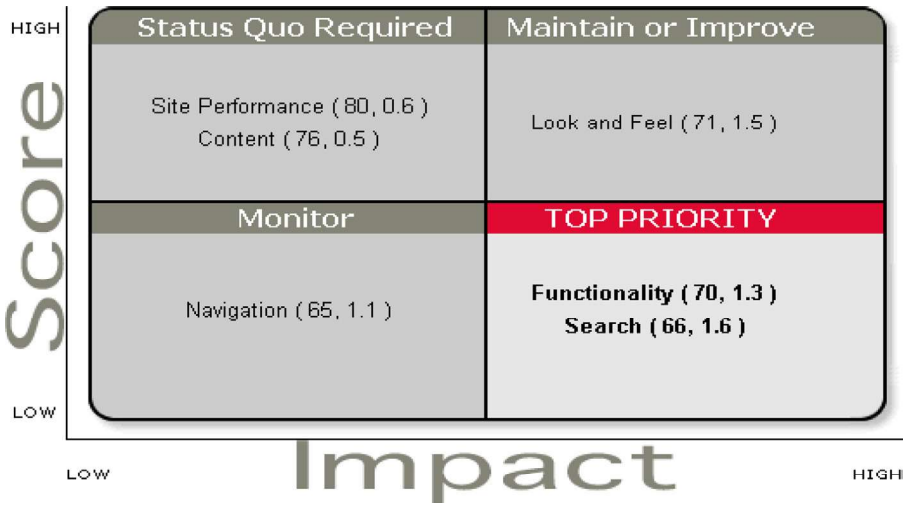


Figure 6.18 Sample results from an ACSI analysis for a website. High and low scores for the six quality areas are represented on the vertical axis, and high and low impact scores are shown on the horizontal axis. The quality areas that fall in the lower right quadrant (Functionality and Search in this example) should be your top priorities for improvement.

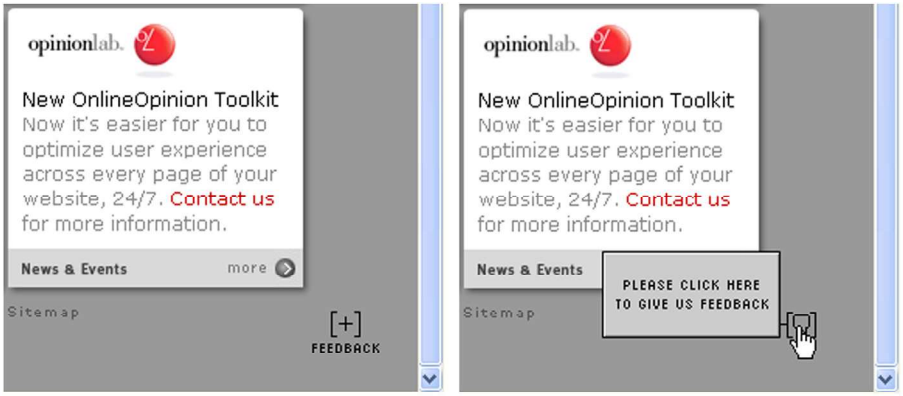


Figure 6.19 An example of a web page containing OpinionLab's feedback mechanism in the lower right corner. This animated icon stays in that position while the user scrolls the page. Moving the mouse over the icon reveals the version shown on the right.

Figure 6.20 Examples of OpinionLab mechanisms for capturing feedback about a web page. The version on the left allows the user to give a quick overall rating of the page. The version on the right allows for more detailed feedback on a few different scales.

visualizing data for a website, including the one shown in Figure 6.21, which allows you to easily spot 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

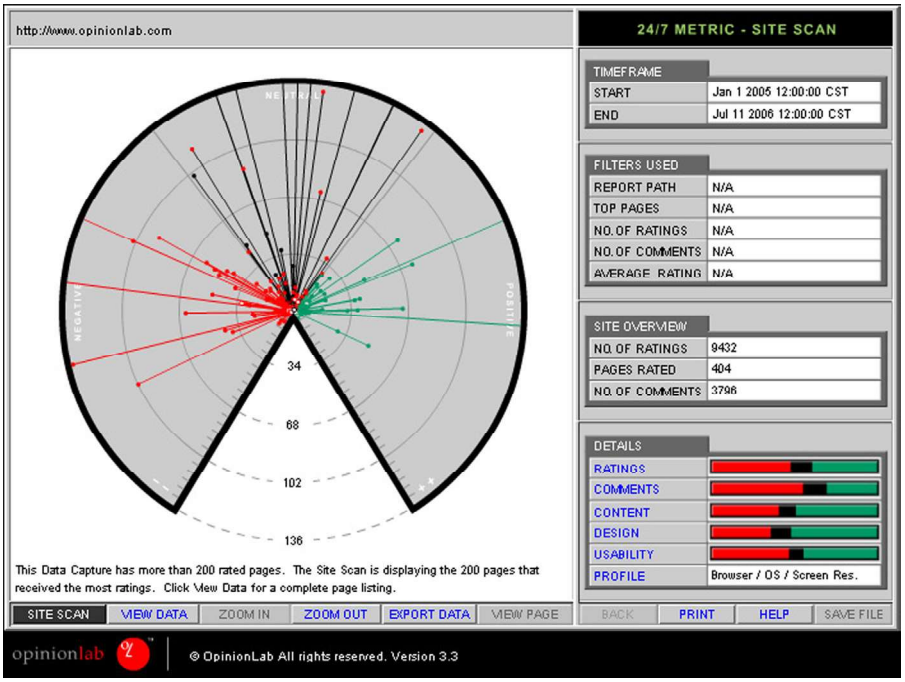


Figure 6.21 OpinionLab provides a variety of techniques for visualizing data for a website. In the visualization shown here, the most-rated 200 pages are represented graphically on the left. Pages receiving the most negative ratings are on the left, those with neutral ratings at the top, and those with the most positive ratings on the right.

OpinionLab keep the number of questions to a minimum. You need to try 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 about 20 is the maximum number of questions you should ask in this type of survey.

- *Self-selection of respondents*. Because respondents make a decision about whether or not to complete the survey, they are self-selecting. You should 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 monitor responses closely 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 or IP address) when the survey has already been offered to someone. As long as the user doesn't clear their cookies and is using the same computer, the survey won't be presented to them 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 the objectives of a usability study, you might want to assess users' reactions to specific *attributes* of the product overall or specific *parts* of the product.

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
- Confidence
- 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 describe a few examples of usability studies that have focused on assessing specific attributes.

Gitte Lindgaard and associates at Carleton University 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 or 500 msec to 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- and the 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.

Bill Albert and associates at Bentley University (Albert, Gribbons, & Almadas, 2009) extended this research to see if users could form an opinion quickly about

their trust of websites based on very brief exposures to images of web pages. They used 50 screenshots of popular financial and health-care websites. After viewing a page for only 50 msec, participants were asked to give a rating of their trust of the site on a 1 to 9 scale. After a break, they repeated the procedure in a second trial with the same 50 images. They found a significant correlation ($r = 0.81$, $p < 0.001$) between the trust ratings in the two trials.

Several years ago, Tullis conducted an online study of 10 different websites to learn more about what makes a website *engaging*. He 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, participants responded to a single rating worded as "This website is: Not At All Engaging … Highly Engaging" using a five-point scale. The two sites that received the highest ratings on this scale are shown in Figure 6.22.

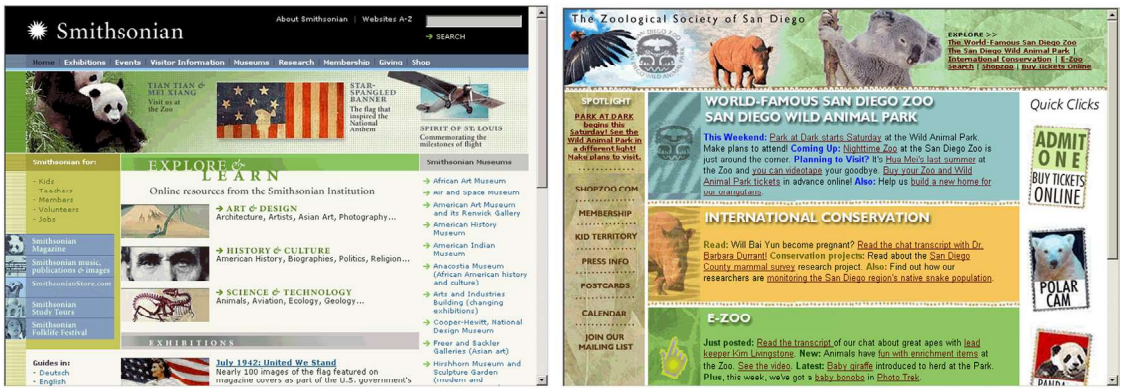


Figure 6.22 Websites rated as the most engaging of 10 sites studied.

One of the techniques often used in analyzing data from subjective rating scales is to focus on the responses that fall onto 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 Box" or "Bottom Box" scores. We used this technique in an online study assessing users' reactions to various load times for an intranet homepage. We manipulated the load time artificially over a range of 1 to 11 seconds. Different load times were presented in a random order, and users were never told what the load time was. After experiencing each load time, users were asked to rate that load time on a five-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 6.23 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 3 and 5 seconds.

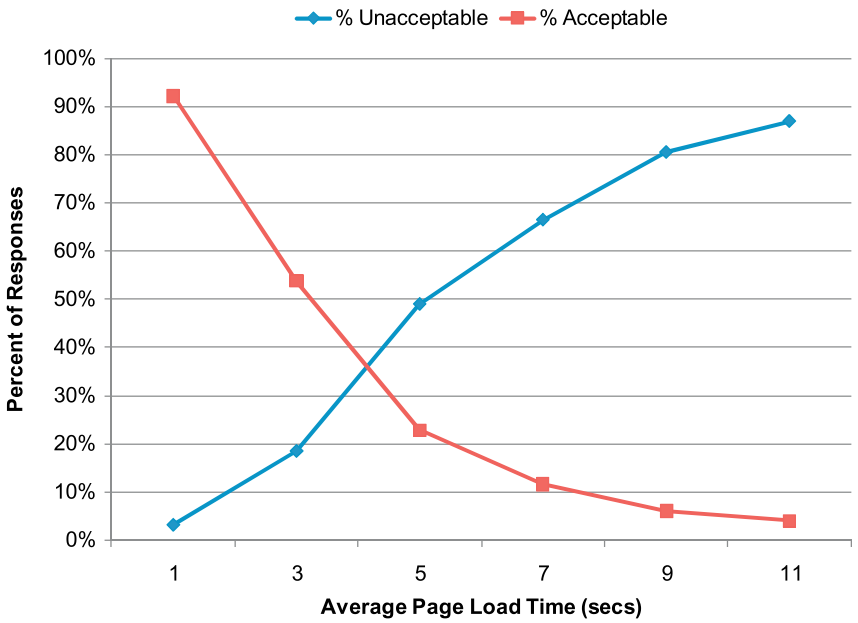


Figure 6.23 Data in which users rated the acceptability of various load times for an intranet homepage presented in a random order. Ratings were on a five-point scale, and data shown here are for the bottom two (Unacceptable) and top two (Acceptable) values only.

B. J. Fogg and 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 seven-point scale from "Much less believable" to "Much more believable," on which 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."

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 user 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 10 different websites. After

interacting with a site, users 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 seven-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 10 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 possible homepage designs for a website. (In fact, the designs were really just templates containing "placeholder" or "Lorem Ipsum" text.) One of the techniques 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 five-point scale (−2, −1, 0, 1, 2) of "Poor" to "Excellent." (Note to self and others: Don't use that scale again. It tends to bias respondents away from the ratings associated with the negative values and zero. But the results are still valid if the main thing we're interested in is the relative *comparison* of the ratings for the different designs.) Results for the five designs are shown in Figure 6.24. The

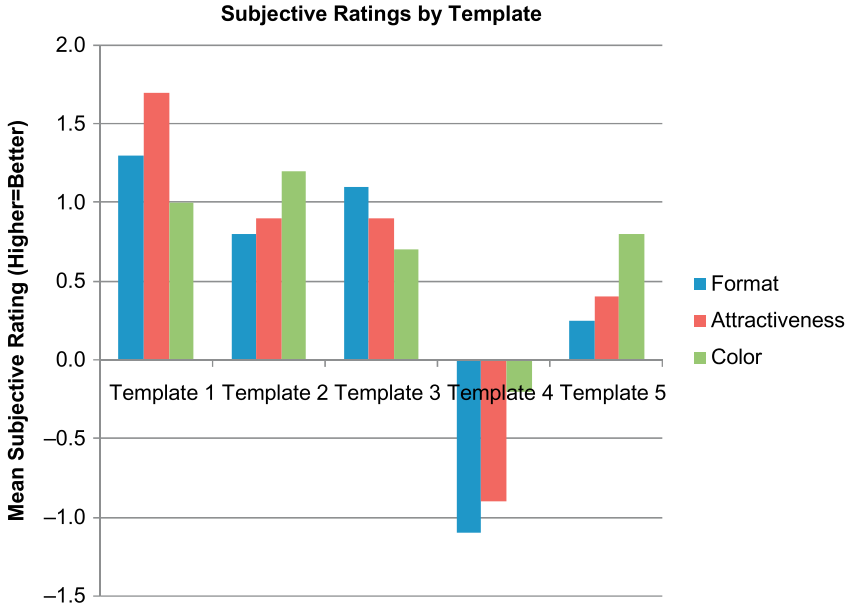


Figure 6.24 Data in which five different designs for a website's homepage were each rated on three scales: format, attractiveness, and use of color. Adapted from Tullis (1998); used with permission.

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. Participants were asked to rank-order the five templates from their most preferred to least preferred. In this study, 48% of the participants ranked Template 1 as their first choice, while 57% 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 user to add comments related to any of the individual rating scales. Although the utility of these comments to the calculation of specific metrics may be limited, they can be very helpful in identifying ways to improve the product.

Another flavor of open-ended question used commonly in usability studies is to ask the users to list three to five things they like the *most* about the product 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. Of course, you could also treat the remarks that participants offer while thinking aloud as these kinds of verbatim comments.

Entire books have been written about analyzing these kinds of verbatim responses using what's generally called text mining (e.g., Miner et al., 2012), and a wide variety of tools are available in this space (e.g., Attensity, Autonomy, Clarabridge, to name a few). We will just describe a few simple techniques for collecting and summarizing these kinds of verbatim comments.

Summarizing responses from open-ended questions is always a challenge. We've never come up with a magic solution to doing this quickly and easily. One thing that helps 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.

Figure 6.25 Word cloud created with Wordle.net of responses in an online study of the NASA website about the Apollo Space Program to a question asking for anything they found particularly frustrating or challenging about the site.

One very simple analysis method that we like is to copy all of the verbatim comments in response to a question into a tool for creating word clouds, such as Wordle.net. For example, Figure 6.25 shows a word cloud of responses to a question asking participants to describe anything they found particularly

challenging or frustrating about using the NASA website about the Apollo Space Program (Tullis, 2008b). In a word cloud, larger text is used to represent words that appear more frequently. It's apparent from this word cloud that participants were commenting frequently on the "search" on the site and the "navigation." (Some frequent words, such as "Apollo," are certainly not surprising given the subject matter.)

EXCEL TIP

Finding All Comments That Include a Specific Word

After studying a word cloud (and the accompanying word frequencies that most of these tools can generate), it's sometimes helpful to find all of the verbatim comments that included specific words anywhere in the comment. For example, after seeing the word cloud in Figure 6.25, it might be helpful to find all the comments that included the word "navigation." This can be done in Excel using the =SEARCH function. You can then sort on the column containing the results of the SEARCH function. Entries containing the target word will have numeric values (actually the character position where the target word starts) and those that don't contain the target word will give a "#VALUE!" error.

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 NASA homepage shown in Figure 6.26. First, the user would be given a chance to explore the site a little and complete a few very general tasks, such as reading the latest news from NASA and finding how to get images from the Hubble Space Telescope. Then, with the site no longer available to the user, a questionnaire is given that lists a variety of specific pieces of content that the site may or may not have had.

These would generally be content *not* related directly to the specific tasks that the user was asked to perform. You're interested in whether some of these other pieces of content "stood out" to the user. The user then indicates which of the pieces of content on the questionnaire he or she remembers seeing on the site. For example, two of the items on the questionnaire might be "When the ISS Crew is Due to Return" and "Satellite observation of the Western wildfires," 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 users' learning and comprehension related to some of the content of the website. After interacting with a

Figure 6.26 This NASA homepage illustrates one technique for assessing how "attention-grabbing" various elements of a web page are. After letting users interact with the site, you ask them to identify from a list of content items which ones were actually on the site.

site, users 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 back to that. When the users 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 users' *awareness* of a specific piece of information or functionality and their perceived *usefulness* of that same piece of information or functionality once they are made aware of it. For example, if a vast majority of users are unaware of some specific functionality, but once they notice it they find it very useful, that suggests you should promote or highlight that functionality in some way.

To analyze awareness–usefulness gaps, you must have both an awareness and usefulness metric. We typically ask users 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 percentage of users who are aware of the functionality next to the percentage of users who found the functionality useful (percent top-2 box). The difference between the two bars is called the *awareness–usefulness gap* (see Figure 6.27).

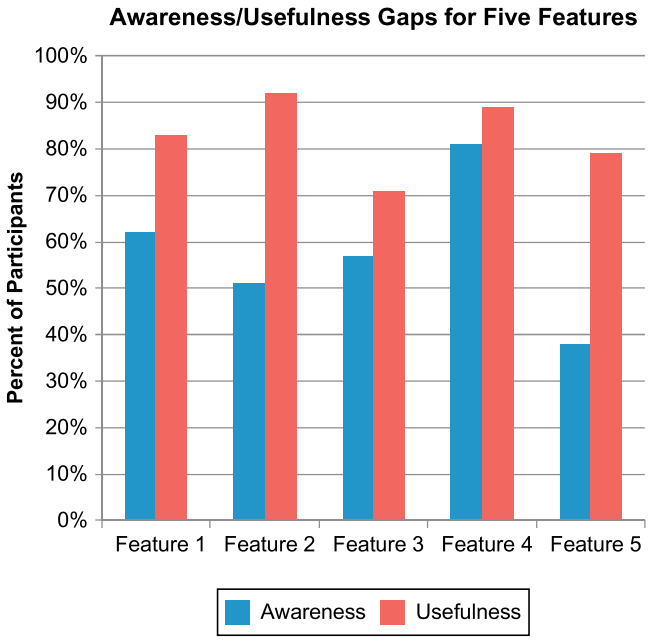


Figure 6.27 Data from a study looking at awareness–usefulness gaps. Items with the greatest difference between awareness and usefulness ratings, such as Tasks 2 and 5, are those you should consider making more obvious in the interface.

6.8 SUMMARY

Many different techniques are available for getting UX metrics from self-reported data. Here's a summary of some of the key points to remember.

1. Consider getting self-reported data at both a task level and at the end of a 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 SUS 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 such as WAMMI or ACSI. The major advantage they provide is the ability to show you how the results for your site compare to a large number of sites in their reference database.
4. Be creative but also cautious 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 such as checking for awareness or comprehension after interacting with the product.