Running Usability Evaluations

Running a usability test with real customers is essential to good design. You may know a lot about your customers, but it is hard to predict how people will react to and interact with a Web site. Usability tests are also effective in ending those endless opinion wars in which members of the design team argue about what people like and don't like. The best way to answer this question is to recruit some participants,[1] run a quick test, and see what they say and do. This appendix lays out the steps for running both formal and informal usability tests, from setting up the test to running the test to analyzing and presenting the results.

Our assumption here is that you want to run a usability test in which both you and the participant are in the same place. You should also be aware that an alternative approach is **remote usability testing**—that is, recruiting and testing many participants online without your having to be there. We discuss how to do this in Appendix E—Online Research.

A.1 Setting Target Goals

What Do You Want to Learn from the Test? • The first thing you have to do is decide what you want to get out of the test. Do you want to find out if people are having problems with a specific part of the Web site? Do you want to see how well a proposed design works? Or do you just want to get general feedback about the existing Web site?

How Will You Get the Information You Want? • After deciding what you want to learn, think about how you will get this information. If people

---

1 In psychology and other fields, the term *subjects* is used instead of *participants,* but we have always felt that the term *subjects* has a slightly sinister tone.

are having problems with a portion of the Web site, the straightforward thing to do is to test tasks that rely on that part and see what the problems are. If you want to test a new design, it is useful to compare it to the old design or to a competitor's Web site. This approach is also useful for getting general feedback about an existing Web site.

Process Data versus Bottom-Line Data • There are two kinds of data that you can get from a usability test: process data and bottom-line data. **Process data** consists of informal, qualitative observations of what people are thinking and doing—an overall feeling of what works and what does not on a Web site. The key things to look for here are **critical incidents**, places on your Web site where participants are confused, frustrated, or even swear. Critical incidents also include cases in which people are pleasantly surprised or say something positive about the site.

In contrast, **bottom-line data** consists of formal, quantitative measurements of what happened, such as the time it takes to complete a task, the number of errors that occur, or the time it takes to learn a task.

In general, you should focus on getting process data first because it gives a good overview of where the problems in a Web site are and because it is easier to get. Process data can also be obtained from low-fidelity paper prototypes, making it a handy technique for the early stages of design.

It takes more work to get and make use of bottom-line data. One reason is that you need to have lots of participants to get statistically reliable results. Another is that bottom-line data does not always tell you what problems need to be fixed, it just tells you that people are going too slowly or are making too many errors. Bottom-line data is better for later phases of design, when you're tuning the performance of an existing Web site. It is also better for comparing two Web site designs, to show that one is superior to another in a particular measurable aspect. Such a comparison can be especially important when you're trying to convince management to make either a major change or a change on an important page, such as the homepage of a high traffic site.

Setting Up the Tasks   A.2

The next step is to choose several representative tasks. By this we mean realistic tasks that your target customers are likely to do on your Web site. Choose some tasks that are simple, some that are of medium difficulty, and some that are hard. Ideally, these tasks will have already been

worked out in the task analysis you carried out when learning to know your customers and can just be taken from there (see Chapter 3—Knowing Your Customers: Principles and Techniques).

3

Simple Tasks Are Short and Performed Often • Simple tasks include things like "Find the latest news article about parenting" or "Find the phone number and e-mail address of the help desk." Success on simple tasks is a binary result: The person either succeeds or fails.

Tasks of Medium Difficulty Are a Little Longer and Harder than Simple Tasks • Examples of medium-difficulty tasks include "Purchase the cheapest printer you can find," "Print out a list of all your previous purchases," and "Add a message to the gourmet cooking community board." These tasks span a few Web pages, but they are reasonable things that people would do. Some medium-difficulty tasks will have binary success metrics; that is, they will either succeed or they will fail. The results of other medium-difficulty tasks will be more open-ended and require further interpretation of the results.

Hard Tasks Span Many Web Pages and Are Fairly Involved • Examples of hard tasks include "Make the Web site show you only the stocks you are interested in," "Buy a digital camera for a friend that he or she will like," and "Buy a toy for your friend's one-year-old child." Most hard tasks are free-form, so it will take some judgment to determine how successful participants are.

Tasks Should Be about What People Want to Do • Be careful not to tell people *how* to do the task. For example, instead of saying, "Go to 'My profile' and find your previous purchases," you should say something like, "Find all of your previous purchases." Again, the task should be worded in the way people would ordinarily think about the problem—that is, *what,* not *how.* Another example of careful wording is, "Make the Web site show you only the stocks that interest you." Not as realistic would be a task like "Customize your profile to show you the stocks that interest you," because the words *customize* and *profile* are not likely to be part of people's regular vocabulary. Another reason is that it might lead people on, especially if there are links labeled "customize" or "profile."

Tasks Should Be Realistic • For example, "Create a new customer account" is something that many people do on a Web site, but not because they

want to. People create an account only because they have to, to get something else done. In other words, creating an account is more of a secondary task that people do to accomplish a primary task.

"Buy a digital camera for a friend that he or she will like" is very open-ended, but it is likely to be the way people approach the problem. It is important that tasks be realistic because you want to find out what people are thinking and see if the design provides the right cues to support them.

Tasks Should Form a Complete Story • Taken as a whole, the tasks should be complete, forming a cohesive and believable story. For example, it does not really make sense if the tasks are given in the following order: "Find previous purchases," "Add a message to a community board," and then "Find the privacy policy." The tasks need to flow together. For example, the following order makes more sense: "Find the privacy policy," "Purchase a printer," and then "Purchase additional ink toner cartridges."

Also be careful not to fragment tasks: "Purchase the best printer for under $300" makes more sense than (1) "Create an account," (2) "Find and compare printers for under $300," and (3) "Purchase the printer you found." Testing fragmented tasks may show that customers can complete the subtasks just fine, but when they are put together in a more realistic situation, the results may not be nearly as good.

The number of tasks to test depends on how extensively you want to test your Web site. Five to ten tasks is about right for most cases—enough to cover a lot of functionality without taking a lot of time for each participant.

Recruiting Participants A.3

After defining some tasks, you should begin recruiting participants. These participants need to be representative of eventual customers in terms of vocabulary, general knowledge, and desired tasks. If the Web site is aimed toward college students, then advertise at a nearby college. If the Web site is for mothers of young children, then get friends of friends who are also mothers or advertise with local mom's groups.

Avoid Friends and Family • One thing to avoid is getting close friends or family to help out unless you're sure that they will give honest feedback. They may be reluctant to criticize something that you have worked so hard on. Also do not get coworkers from down the hall. They are likely to

know too much about what you are doing. It's OK to use people like this for a first pass, as a way of getting quick comments on a design and piloting your experimental procedures, but do not rely solely on feedback from these tests. Again, get people who would realistically use the Web site.

Buy Participants' Time with Gifts and Prizes • One way to recruit people is to compensate them for their time. You might be surprised what some people will do for a free T-shirt.[2] Some other ways of drawing in people include giving small toys, coffee mugs, gift certificates, or some money, or giving a large cash prize of $200 to $300 to the participant who "does the best." This last type of compensation works well for experiments where creative performance is important. For straight cash payments, we normally offer about $20 per hour for university students and about $50 per hour for other participants.

  If you don't have the time to recruit participants, several market research firms can recruit participants who meet the profile you need for about $100/participant, not including the compensation you must pay each participant. Many usability practitioners and designers go this route, although this approach may double your direct costs for running the tests.

Getting the Right Number of Participants • You do not need many participants to get process data. If you are in the early stages of design, five or six people will be fine, especially for paper prototypes. You will need more people—often about 10 to 20 participants—in the later stages of design to evaluate the site. However, you will need to increase these numbers if you have a large and diverse audience to cover, or if your Web site is very large. Getting so many people right at the outset might seem expensive, but consider how much trouble this investment will save you later when you have created a more useful and usable Web site for your customers.

  Getting bottom-line data requires a lot more people. Ten to 20 people can provide initial data, but most tasks will still have a large amount of variability. Section A.5, Analyzing the Data, will give more details about the relationship between the number of people and variability in the data.

  When you're recruiting participants, get a few more people than are really needed. The first few tests you run may be a little rough, and you may have to make some changes to make the evaluation flow smoothly. In addition, not everyone remembers to show up.

----

2 Then again, if you've been working in the computer industry, you might not!

There are two things you should do when recruiting people. First, give them a general overview of the experiment, describing what the Web site is about, what they will do, and approximately how long the whole thing will take. Do not provide too many details because you do not want to bias the test. Second, tell them about any prizes or compensation that will be given for participating. If a person agrees to be a participant, schedule a time and place for the test, and then get his or her name and either a phone number or an e-mail address so that you can provide a reminder before the test.

Choosing Between-Groups versus Within-Groups Experimental Design • One important consideration in experimental design is whether each participant participates in more than one experimental condition. Say, for example, that you're testing two versions of a Web site to compare them. In this scenario, there are two experimental conditions. In a **between-groups experiment**, you break your pool of test participants into two groups and each group uses only one of the Web sites. In contrast, in a **within-groups experiment** you have only one group of test participants, and each participant uses both sites.

These two types of experimental design have trade-offs. For example, a within-groups experiment may not require as many test participants before producing statistically significant results. If you're after bottom-line data, the within-groups approach can save you considerable time and money. On the other hand, a within-groups experiment can raise issues of validity if learning effects are involved. For example, if you test the same tasks on two versions of the same Web site, your participants might be quicker completing a task the second time because they learned how to do it on the first site. You can alleviate some of these problems by randomizing or counterbalancing the order of sites tested and other experimental conditions.

In general, within-groups experiments work better when a low-level interaction technique is being tested, such as finding the best position for a particular button on the page. Use between-groups experiments when you want to compare tasks on two versions of a site or between two competitive sites. Try to make sure that the participants in the two groups match as well as possible in terms of demographics, Internet experience, and familiarity with the problem domain.

A.4 # Running the Test

Several considerations about the test itself are important—from where you run it to what you say to the participants. In this section we look at these issues.

Setting Up the Test Location

If you are evaluating a paper prototype, you can conduct the test practically anywhere. All you need is a large table and places for everyone to sit. For online prototypes, the testing location just needs to be a quiet place with a networked computer.

Video cameras and audio recorders are useful to have in both cases, but they are not required. You can accomplish some tests simply by taking notes on paper, though audio and video recordings make it easy to clarify specific issues later. In contrast, some companies have special rooms for testing, complete with expensive recording equipment, eye-tracking devices, and one-way mirrors for observers. These kinds of setups are useful for gathering bottom-line data but are not necessary for process data.

Ethical Considerations

Tests can be a grueling experience for some people. Participants have been known to leave in tears, embarrassed by their mistakes or their inability to complete the tasks successfully. You have a responsibility to alleviate these kinds of problems. One way is to avoid pressuring people to participate. You need to get participants' informed consent regarding the subject matter of the test, and then make it clear that the test is voluntary and that participants can stop the test at any time for any reason (see Appendix C—Sample Consent Form). You also need to stress that you're testing the Web site and not the participants themselves, and that they are really helping you by finding problems with the site. If they are having problems, then it is the Web site's fault, not theirs.

If other people will see the collected data, then the data should also be made as anonymous as possible. Names and other pieces of identifying information should be removed, and people's faces should be blurred out in any pictures and video footage. In some cases a videotape of a person struggling with a human–computer interface has been played for an audience that included that very person! Making the data anonymous will help prevent any potentially awkward situations in the future.

Test Roles

The key role in running a usability test is the facilitator. The facilitator greets participants, introduces any other people in the room, explains the procedure for the test, and answers participants' questions.

The other people act simply as observers, watching what participants do. Their role is to take notes and keep quiet. Observers can also be remote if the setup allows them to view things from another location or through a two-way mirror.

If you're running a test on a paper prototype, another role you'll need is the computer. The job of the person playing computer is to run the interface, updating the paper interface as needed.[3]

Running a Pilot Test

Before running the tests with actual participants, you should carry out a pilot test with two or three people. In this case, coworkers and friends are OK. The key is to get used to the procedure of running a test and to work out any bugs in your procedure. A pilot also helps you figure out how long the test will take so that you know whether you need to cut or possibly add more tasks.

After you have finished the pilot tests, try analyzing the collected data. This data should not be used in the final analysis, but you should do the analysis to make sure that you are collecting the right data. For example, once when we were evaluating a Web site, we asked people to sort a list of features according to importance. Although the exercise was useful, it turned out that the data we had gathered was extremely difficult to analyze properly. One person mentioned that only the top two things in the list were really important to her; another identified the top four items as important. In retrospect, a better way of gathering this information would have been to ask people to rate the importance of each feature from 1 (not important) to 7 (very important). If we had tried analyzing data from the pilot test, we would have caught the problem before conducting the real test.
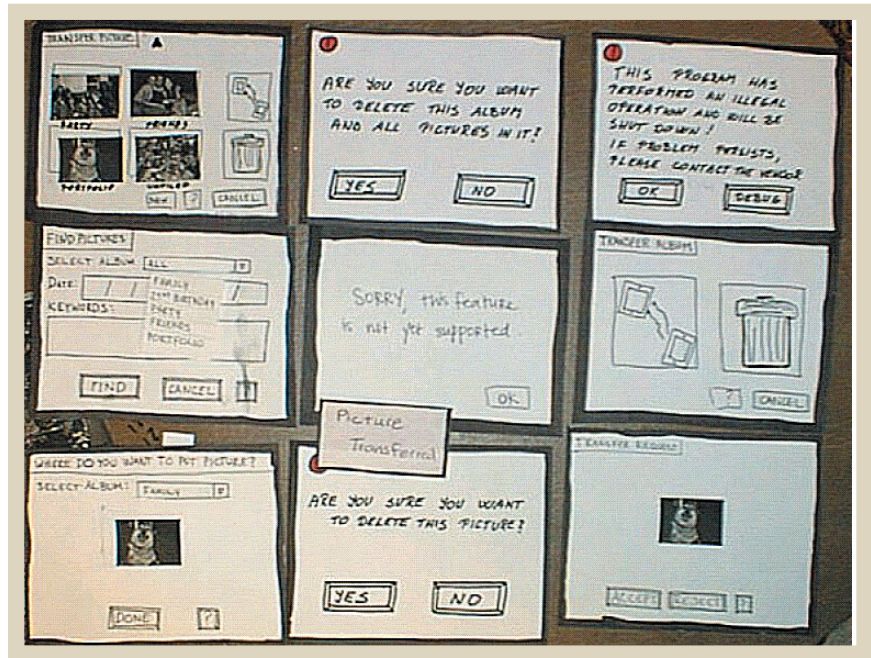
Testing Paper Prototypes

Paper prototypes are useful for obtaining data early in the process, but you should not use them for bottom-line data because they are too far removed from the final implementation. Most people have not seen

---

3 A secondary job is to just smile at all the bad jokes about being slow and needing to upgrade the computer.

Figure AppA.1

It is easier to run usability tests with oversized paper prototypes because everyone can see what's happening.

AppA.1

paper prototypes before. You will have to explain the concept, but the majority of people catch on pretty quickly.

Ask participants to point at things with their finger, using it as a mouse. If they click on a link, the person playing the computer just switches to another piece of paper representing the next page. If they click on a drop-down menu, the computer can place an index card with the choices on top. Note that having a paper prototype that is larger than it would be in reality makes it easier for everyone to see what the participant is pointing at (see Figure AppA.1).

It is difficult to simulate highly interactive elements such as mouse roll-overs and animations with paper prototypes. In most cases, this is a good thing because it forces design teams to focus on the core issues first. Be aware that this is a limitation of paper prototypes, and plan accordingly.

Testing Online Prototypes

Online computer–based prototypes can be used for obtaining either process data or bottom-line data. If you're testing a high-fidelity prototype, you need to make it clear to the participants that they will be testing an

early design and not the final Web site. They might mistakenly believe that the Web site is nearly done when it is really in the early stages of design. Setting their expectations properly will help them give you the type of high-level feedback you need at this stage, rather than comments on the visuals, such as colors and fonts. Later you can test again to evaluate these details.

Before starting a test, be sure to clear the Web browser's history and cache so that it will be as if the person had never been to the Web site before, making all the links unvisited.

Starting and Carrying Out the Test

Greet the Participant • Tests are generally broken into three major phases: preliminary instructions and paperwork, performing the task, and a debriefing. Start by introducing yourself and the rest of the team. Then describe the purpose of the test at a high level, and be sure to emphasize that you're testing the Web site and that you're not testing them in any way. Say something like, "We're asking you to help us improve the Web site by helping us find problems with it. We're testing the Web site and not you."

Also make it clear that you will not provide help as they go through the tasks because you want to see how they would go through the Web site normally. However, emphasize that it's all right for them to stop the test at any time for any reason.

This is also a good time to put a "Do Not Disturb" sign on the door saying that there is a customer research study in progress. You do not want any interruptions (unless this is one of the things you want to observe).

Fill Out the Paperwork • After greeting the participants, have them fill out any paperwork you may have. The paperwork may request such things as basic demographic information, a name and address to which you can send a check if you're paying them, and consent. Any consent forms should explain what the test is, what kinds of data will be collected, and how the data will be used. Make sure you have two copies of the consent forms—one for you and one for the participant to keep.

Ask Participants to Think Aloud • If you're gathering process data, ask the participants to think aloud, to say what they're looking for and what they're trying to do. Although some people are really good at this, others find it a little awkward. The facilitator should prompt participants every so often if they stop talking, asking things like, "So what are you looking for now?" or "What are you trying to do now?"

Do not do this if you are collecting bottom-line data because thinking aloud may cause participants to make more errors or to go through the Web site more slowly.

Instruct the Participants How to Start • Ask the participants if they have any questions before starting. Then hand them any instructions you may have, any special information, such as a fake credit card number to use, and the first task to complete. Ask them to read the task aloud. Doing this will help them start thinking aloud.

You may want to have participants fill out a very short survey after completing each task. You can ask questions like how easy or hard they thought the task was. You can also ask questions to make sure that they found the right piece of information. For example, if the task is to find and add a specific item to the shopping cart, you can ask them how much the item cost. This is just a redundant check, to make sure that they really did complete the task successfully.

Take Good Notes during Each Task • The observers should be taking notes about what each participant says and does during the test. It also helps to record audio and video if possible. Use a digital watch or a clock to keep track of time, too. If a certain task takes far too long, tell the participant that it's OK to move on to the next task.

If you're measuring bottom-line data, make sure everyone knows what to measure. For example, is it an error if someone hits the **Back** button on the browser? Is it an error if someone goes back to the homepage? The criteria need to be agreed on beforehand. And what happens if someone

---

Some Common Mistakes in Running Usability Tests

- Testing a Web site using unrealistic tasks
- Using significantly leading or biased tasks when comparing Web sites
- Recruiting participants who do not represent your expected customers
- Forgetting to clear out the Web browser file cache and history list before starting
- Using only a computer that has a fast network connection, a high resolution monitor, and a fast processor (unless all or most of your customers really do have these)

does not finish a task? There are no hard rules here, but a common technique is to throw out the data for that participant with a clarifying note in the final report or to assign a very large time and a large number of errors, just to keep everything numerical.

Watch Closely • Yes, it will be frustrating to watch people struggle with something you put so much time into, clicking on the wrong link or not seeing the text right in front of them. But bite your lip and keep your mouth closed: You are here to watch and to learn how to improve the Web site. Make sure that none of the observers laugh, groan, or make any other inappropriate response. These are the types of things that can unnerve your participants.

If a participant does something really interesting, ask a follow-up question. Ask open-ended questions, such as, "What are you looking for?" Let the participants know that things are going all right. Prompt them to keep speaking and tell you what they're thinking. Also look out for nonverbal cues, such as a furrowed brow or a puzzled look.

Answer any general questions that participants may have, but do not help them with the tasks. Also do not help some participants more than others. Plan in advance what you will and will not help with. For example, it is common to decide that you will help participants when they run into known bugs or functionality that has not been implemented yet. Simply get them back on track.

Follow Up with a Quick Survey • After all the tasks have been completed, follow up with a short survey. You want to get your participants' overall impressions and comments about the Web site, seeing what they liked and disliked about the Web site. Also ask them where they felt they had problems with the site and where they thought it worked well.

Debrief the Participants after the Test • Wrap up by debriefing each participant, telling them what you were looking for, as well as discussing any interesting behavior the participant had. People often do not remember specific actions, so it may be useful to go through the Web site again or to show video segments to help prompt their memory.

Ask participants if they have any thoughts on how to fix any problems they encountered. Take these comments with a grain of salt because participants usually do not have an understanding of design or the underlying technology. Nevertheless, these comments are useful to hear. Afterward, finish up by asking if they have any final questions, and then thank them for their time.

A.5   Analyzing the Data

Analyzing Process Data • Think about what you saw and what the partici-
pants said. Did they understand the things you thought they would?
Were they confused by any terms or concepts? If so, maybe things need
K11 to be renamed to use FAMILIAR LANGUAGE (K11) or explained in greater
detail. If the concept in question is a concept fundamental to the Web
site, be sure to make that fact clear on the homepage because otherwise
people might leave without ever bothering to figure it out.

What errors could they recover from? For example, did they click on a
link but then quickly realize that it was the wrong one? It is important to
minimize these kinds of problems, but these are usually just minor
annoyances. A bigger problem would be indicated by systematic "ping-
ponging"—that is, repeated back and forth attempts from one page down
unfruitful paths. Such behavior would suggest a need for more DESCRIP-
K9 TIVE, LONGER LINK NAMES (K9), which would give the participants more
"information scent" to find the page they're looking for.

Focus first on the errors from which participants could not recover. Did
participants have problems finding items on the Web site? Did they have
trouble understanding the overall structure of the Web site? What about
navigation? Could they make their way through the site adequately? Did
they make any errors and not even notice that there was a problem?[4]
These could be fundamental problems of the site and should be addressed
first when you're fixing the site.

The most important question to ask is why the error occurred. Was the
navigation too confusing, making it difficult to go to other pages? Was the
information disorganized, making it hard to find things across pages? Was
the Web page too cluttered, making it hard to find anything on a page?
Was the site too slow, causing participants to lose track of what they
wanted to do? Just like a doctor, you get to see only the symptoms, but
you need to keep asking yourself if any fundamental issues are causing all
of these problems.

Another thing to keep in mind is that people do not give up in usability
tests as easily as they would in the real world. You have to realize that no
matter what you do, you are still putting an implicit amount of pressure
on participants to try their best to successfully complete the task. People

---

4   A perfect example of this is the infamous butterfly ballot used in the 2000 U.S. presiden-
tial election. Statistical analysis of neighboring counties suggests that at least several
thousand citizens unintentionally voted for the wrong candidate, but the voting process
lacked a verification process that would let people check their votes.

are more attentive and willing to go through a few more pages when they know they are being observed.

Analyzing Bottom-Line Data • Be careful when analyzing bottom-line data. For example, suppose the target goal is to ensure that a person new to a Web site can find and purchase an item in 20 minutes or less. When running our test, we get times of 20, 15, 45, 10, 5, and 25 for our six participants. The mean or average time for this is 20 minutes. Looks pretty good! The median for this set of numbers is 17.5—even better!

However, the problem is that there is very little certainty here because there are only six participants and the results are highly variable. If you calculate the standard deviation, a measure of how variable the numbers in this set are, you will find that the value is approximately 14. If we divide the standard deviation by the square root of the number of samples we have (6), we get 5.8. This is the *standard error of the mean,* and it tells us how much variation we can expect in the typical value. It is plausible that the typical value is as small as the mean minus twice the standard error of the mean, resulting in a lower bound of 8.5, or as large as the mean plus twice the standard error of the mean, or 32. This latter value would clearly be far from our stated goal of 20 minutes!

We can say more precisely what we mean by *plausible*. The best thing to do here is to use statistical techniques. Cranking through basic statistical methods, you can calculate with 95 percent confidence that the actual average time will be 20, plus or minus 11, minutes. In other words, you are 95 percent likely to be correct in saying that the actual time will be in this range, but 5 percent of the time you'll be wrong.[5]

Usability test data is often quite variable, which means that you need lots of participants to get good estimates of typical values. In addition, the breadth of range depends on the square root of the number of participants. In other words, if you have 4 times as many participants, you narrow the range by an average factor of only 2. Continuing the example, in general, quadrupling the number of participants from 6 to 24 will narrow the spread of the average time from 20, plus or minus 11, minutes to 20, plus or minus 6, minutes (assuming that the mean and the standard deviation stay about the same). This is where online usability evaluation methods, as described in Appendix E—Online Research, become useful

---

5   If you use Microsoft Excel, you can calculate this range using the CONFIDENCE function. If you use a more advanced tool, like SPSS, you probably already know how to calculate this.

because they make it easier to scale up the number of participants and thus tighten your confidence intervals.

Basic statistics is beyond the scope of this book, but a great introduction on the topic is *The Cartoon Guide to Statistics* by Larry Gonick, Jr., and Woolcott Smith. This book covers the main concepts that you will want to be familiar with when doing basic statistical analyses, including mean, variance, standard deviation, correlation, regression, *t*-test, and ANOVA.

A.6   Presenting the Results

After the data has been collected and analyzed, the results need to be presented to the design team or to the clients. Results can be reported in the form of a written report or an oral presentation. Here's a short outline of the sections your report should include:

- Executive Summary
- Tasks
- Participants
- Problems Found
- Participant Feedback
- Suggested Improvements
- Appendices

Start with the Executive Summary, which gives a quick overview of what you did in the test, a summary of the results, and a rundown of the recommendations for improvement. Next, in Tasks, talk about the tasks that you had participants carry out, describing why these tasks were chosen. Continue with Participants, a short description of the number of participants, general demographics, and any defining characteristics.

In the next section, Problems Found, list the problems encountered, prioritized by severity. Use screen shots of problem Web pages, using circles and arrows to point out critical incidents. Graphs showing the success rates of participants at completing tasks will also help people understand the results. If you're presenting the results orally, this is a good time to show video clips to help convey your message. Video is extremely valuable for convincing skeptical programmers and management that there are problems with the Web site. You can also include video clips in written reports that you plan to put online.

The Participant Feedback section contains both positive and negative feedback from participants. This section can include summaries of surveys taken by participants after they finished the test, or direct quotes from them during the test.

The next section, Suggested Improvements, outlines what needs to be changed to improve the Web site. The improvements should be triaged into "must do," "should do," and "could do" categories. The "must do" improvements are the show stoppers, the ones that caused serious problems from which people could not recover. They also include really simple improvements that take only a short time to fix, such as misspellings or broken links. The "should do" improvements represent problems that are annoying but tolerable—problems that most people can figure out. The "could do" improvements are changes that will take too much effort to implement for the resulting benefits. Keep these ideas on the back burner for the next iteration.

The last section, Appendixes, contains any test materials used during the experiment, such as demo scripts and instructions, as well as all of the raw data in a cleaned-up form.

Your evaluation plan can often be used as the basis for your usability test report. See Appendix B—Sample Web Site Evaluation Plan.