CHAPTER

Planning a Usability Study

3

The Boy Scouts' motto is "Be Prepared." This is true not only when you head off into the wilderness but also when you collect usability data. Preparation is the key to any successful usability study. If nothing else, we hope this chapter convinces you to plan ahead when running a usability study, particularly where data collection is involved.

A few high-level questions must be answered when planning any usability study. First, you need to understand the goals of the study. For example, are you trying to ensure optimal usability for a new piece of functionality, or are you benchmarking the user experience for an existing product? Next, you need to understand the goals of the users. Are users looking to simply complete a task and then stop using the product, or will they use the product many times on a daily basis? Knowing both the study goals and the user goals will lead toward choosing the right metrics.

Many practical details come into play as well. For example, you must decide on the most appropriate evaluation method, how many participants are enough to get reliable feedback, how collecting metrics will impact the timeline and budget, and how the data will be collected and analyzed. By answering these questions, you will be well prepared to carry out any usability study involving metrics. In the end, you will likely save time and money and have a greater impact on the product.

3.1 STUDY GOALS

The first decision to make when planning a usability study is how the data will ultimately be used within the product development life cycle. There are essentially two general ways to use data: formative and summative.

3.1.1 Formative Usability

When running a formative study, a usability specialist is much like a chef who periodically checks a dish while it's being prepared and makes adjustments to positively impact the end result. The chef might add a little salt, then a few more

spices, and finally a dash of chili pepper right before serving. The chef is periodically evaluating, adjusting, and reevaluating. The same is true in formative usability. A usability specialist, like a chef, periodically evaluates a product or design, identifies shortcomings, makes recommendations, and then repeats the process, until, ideally, the product comes out as close to perfect as possible.

What distinguishes formative usability is the iterative nature of the testing. The goal is to make improvements in the design. This means identifying or diagnosing the problems, making recommendations, and then evaluating again. Formative usability is always done before the design has been finalized. In fact, the earlier the formative evaluation, the more impact the usability evaluations will have on the design.

Here are a few key questions you will be able answer with a formative approach:

- What are the most significant usability issues that are preventing users from completing their goals or that are resulting in inefficiencies?
- What aspects of the product work well for the users? What do they find frustrating?
- What are the most common errors or mistakes users are making?
- Are improvements being made from one design iteration to the next?
- What usability issues can you expect to remain after the product is launched?

The most appropriate situation to run a formative usability study is when an obvious opportunity to improve the design presents itself. Ideally, the design process allows for multiple usability evaluations. If there's no opportunity to impact the design, then running a formative test is probably a waste of time. Generally, though, selling the value of formative usability shouldn't be a problem. Most people will see the importance of it. The biggest obstacles tend to be a limited budget or time rather than a failure to see the value.

3.1.2 Summative Usability

Continuing with our cooking metaphor, summative usability is about evaluating the dish after it comes out of the oven. The usability specialist running a summative test is like a food critic who evaluates a few sample dishes at a restaurant or perhaps compares the same meal in multiple restaurants. The goal of summative usability is to evaluate how well a product or piece of functionality meets its objectives. Summative testing can also be about comparing several products to each other. Although formative testing focuses on identifying ways of making improvements, summative testing focuses on evaluating against a certain set of criteria. Summative usability evaluations answer these questions:

- Did we meet the usability goals of the project?
- How does our product compare against the competition?
- Have we made improvements from one product release to the next?

Running a successful summative usability test should always involve some follow-up activities. Just knowing is usually not enough for most organizations. Potential outcomes of a summative usability test might be securing funding to enhance functionality on your product, or launching a new project to address some outstanding usability issues, or even benchmarking changes to the user experience against which senior managers will be evaluated. We recommend that follow-up actions be planned along with any summative usability study.

3.2 USER GOALS

When planning a usability study, you need to know something about the users and what they are trying to accomplish. For example, are users forced to use the product every day as part of their job? Are they likely to use the product only once or just a few times? Are they using it frequently as a source of entertainment? It's critical to understand what matters to the user. Does the user simply want to complete a task, or is their efficiency the primary driver? Do users care at all about the design aesthetics of the product? All these questions boil down to measuring two main aspects of the user experience: performance and satisfaction.

3.2.1 Performance

Performance is all about what the user actually *does* in interacting with the product. It includes measuring the degree to which users can successfully accomplish a task or set of tasks. Many measures related to the performance of these tasks are also important, including the time it takes to perform each task, the amount of effort to perform each (such as number of mouse clicks or amount of cognitive effort), the number of errors committed, and the amount of time it takes to become proficient in performing the tasks (learnability). Performance measures are critical for many different types of products and applications, especially those where the user doesn't really have much choice in how they are used (such as a company's internal applications). If users can't successfully perform key tasks when using a product, it's likely to fail. Chapter 4 reviews different types of performance measures.

3.2.2 Satisfaction

Satisfaction is all about what the user *says* or *thinks* about his interaction with the product. The user might report that it was easy to use, that it was confusing, or that it exceeded his expectations. The user might have opinions about the product being visually appealing or untrustworthy. User satisfaction has many different aspects. Satisfaction, and many other self-reported metrics, is important for products where the user has some choice in their usage. This would certainly be true

for most websites, software applications, and consumer products. Satisfaction metrics are reviewed in Chapter 6.

DO PERFORMANCE AND SATISFACTION ALWAYS CORRELATE?

Perhaps surprisingly, performance and satisfaction don't always go hand-in-hand. We've seen many instances of a user struggling to perform key tasks with an application and then giving it glowing satisfaction ratings. Conversely, we've seen users give poor satisfaction ratings to an application that worked perfectly. So it is important that you look at both performance and satisfaction metrics to get an accurate overall picture of the user experience.

3.3 CHOOSING THE RIGHT METRICS: TEN TYPES OF USABILITY STUDIES

You should consider many issues when choosing metrics for a usability study, including the goals of the study and the user, the technology that's available to collect and analyze the data, and the budget and time you have to turn around your findings. Because every usability study has unique qualities, we can't prescribe the exact metrics to use for every type of study. Instead, we've identified ten prototypical categories of usability studies and developed recommendations about metrics for each. The recommendations we offer are simply suggestions that you should consider when running a usability study with a similar set of characteristics. Conversely, metrics that may be essential to your study may not be on the list. Also, we strongly recommend that you explore your raw data and develop new metrics that are meaningful to your project goals.

Ten common usability study scenarios are listed in Table 3.1. (Many of the usability metrics in Table 3.1 are discussed in detail later in the book.) The metrics that are commonly used or are appropriate for each of the usability study scenarios are indicated. The sections that follow discuss each of the ten scenarios.

3.3.1 Completing a Transaction

Many usability studies are aimed at making transactions run as smoothly as possible. These might take the form of a user completing a purchase, registering a new piece of software, or selling a stock. A transaction usually has a well-defined beginning and end. For example, on an e-commerce website, a transaction may start when a user places something in his shopping cart and ends when he has completed the purchase on the confirmation screen.

Perhaps the first metric that you will want to examine is *task success.* Each task is scored as a success or failure. Obviously the tasks need to have a clear end-state, such as reaching a confirmation that the transaction was successful.

Table 3.1 Ten Common Usability Study Scenarios and Their Most Appropriate Metrics

Usability Study Scenario	Task Success	Task Time	Errors	Efficiency	Learn-ability	Issues-Based Metrics	Self-Reported Metrics	Behavioral and Physiological Metrics	Combined and Comparative Metrics	Live Website Metrics	Card-Sorting Data
1. Completing a transaction	X			X		X	X			X	
2. Comparing products	X			X			X		X		
3. Evaluating frequent use of the same product	X	X		X	X		X				
4. Evaluating navigation and/or information architecture	X		X	X							X
5. Increasing awareness							X	X		X	
6. Problem discovery						X	X				
7. Maximizing usability for a critical product	X		X	X							
8. Creating an overall positive user experience							X	X			
9. Evaluating the impact of subtle changes										X	
10. Comparing alternative designs	X	X				X	X		X		

Reporting the percentage of participants who were successful is an excellent measure of the overall effectiveness of the transaction. If the transaction involves a website, some *live website metrics*, such drop-off rate from the transaction, can be very useful. By knowing where users are dropping off, you will be able to focus your attention on the most problematic steps in the transaction.

Calculating *issue severity* can help narrow down the cause of specific usability problems with a transaction. By assigning a severity to each usability issue, you will be able to focus on the high-priority problems with any transaction. Two types of *self-reported metrics* are also very useful: likelihood to return and user expectations. In cases where users have a choice of where to perform their transactions, it's important to know what they thought of their experience. One of the best ways to learn this is by asking participants whether they would use the same product again and whether the product met or exceeded their expectations. *Efficiency* is an appropriate metric when a user has to complete the same transaction many times. Efficiency is often measured as task completion per unit of time.

3.3.2 Comparing Products

It's always useful to know how your product compares to the competition or to previous releases. By making comparisons, you can determine your product's strengths and weaknesses and whether improvements have been made from one release to another. The best way to compare different products or releases is through the use of various metrics. The type of metrics you choose should be based on the product itself. Some products aim to maximize efficiency, whereas others try to create an exceptional user experience.

For most types of products, we recommend three general classes of metrics to get an overall sense of usability. First, we recommend looking at some task success measures. Being able to complete a task correctly is essential for most products. It's also important to pay attention to efficiency. Efficiency might be task completion time, number of page views (in the case of some websites), or number of action steps taken. By looking at efficiency, you will get a good sense of how much effort is required to use the product. Some self-reported metrics of satisfaction provide a good summary of the user's overall experience. Satisfaction measures make the most sense with products where people have choices. Finally, one of the best ways to compare usability across products is by *combined and comparative metrics*. This will give an excellent big picture of how the products compare from a usability perspective.

3.3.3 Evaluating Frequent Use of the Same Product

Many products are intended to be used on a frequent or semifrequent basis. Examples might include microwave ovens, DVD players, web applications used as part of your job, and even the software program we used to write this book. These products need to be both easy to use and highly efficient. The amount of

effort required to burn a DVD or *not* burn popcorn in a microwave needs to be kept to a minimum. Most of us have very little time or patience for products that are difficult to use.

The first metric we would recommend is task time. Measuring the amount of time required to complete a set of core tasks will reveal the effort involved. For most products, the faster the completion time, the better. Because some tasks are naturally more complicated than others, it may be helpful to compare task completion times to expert performance. Other efficiency metrics such as the number of steps or page views (in the case of some websites) can also be helpful. The time for each step may be short, but the separate decisions that must be made to accomplish a task can be numerous.

Learnability metrics assess how much time or effort is required to achieve maximum efficiency. Learnability can take the form of any of the previous efficiency metrics examined over time. In some situations, consider self-reported metrics, such as awareness and usefulness. By examining the difference between users' awareness and perceived usefulness, you will be able to identify aspects of the product that should be promoted or highlighted. For example, users may have low awareness for some parts of the product, but once they use it, they find out it is extremely useful.

3.3.4 Evaluating Navigation and/or Information Architecture

Many usability studies focus on improving the navigation and/or information architecture. This is probably most common for websites, software programs, or consumer electronics. It may involve making sure that users can quickly and easily find what they are looking for, easily navigate around the product, know where they are within the overall structure, and know what options are available to them. Typically, these studies involve the use of wire-frames or partially functional prototypes because the navigation and information mechanisms and information architecture are so fundamental to the design that they have to be figured out before almost anything else.

One of the best metrics to evaluate navigation is task success. By giving participants tasks to find key pieces of information (a ''scavenger hunt''), you can tell how well the navigation and information architecture works for them. Tasks should touch on all the different areas of the product. An efficiency metric that's useful for evaluating navigation and information architecture is *lostness,* which looks at the number of steps the participant took to complete a task (e.g., web page visits) relative to the minimum number to complete the task.

Card sorting is a particularly useful method to understand how participants organize information. One type of card-sorting study is called a closed sort, which has participants put items into predefined categories. A useful metric to come from a closed card sort study is the percentage of items placed into the correct category. This metric indicates the intuitiveness of the information architecture.

3.3.5 Increasing Awareness

Not every design that goes through a usability evaluation is about making something easier or more efficient to use. Some design changes are aimed at increasing awareness of a specific piece of content or functionality. This is certainly true for online advertisements, but it's also true for products that have important but underutilized functionality. There can be many reasons why something is not noticed or used, including some aspect of the visual design, labeling, or placement.

First, we recommend monitoring the number of interactions with the element in question. This is not foolproof, since a participant might notice something but not click on it or interact with it in some way. The opposite would not be very likely: interaction without noticing. Because of this, the data can help confirm awareness but not demonstrate lack of awareness. Sometimes it's useful to ask for self-reported metrics about whether the participants noticed or were aware of a specific design element. Measuring noticeability involves pointing out specific elements to the participants and then asking whether they had noticed those elements during the task. Measuring awareness involves asking the participants if they were aware of the feature before the study began. However, it's unclear how reliable these data are. Not everyone has a good memory, and some people try to save face and say they saw something when they didn't. Therefore, we don't recommend that this be your sole measure; you should complement it with other data sources.

Memory is another useful self-reported metric. For example, you can show participants several different elements, only one of which they had actually seen previously, and ask them to choose which one they saw during the task. If they noticed the element, their memory should be better than chance. But perhaps the best way to assess awareness, if you have the technology available, is through the use of *behavioral and physiological metrics* such as eye-tracking data. Using eye-tracking technology, you can determine the average time spent looking at a certain element, the percentage of participants who looked at it, and even the average time it took to first notice it. Another metric to consider, in the case of websites, is a change in live website data. Looking at how traffic patterns change between different designs will help you determine relative awareness. Simultaneous testing of alternative designs (A/B testing) on live sites is an increasingly common way to measure how small design changes impact user behavior.

3.3.6 Problem Discovery

The goal of problem discovery is to identify major usability issues. In some situations you may not have any preconceived ideas about what the significant usability issues are with a product, but you want to know what annoys users. This is often done for a product that is already built but has not gone through usability evaluation before. A problem discovery study also works well as a periodic checkup to get back in touch with how users are interacting with your product. A discovery study is a little different from other types of usability studies because it is generally

open-ended. Participants in a problem discovery study may be generating their own tasks, as opposed to being given a list of specific tasks. It's important to strive for realism as much as possible. This might involve using the live product and their own accounts and performing tasks that are relevant only to them. It might also include evaluating the product in the participants' environments, such as homes or workplaces.

Because they may be performing different tasks and their contexts of use may be different, comparing across participants may be a challenge. *Issues-based metrics* may be the most appropriate for problem discovery. Assuming you capture all the usability issues, it's fairly easy to convert those data into frequency and type. For example, you might discover that 40 percent of the usability issues pertain to high-level navigation and 20 percent of the issues to confusing terminology. Even though the exact problems encountered by each participant might be different, you are still able to generalize into a higher-level category of issue. Examining the frequency and severity of specific issues will reveal how many repeat issues are being observed. Is it a one-time occurrence or part of a recurring problem? By cataloging all the issues and assigning severity ratings, you may come away with a quick-hit list of design improvements.

3.3.7 Maximizing Usability for a Critical Product

Although some products *strive* to be easy to use and efficient, such as a cell phone or washing machine, other products *have* to be easy to use and efficient, such as a defibrillator machine, voting machine, or emergency exit instructions on an airplane. What differentiates a critical product from a noncritical product is that the entire reason for the critical product's existence is for the user to complete a very important task. Not completing that task will have a significant negative outcome.

Measuring usability for any critical product is essential. Just running a few participants through the lab is rarely good enough. It's important that user performance be measured against a target goal. Any critical product that doesn't meet its target usability goal should undergo a redesign. Because of the degree of certainty you want from your data, you may have to run relatively large numbers of participants in the study. One very important metric is user *errors*. This might include the number of errors or mistakes made while performing a specific task. Errors are not always easy to tabulate, so special attention must be given to how you define an error. It's always best to be very explicit about what constitutes an error and what doesn't.

Task success is also important to measure. We recommend using a binary approach to success in this situation. For example, the true test of a portable defibrillator machine is that someone can use it successfully by themselves. In some cases, you may wish to tie task success to more than one metric, such as completing the task successfully within a specific amount of time and with no errors. Other efficiency metrics are also useful. In the example of the defibrillator machine, simply using it correctly is one thing but doing so in a timely manner is altogether different. Self-reported metrics are relatively less important with critical products. What users think about their use of the product is much less important than their actual success.

3.3.8 Creating an Overall Positive User Experience

Some products strive to create an exceptional user experience. It's simply not enough to be usable. These products need to be engaging, thought-provoking, entertaining, and maybe even slightly addictive. The iPod and TiVo are two such products that come to mind. These are products that you tell a friend about and are not embarrassed to mention at a party. Their popularity usually grows at phenomenal rates. Even though the characteristics of what constitutes a great user experience are subjective, they are still measurable.

Although some performance metrics may be useful, what really matters is what the user thinks, feels, and says with respect to his or her experience. In some ways, this is the opposite perspective of measuring usability of a critical product. If the user struggles a little at first, it may not be the end of the world. What matters is how the user feels at the end of the day. Many self-reported metrics must be considered when measuring the overall user experience.

Satisfaction is perhaps the most common self-reported metric, but it may not always be the best one. Being ''satisfied'' is usually not enough. One of the most valuable self-reported metrics we've used relates to the participant's expectation. The best experiences are those that *exceed* a participant's expectations. When the participant says something is much easier, more efficient, or more entertaining than expected, you know you are onto something.

Another set of self-reported metrics relates to future use. For example, you might ask questions related to likelihood to purchase, recommend to a friend, or use in the future. Another interesting set of metrics relates to subconscious reactions that users may be having. For example, if you want to make sure your product is engaging, you can look at *physiological metrics*. Changes in pupil diameter can be used to gauge the level of arousal, or if you're trying to eliminate stress as much as possible, you can measure heart rate or skin conductance.

3.3.9 Evaluating the Impact of Subtle Changes

Not all design changes have an obvious impact on user behavior. Some design changes are much more subtle, and their impact on user behavior is less clear. Small trends, given enough users, can have huge implications for a large population of users. The subtle changes may involve different aspects of the visual design, such as font choice and size, placement, visual contrast, color, and image choice. Nonvisual design elements, such as subtle changes to content or terminology, can also have an impact on the user experience.

Perhaps the best way to measure the impact of subtle design changes is through *live-site metrics* from A/B tests. A/B testing involves comparing a control design against an alternative design. For websites, this usually involves diverting a portion of web traffic to an alternative design and comparing metrics such as traffic or purchases to a control design. An online usability study with a large population can also be very useful. If you don't have access to the technology to run A/B tests or

online studies, we recommend using e-mail and online surveys to get feedback from as many representative participants as you can.

3.3.10 Comparing Alternative Designs

One of the most common types of usability studies involves comparing more than one design alternative. Typically, these types of studies take place early in the design process, before any one design has been fully developed. (We often refer to these as "design bakeoffs.") Different design teams put together semifunctional prototypes, and we evaluate each design using a predefined set of metrics. Setting up these studies can be a little tricky. Because the designs are often similar, there is a high likelihood of a learning effect from one design to another. Asking the same participant to perform the same task with all designs usually does not yield valuable information, even when counterbalancing design and task order.

There are two solutions to this problem. You can set up the study as purely between-subjects whereby each participant only uses one design, which provides a clean set of data but requires significantly more participants. Alternatively, you can ask participants to perform the tasks using one primary design (counterbalancing the designs) and then show the other design alternatives and ask for their preference. This way you can get feedback about all the designs from each participant.

The most appropriate metrics to use when comparing multiple designs may be issues-based metrics. Comparing the frequency of high-, medium-, and low-severity issues across different designs will help shed light on which design or designs are more usable. Ideally, one design ends up with fewer issues overall and fewer high-severity issues. Performance metrics such as task success and task times can be useful, but because sample sizes are typically small, these data tend to be of limited value. A couple of self-reported metrics are particularly relevant. One is asking each participant to choose which prototype they would most like to use in the future (as a forced choice comparison). Also, asking each participant to rate each prototype along dimensions such as ease of use and visual appeal can be insightful.

3.4 OTHER STUDY DETAILS

Many other details must be considered when planning a usability study. Several important issues to consider are budget/timelines, evaluation methods, participants, and data cleanup.

3.4.1 Budgets and Timelines

The time and cost of running a usability study with metrics depends on the evaluation method, metrics, participants, and available tools. It's impossible for us to give even approximate costs or time estimates for any particular type of

usability study. The best we can do is provide a few general rules of thumb for estimating time and costs for some common types of studies. When making these estimates, we recommend that you carefully consider all the variables that go into any usability study and communicate those estimates to business sponsors (or whoever is funding the study) as early as possible. Also, it's wise to add at least a 10 percent buffer for both time and costs, knowing that there may be some unforeseen costs and delays.

If you are running a formative study with a small number of participants (ten or fewer), collecting metrics should have little, if any, impact on the overall timeline or budget. Collecting and analyzing basic metrics on issue frequency and severity should at most add a few hours to any study. Just allow yourself a little time to analyze the data once the study is complete. If you're not yet very familiar with collecting these metrics, give yourself some extra time to set up tasks and agree on severity ratings prior to starting the test. Because it is a formative study, you should make every attempt to get the findings back to the stakeholders as quickly as possible to influence the next design iteration and not slow down the project.

In the case of running a lab test with a larger number of participants (usually more than a dozen), including metrics may have more of an impact on the budget and timeline. The most significant cost impact may be any additional costs for recruiting and compensating the participants. These costs depend on who they are (e.g., internal to your company versus external) and how participants are recruited. The most significant impact on the timeline is likely to be the additional time required to run the larger number of participants. Depending on your billing or cost-recovery model, there may also be additional costs because of the increased time for the usability specialists. Keep in mind that you will also need extra time to clean up and analyze the data.

Running an online study is quite different in terms of costs and time. Typically, about half of the time is usually spent setting up the study, from identifying and validating tasks, creating questions and scales, evaluating the prototypes or designs, identifying and/or recruiting participants, and developing the online script. Unlike traditional lab tests where a lot of time is spent collecting the data, running an online study requires little, if any, time on the part of the usability specialist for data collection. With most online usability testing technologies you simply flip the switch and then monitor the data as they pour in.

The other half of the time is spent cleaning up and analyzing the data. It's very common to underestimate the time required for this. Data are often not in a format that readily allows analysis. For example, you will need to filter out extreme values (particularly when collecting time data), check for data inconsistencies, and code new variables based on the raw data (such as creating top-2-box variables for self-reported data). We have found that we can run an online study in about 100 to 200 person-hours. This includes everything from the planning phase through data collection, analysis, and presentation. The estimate can vary by up to 50 percent in either direction based on the scope of the study.

3.4.2 Evaluation Methods

One of the great features of collecting usability metrics is that you're not restricted to a certain type of evaluation method (e.g., lab test, online test). Metrics can be collected using almost any kind of evaluation method. This may be surprising because there is a common misperception that metrics can only be collected through large-scale online studies. As you will see, this is simply not the case.

Choosing an evaluation method to collect metrics boils down to how many participants are needed and what metrics you're going to use. The most common usability method is a lab test that requires a relatively small number of participants (typically four to ten). The lab test involves a one-on-one session between a moderator (usability specialist) and a test participant. The moderator asks questions of the participants and gives them a set of tasks to perform on the product in question. The test participant is likely to be thinking aloud as she performs the various tasks. The moderator notes the participant's behavior and responses to questions. Lab tests are best used in formative studies where the goal is to make iterative design improvements. The most important metrics to collect are about issues, including issue frequency, type, and severity. Also, collecting performance data such as task success, errors, and efficiency may also be helpful.

Self-reported metrics can also be collected by having participants answer questions regarding each task or at the conclusion of the study. However, we recommend that you approach performance data and self-reported data very carefully because it's easy to overgeneralize the results to a larger population without an adequate sample size. In fact, we typically only report the frequency of successful tasks or errors. We hesitate even to state the data as a percentage for fear that someone (who is less familiar with usability data or methods) will overgeneralize the data.

Usability tests are not always run with a small number of participants. In some situations you might want to spend some extra time and money by running a larger group of participants (perhaps 10–50 users). The main advantage of running a test with more participants is that as your sample size increases, so does your confidence in the data. This will afford you the ability to collect a wider range of data. In fact, all performance, self-reported, and physiological metrics are fair game. But there are a few metrics that you should be cautious about. For example, inferring website traffic patterns from usability-lab data is probably not very reliable, nor is looking at how subtle design changes might impact the user experience. In these cases, it is better to test with hundreds or even thousands of participants in an online study.

Online studies involve testing with many participants at the same time. They are an excellent way to collect a lot of data in a relatively short amount of time. Online studies are usually set up similarly to a lab test in that there are some background questions, tasks, and follow-up questions. Participants go through a well-defined script of questions and tasks, and all their data are collected automatically using the particular tool. You can collect a wide range of data, including

many performance metrics and self-reported metrics. It may be difficult to collect issues-based data because you're not directly observing participants. But the performance and self-reported data can point to issues, and verbatim comments can help infer their causes.

Online surveys are also appropriate for capturing data on more subtle designs. For example, you can gauge emotional responses to specific visual design elements. The main drawback of online usability studies is that the data you get from each participant is less rich than what you can get in a lab, but that may be offset by the larger number of participants. Another limitation is that online studies only work well with websites or software. Products such as consumer electronics can't easily be tested using an online study.

Focus groups are a great way to get at people's perceptions and attitudes about any particular product or concept. In most cases, there is no direct interaction with the product. All the data from a focus group are in the form of self-reported metrics. Some usability specialists prefer to administer short questionnaires before the focus group begins or at its conclusion. Some of the more useful questions revolve around the likelihood to use the new functionality or to recommend the product to friends. Typical focus groups include about eight to ten participants. We recommend that you conduct at least three groups whenever possible. Data from only one focus group of eight or ten participants may not be very reliable, partly because of the possibility of only one particularly vocal participant swaying the group's opinions.

FOCUS GROUPS VERSUS USABILITY TESTS

When some people first hear about usability testing, they believe it is the same as a focus group. But in our experience, the similarity between the two methods begins and ends with the fact that they both involve representative participants. In a focus group, the participants commonly watch someone demonstrate or describe a potential product, and then react to it. In a usability test, the participants actually try to use some version of the product themselves. We've seen many cases where a prototype got rave reviews from focus groups and then failed miserably in a usability test.

3.4.3 Participants

The participants in any usability study have a major impact on its findings. It's critical that you carefully plan how to include the most representative participants as possible in your study. The steps you will go through in recruiting participants are essentially the same whether you're collecting metrics or not.

The first step is to identify the recruiting criteria that will be used to determine whether a specific person is eligible to participate in the study. Criteria should be as specific as possible to reduce the possibility of recruiting someone who does not fit the profile(s). We often recruit participants based on many characteristics, including

their experience with the web, years away from retirement, or experience with various financial transactions. As part of identifying the criteria, you may segment participant types. For example, you may recruit a certain number of new participants as well as ones who have experience with the existing product.

After deciding on the types of participants you want, you need to figure out how many you need. As you saw in section 2.1.2, the number of participants needed for a usability test is one of the most hotly debated issues in the field. Many factors enter into the decision, including the diversity of the user population, the complexity of the product, and the specific goals of the study. But as a general rule of thumb, testing with about six to eight participants for each iteration in a formative study works well. The most significant usability findings will be observed with the first six or so participants. If there are distinct groups of users, it's helpful to have at least four from each group.

For summative usability studies, we recommend having data from 50 to 100 representative users. If you're in a crunch, you can go as low as 20 participants, but the variance in the data will be quite high, making it difficult to generalize the findings to a broader population. In the case of studies where you are testing the impact of potentially subtle design changes, having at least 100 participants is advisable.

After determining the sample size, you will need to plan the recruiting strategy. This is essentially how you are actually going to get people to participate in the study. You might generate a list of possible participants from customer data and then write a screener that a recruiter uses when contacting potential participants. You might send out requests to participate via e-mail distribution lists. You can screen or segment participants through a series of background questions. Or you might decide to use a third party to handle all of the recruiting. Some of these companies have quite extensive user panels to draw on. Other options exist, such as posting an announcement on the web or e-mailing a specific group of potential participants. Different strategies work for different organizations.

3.4.4 Data Collection

It's important to think about how the data are going to be collected. You should plan out well in advance how you are going to capture all the data that you need for your study. The decisions you make may have a significant impact on how much work you have to do further down the road when you begin analysis.

In the case of a lab test with a fairly small number of participants, Excel probably works as well as anything for collecting data. Make sure you have a template in place for quickly capturing the data during the test. Ideally, this is not done by the moderator but by a note taker or someone behind the scenes who can quickly and easily enter the data. We recommend that data be entered in numeric format as much as possible. For example, if you are coding task success, it is best to code it as a ''1'' (success) and ''0'' (failure). Data entered in a text format will eventually have to be converted, with the exception of verbatim comments.

The most important thing when capturing data is for everyone on the usability team to know the coding scheme extremely well. If anyone starts flipping scales (confusing the high and low values) or does not understand what to enter for certain variables, you will have to either recode or throw the data out. We strongly recommend that you offer training to others who will be helping you collect data. Just think of it as cheap insurance to make sure you end up with clean data.

For studies involving larger numbers of participants, consider using a data-capture tool. If you are running an online study, data are typically collected automatically. You should also have the option of downloading the raw data into Excel.

3.4.5 Data Cleanup

Data rarely come out in a format that is instantly ready to analyze. Some sort of cleanup is usually needed to get your data in a format that allows for quick and easy analysis. Data cleanup might include the following:

Filtering data. You should check for extreme values in the data set. The most likely culprit will be task completion times (in the case of online studies). Some participants may have gone out to lunch in the middle of the study, and their task times will be unusually large. Also, some participants may have taken an impossibly short amount of time to complete the task. This is likely an indicator that they were not truly engaged in the study. Some general rules for how to filter time data are included in section 4.2. You should also consider filtering out data for participants who do not reflect your target audience or where outside factors impacted the results. We've had more than a few usability testing sessions interrupted by a fire drill!

Creating new variables. Building on the raw data set is very useful. For example, you might want to create a top-2-box variable for self-reported rating scales by counting the number of participants who gave one of the two highest ratings. Perhaps you want to aggregate all the success data into one overall success average representing all tasks. Or you might want to combine several metrics using a *z*-score transformation (described in section 8.1.3) to create an overall usability score.

Verifying responses. In some situations, particularly for online studies, participant responses may need to be verified. For example, if you notice that a large percentage of participants are all giving the same wrong answer, this should be investigated.

Checking consistency. It is important to make sure that data are captured properly. A consistency check might include comparing task completion times and successes to self-reported metrics. If many participants completed a task in a relatively short period of time and were successful but gave the task a very low rating, there may be a problem with either how the data were captured or participants confusing the scales of the question. This is quite common with scales involving self-reported ease of use.

Transferring data. It's common to capture and clean up the data using Excel, then use another program such as SPSS to run some statistics (although all the basic statistics can be done with Excel), and then move back to Excel to create the charts and graphs.

Data cleanup can take anywhere from one hour to a couple of weeks. For pretty simple usability studies, with just a couple of metrics, cleanup should be very quick. Obviously, the more metrics you are dealing with, the more time it will take. Also, online studies can take longer because more checks are being done. You want to make sure that the technology is correctly coding all the data.

3.5 SUMMARY

Running a usability study including metrics requires some planning. The following are some of the key points to remember.

1. The first decision you must make is whether you are going to take a formative or summative approach. A formative approach involves collecting data to help improve the design before it is launched or released. It is most appropriate when you have an opportunity to positively impact the design of the product. A summative approach is taken when you want to measure the extent to which certain target goals were achieved. Summative testing is also sometimes used in competitive usability studies.

2. When deciding on the most appropriate metrics, two main aspects of the user experience to consider are performance and satisfaction. Performance metrics characterize what the user *does* and include measures such as task success, task time, and the amount of effort required to achieve a desired outcome. Satisfaction metrics relate to what users *think* or *feel* about their experience.

3. Budgets and timelines need to be planned out well in advance when running any usability studies involving metrics. If you are running a formative study with a relatively small number of participants, collecting metrics should have little, if any, impact on the overall timeline or budget. Otherwise, special attention must be paid to estimating and communicating costs and time for larger-scale studies.

4. Three general types of evaluation methods are used in collecting usability data. Lab tests with small numbers of participants are best in formative testing. These studies typically focus on issues-based metrics. Lab tests with large numbers of participants (more than a dozen) are best to capture a combination of qualitative and quantitative data. These studies usually measure different aspects of performance such as success, completion time, and errors. Online studies with very large numbers of participants (more than one hundred) are best to examine subtle design changes and preferences.

5. Clearly identify the criteria for recruiting participants for the usability test, making sure they are truly representative of the target users. In a formative usability study, testing with about six to eight participants for each iteration is usually enough. If there are distinct groups of users, it's helpful to have at least four from each group. For summative usability studies, we recommend collecting data from 50 to 100 representative users, if possible.
6. Plan well in advance how you are going to capture all the data you need for your study. Make sure you have a template in place for quickly capturing the data during the test and that everyone who will be assisting with the data collection is familiar with any coding conventions. Consider using data-logging tools for larger-scale studies.
7. Data cleanup involves manipulating the data in a way to make them usable and reliable. For example, filtering data means removing extreme values or removing records that are problematic. Consistency checks and verifying responses are important steps in making sure that participants' intentions map to their responses.