CHAPTER 3
Designing the Study

This chapter walks you through the steps in designing an online usability study. We cover all the major sections of a study in the order a participant would likely encounter them: introductory content, screening and starter questions, tasks, follow-up questions, and wrap-up. The end of the chapter includes a discussion of some special topics, including progress indicators and speed traps.

With online testing it’s especially important to put thought into the study design since you can’t change it on the fly. When creating questions and tasks, it’s important to anticipate how you may want to structure the study results and tie them back to your business goals. The last thing you’ll want is to have clients or business partners asking for data that you didn’t collect.

We also want to emphasize that all studies and user groups are different. Just as a seasoned designer never gets a design right the first time, don’t expect that you’ll get the study design right the first time either, so be prepared to pilot test (pretest) the study, which is discussed more extensively in Chapter 4.

3.1 INTRODUCING THE STUDY

When creating an online usability study, you should consider how to set expectations for the participants so that they are not likely to leave the study early—often referred to as bailing or “drop-off.” Especially these days when there are tons of spam, it’s hard for people to know when to trust something that asks for their time and clicks.

To set these expectations, you should start with one or two concise introductory pages that address the

- Purpose of the study
- Sponsor/contact information
- Time estimate
- Incentive
- Technical requirements
- Legal information/consent
- Instructions

© 2010 Elsevier Inc.
Ideally, you want to limit the text as much as possible and format it to enhance scanning. Using bullet points instead of dense paragraphs is a good way to do this.

### 3.1.1 Purpose, sponsor information, motivation, and incentive

First, you should briefly describe the purpose of the study, without giving away any information that may bias the behavior of the participants. For example, if the focus of the study is on the effectiveness of ads, you may want to be more general up front if you suspect that they may look at the ads unnaturally throughout the test. Tell them something less specific, such as you are studying how users interact with different types of Web pages. However, with any research it’s ethically important to disclose the purpose of the study at the end of the study if you were being more general toward the beginning.

Try to use this introduction to engage the participants’ interest so that they are not tempted to fly through the study incentive hungry and with little care. Emphasize that the purpose of the study is to gain their valuable feedback so that the product can be better for them. Make them feel they have a stake in the quality of the product. When people feel as though they’re not just another number, they will take it seriously and the incentive will become secondary.

Just as it’s important to make people feel engaged enough to participate in the study, it’s equally important to gain their trust. Establishing trust will help get participants to take part in the study and also produce more reliable data. In the introduction, you should establish the legitimacy of the study so that participants aren’t concerned that the study is a façade for a virus that will eat away at their computer or that you’re secretly trying to sell something or raise money (often referred to as “sugging” and “frugging,” respectively). One way to establish trust is to make it clear who’s sponsoring the study and who to contact with questions. If you’re conducting a competitive study and don’t want to reveal the sponsoring organization’s name, still give contact information for a third party, including a name, phone number, and email address.

If you’re using an existing customer base or user panel or have conducted surveys with the same population in the past, it’s not as important to
emphasize the sponsorship. However, it’s always a good idea to provide contact information, even if the participant is just having technical trouble or needs some assistance. It all goes back to establishing trust with the participant.

As with establishing trust, it’s important to elicit intrinsic motivation at the outset of the study, as most drop-offs occur during the first half of a survey or study (Brazil et al., 2008). You may want to appeal to their sense of belonging in a group of special users, who can make a difference for others, and in a way that will also benefit them. Make them feel as though they are contributing to their fellow customers, co-workers, their profession, or society in general.

As discussed in Chapter 2.8, you may want to provide an incentive to participants. If so, make sure to emphasize it in the introduction (without sounding gimmicky and cliché) and again at the end of the study, including how and when participants will receive it.

### 3.1.2 Time estimate

Providing the participants with a time estimate is especially important in setting expectations. We recommend that you express the time estimate as a range, such as “5–10” minutes or “15–20 minutes” or you can use descriptors such as “about” or “approximately” to describe the time required to complete the study, such as “this study will take about 10 minutes to complete.” It’s critical to be realistic in the estimate and, if possible, to base it on pilot testing. If you underestimate, the drop-off rate will be much higher, as participants may realize that it’s starting to take longer than they wanted to spend. If you overestimate, the number of participants willing to take the study may be lower.

Several factors influence how long people are likely to spend with an online usability study, but they generally boil down to one overriding factor: motivation. The more motivated the participants are, the more time they’ll devote. We’ve conducted online usability studies that participants completed in only 5–10 minutes and others that took close to an hour. Of course, the other general rule is the shorter the study, the more participants you’re likely to get. Obviously, getting someone to devote 5–10 minutes doesn’t take much motivation and, from what we’ve seen, may increase the participation rate significantly if you’re able to keep it that short and sweet. But getting someone to devote an hour does take some serious motivation.

An interesting trend we’ve seen is that there seems to be a threshold in many peoples’ minds around 15–20 minutes. If a study is expected to take less than 15 minutes, many people seem more willing to just go ahead and do it. However, if a study is expected to take more than 20 minutes, many people seem to view this as a more serious time commitment.
Beyond the Usability Lab

52

Similarly, Hogg and Miller (2003) found that the drop-off rate for a survey increased significantly when a survey took more than 15 minutes to complete. What’s more, the participants who dropped off from the longer surveys were shown to have less intrinsic motivation (inherent interest) in the concept of the survey. Although there may be something lost in translation from an online survey to an online usability study, it is not a stretch to assume that the same thing may apply here; not only will a longer study produce more drop-off, it may lead to a type of self-selection bias, where those who are left completing the survey may be more interested or knowledgeable about the concept being tested.

Figure 3.1 shows a simple example of displaying the purpose and time estimate from an online usability study of the National Cancer Institute Web site.

Because time on task is collected as part of this study, note that the participants are asked to plan for participating when they’ll be least interrupted.

A STUDY WITHIN A STUDY

In one online usability study we conducted, it ran for 2 days and an email message was sent to all the potential participants each morning. On the first day, our email message and the introductory page of the study gave an estimate of about 20 minutes to complete the study, which was our best estimate at the time. After we looked at data from the first day, we saw that the average time was actually between 10 and 15 minutes. As a result, we revised the time estimate to 10–15 minutes for the second day. The participation rate improved significantly.

Similarly, Hogg and Miller (2003) found that the drop-off rate for a survey increased significantly when a survey took more than 15 minutes to complete. What’s more, the participants who dropped off from the longer surveys were shown to have less intrinsic motivation (inherent interest) in the concept of the survey. Although there may be something lost in translation from an online survey to an online usability study, it is not a stretch to assume that the same thing may apply here; not only will a longer study produce more drop-off, it may lead to a type of self-selection bias, where those who are left completing the survey may be more interested or knowledgeable about the concept being tested.

Figure 3.1 shows a simple example of displaying the purpose and time estimate from an online usability study of the National Cancer Institute Web site.

Because time on task is collected as part of this study, note that the participants are asked to plan for participating when they’ll be least interrupted.

3.1.3 Technical requirements

Ideally you’d want the study to work with any operating system, Web browser, and screen resolution, but sometimes that may not be the case. Regardless, you should still let participants know up front what the technical requirements are, as in the examples shown in Figures 3.2 and 3.3. If you’re using a commercial service to administer the test, this information may be built into their tool or instructions already.

Most commercial tools offer the option to ask the participant to download a browser companion...
or scripting component in order to track URLs, clickstreams, task times, etc. If you have any control over the instructional text, again be sure to explain in plain terms what the user will be downloading and whether it’s of any risk to their computer or to the protection of their personal information.

### 3.1.4 Legal information and consent

Depending on the subject matter of the study, you may need to consider creating a legal disclaimer or nondisclosure clause. For example, you may be showing an external audience some proprietary information or new designs that you want kept confidential. In some cases you may require the participants to

![Figure 3.2](image1.png)

*FIGURE 3.2* Example showing technical requirement to disable pop-up blocker.

![Figure 3.3](image2.png)

*FIGURE 3.3* The Keynote Connector download window. Copyright © 2009 Keynote Systems, Inc. All rights reserved.

If using a commercial service, check with your vendor to see what participants will have to download to use it. Some tools are built to uninstall any downloaded components upon completion of the study.
acknowledge that they will not share the information they see with anyone else. This is especially important because anyone can easily take a picture or screen capture of something they’re looking at and email it out to others.

Some type of informed consent agreement is quite common in usability studies and is the ethical responsibility of the researcher to administer. It’s more often used for face-to-face usability testing so that you get the participant’s permission to record video of them and to explain their rights as a participant in the study. In any user research, though, you want to assure them that their data are strictly confidential and that their personal information will not be used. If you are having them access their own accounts or tracking them while they make a purchase on an ecommerce site, the consent form helps assure participants that you will not be capturing any of their own data or, if you are, that data will be kept strictly confidential. This can be tricky so be careful how you are capturing and using data. Figure 3.4 is one example of how consent may be presented to the user.

Before you throw your hands up and say that you don’t want to be entangled in consultations with lawyers, please note that we’re erring on the conservative side to raise awareness of potential risks. If you’re doing a simple study that doesn’t involve any personal information or reveal any company secrets, then legal consultation may be unnecessary. Assess your personal and institutional risk, consult with any corporate policies that apply, and decide if this is important enough to pursue. We do, however, think it’s always your responsibility to inform participants that their data will be treated confidentially, regardless of whether the text is vetted formally through a lawyer.

**TIP**

Some services may have already covered things such as a nondisclosure agreement and treatment of data when panelists registered to participate in studies through them. Check with your vendor, if you’re using one.
3.1.5 Instructions

Instructions are an important part of beginning any study. They need to be clear and concise. Some of the points you may need to cover include the following.

■ An assurance that the study is an evaluation of the Web site, not of the user, and that they just need to put forth their best effort.
■ If you’re giving participants tasks to do, ask them to try to find the answer regardless of whether they already know it.
■ Ask them to respond naturally, as if they were using the Web site on their own.
■ Special instructions on how to advance through and answer the tasks, access help/instructions, pause, or quit the study at any time.

If you find that you’re covering a great deal of information from the preceding sections in one screen or window, consider presenting instructions on a subsequent screen so that users won’t be overwhelmed. Figure 3.5 provides one example of how instructions might be presented to the participant.

We know that users can’t always rely on their short-term memory nor do they find it fruitful to try to learn instructions. If the instructions page feels fairly lengthy, consider adding a way for participants to access the page again at any time during the study. Some commercial services allow you to do this.

![An instructions screen customized using UserZoom.](image)
3.2 SCREENING QUESTIONS

In any kind of online usability study, you may include some questions at the beginning to automatically screen out participants who don’t meet basic criteria for the study. The idea is that if someone doesn’t answer the screening questions in a desired way, they are not asked to participate in the study. It’s simply a way to ensure that you are only reaching the intended audience.

3.2.1 Types of screening questions

The types of screening questions to use will differ depending on the Web site you’re testing. In general, however, consider whether to screen out at least two groups of people:

- People who clearly do not make up the target user groups. For example, if you’re testing a teenage girl’s fashion Web site, it’s pretty safe to say that men over 50 are not a viable user group. Similarly, if you’re testing a Web site that allows people to rent jet planes for everyday personal use, you will probably only be targeting participants at a very high income level or net worth.

- People who may bias the results of the study. The most common example of this is to screen out participants who have higher technical expertise on average, such as developers or Web designers. Unless the Web site is targeted to that user group, these types of professionals are more likely to think like the designers did and “get” what average users may not have.

When constructing screening questions, the first thing you want to do is identify all of your criteria for participating in the study. Then look at the criteria and do a sensitivity check. Are you screening by sensitive information, such as race, nationality, age, gender, or income? If so, make sure you have a valid reason to do so (e.g., high net worth individuals for the jet-renting example). Be aware that asking these kinds of questions may make participants suspicious and therefore unlikely to participate.

Note that you should also try to make a distinction between what criteria should screen out participants versus sort them. If you’re unsure or skeptical about whether to exclude certain groups of people, sorting them will allow you to still get representation from them without biasing targeted user group data, as they’re being categorized separately. The decision of whether to screen out or sort also has implications for the automated screener. For example, you may want about half males and half females. If you have a quota, such as a 500-person cap for each gender, the screener tool should include functionality to keep track of how many males and how many females participate and to screen out one gender once the quota is filled. Many vendor tools allow you to do this.

One common example of a screening question might be to see how familiar users are with a particular topic or Web site. A very simple example is illustrated in Figure 3.6 using the TerraPass Web site example. Participants are simply asked how familiar they are with carbon offsets. If the designers were targeting groups of people who are newer to the
concept and may come to the site to learn more, they might want to screen out participants who are already very familiar with the concept.

We make an intentional distinction between screening questions and starter questions (Section 3.3), as the purpose of each is different. A screening question is meant to include, exclude, or sort a participant based on their answers to a question, whereas a starter question is meant to provide background information about participants that will be helpful in slicing, dicing, and informing the data analysis.

### 3.2.2 Misrepresentation checks

There may be some people trying to participate in studies just for an incentive. Your responsibility to mitigate this problem therefore starts by making sure that people are who they say they are in their screening questions, which we refer to here as “misrepresentation checks.” The exact checks will depend on the subject matter you’re dealing with. The goal is to devise sets of questions where certain combinations of answers are either impossible or statistically very unlikely. An example of this comes from Burke R&D’s Quality Panel Research (2008). They had participants from various user panels participate in a survey in order to measure the incidence of misrepresentation across them. Panelists were asked at the beginning of a survey whether they were left-handed, right-handed, or ambidextrous (“uses both hands equally”). See Figure 3.7. Whereas the percentage of the
population who are ambidextrous is only about 1%, respondents in five different panels claimed that they were ambidextrous 5 or 6% of the time. If a participant answered “ambidextrous,” it is likely that they were just trying to answer anything to qualify for the survey.

Asking handedness doesn’t quite translate to a related screening question for an online user experience study, but perhaps you can devise a set of questions in the context of the Web site domain that can achieve the same goal. The example in which the Burke Institute asked respondents whether they had heard of a nonexistent brand is one type of question you may ask. In our example study of pet Web sites (PETCO and PetSmart), one might ask participants which brands of pet food they buy and then add some brands in the list that don’t exist. If someone claims that they have pets and buy pet food and then choose made-up brand names, they may be misrepresenting themselves.

3.2.3 Exit strategy
It’s important to be graceful when screening out a participant. Just because participants are dealing with a computer, it doesn’t give you license to give people a “so long, loser!” message if they don’t meet screening criteria. This is placing the blame on the user and potentially damaging your reputation with them in the future. Therefore it’s important to inform the users gently when they do not meet screening criteria.

It is also important to try not to reveal why the user was screened out of the study. This is another measure against misrepresentation; if a user has just checked off that he is a male and is immediately terminated thereafter, he may be tempted to go back and check off “female” to become eligible to participate and get the incentive. For this reason, it’s helpful to group screening questions together. This prevents users from immediately understanding what screened them out.

An example of a generic but gentle exit message might be, “We’re sorry, but we have filled our quota for people with your background. Thank you for your interest, and we appreciate your time.”

3.3 STARTER QUESTIONS
Along with screening questions, you may want to collect some data about what type of experience or expectation users have with the product or how much technical experience they have. This can help provide insight when looking at performance data. You might see that less experienced participants perform differently from the experts. Also, you may want to use how they answer the questions to customize the study. For example, you may want to use branching logic and choose to give only certain tasks to specific participants based on their interest or experience.

Although we want to focus this chapter more on task construction than question construction, we didn’t want to leave you without at least a few high-level tips to keep in mind when writing screening and starter questions.
Closed questions (providing participants with answer choices) are easier than open-ended questions to categorize and analyze. The only time you may want to use open-ended questions (e.g., free-form text boxes) is when you think there will be answers you didn’t anticipate, you feel that providing participants with answers may bias their responses, or you are looking for a general opinion, which shouldn’t be the case for screening questions, but perhaps you may incorporate it as a starter question.

- Avoid “double-barreled” questions that combine two questions in one, as participants may not know which question to answer. For example, “To what extent is this Web site easy to use and important to you?”

- Avoid asking yes/no questions as they can sometimes lead or bias a participant into an extreme statement. For example, asking “Do you use this Web site all the time?” can mean different things to different people and may feel somewhat bullying into that position. A better question would be using a multiple-choice question or rating scale offering different intervals for frequency of use.

- Be specific where possible; participants will interpret “daily,” “weekly,” and so on more easily and consistently than “always,” “sometimes,” and so on.

- Avoid asking hypothetical questions. For example, “Would you use this Web site every day for your job if we added a way to buy office supplies?” Participants are not great at knowing how they’ll behave in a situation and sometimes confuse fantasy with reality when faced with hypothetical questions.

- As always, try to stay neutral. Avoid using language that will lead, bias, or clue the participant in to your feelings or motivations as the researcher. This also means avoiding “presupposition” questions, i.e., assuming something about the participant that may not necessarily be true. For example, asking “What is the most fun part of the Web site?” presupposes that the participant finds any part of the Web site “fun,” which may not be the case.

3.3.1 Product, computer, and Web experience

Perhaps the most important starter questions you’ll want to ask are those involving participants’ usage and experience with the product. These factors are likely to correlate with performance metrics during the study.

The trick with these types of questions is to keep them clear and focused. Don’t assume that the amount of time someone has had a product, been a customer, or has been registered on a Web site equals how long they’ve been using the product or Web site. You can ask those questions, but make sure you ask separate questions about their usage. Also don’t assume that just because someone uses something all the time that they think it’s great. Some people have to use something because it’s their job, they can’t afford competitor products, it’s all they know
how to use, or for some other reason. If you want to know something about usage or experience with a product, ask it directly so that you’re not making inferences!

Figure 3.8 shows some examples of past experience questions. Note that rather than asking in general what people use the SuperWidget for, we ask the last time they used it and what they used it for in that instance. Giving a concrete instance to think about helps ground people so that they’re not taxing their memory or going “checkbox crazy” assuming they’ve used the SuperWidget for everything under the sun at some point.

Commonly, we also ask for participants’ familiarity with the Web site. Note, however, that although familiarity may be a useful measure to provide insight into the analysis, it does not imply frequency of use (again, no inferences!). Figure 3.9 is an example using PetSmart and PETCO in which we asked users how many pets they have.

In this study, this question was used for branching purposes (logic built in that routes respondents to different questions based on their answers). If a participant entered “0,” they went on to start the study. If a participant entered “1” through “4 or more” they were brought to another question, as seen in Figure 3.10.

How computer and Web savvy a participant is can also impact their success in using the product. One thing to note is that Web experience may not be the same thing as Web expertise. Chadwick-Dias, Tedesco, and Tullis (2004) found that Web expertise is not always correlated to how much someone has used the Web, but is strongly correlated with opportunities for learning collaboratively when watching or working with others (family members, co-workers, classmates) on computers. You can ask the standard questions around how long and frequently people use the Web and computers, but be aware that this may not make up the entire picture. You may also want to ask them some more questions about what they do on the Web and computers, how they learned it, and in what locations they most often use it.

### 3.3.2 Expectations

One metric that you may want to use in the study is a measure of how well the product actually meets the participants’ expectations. Sometimes you’ll find that certain tasks are difficult for participants, but if they expected it to be difficult that may not be something to worry as much about as the case where they expected it to be easy.
Albert and Dixon (2003) used a methodology where, before the study or task itself, participants would rate how easy or difficult they expected each of the tasks to be on a rating scale. Then participants were asked after each task to rate how easy or difficult it actually was on the same scale. Users don’t necessarily require prior experience on the Web site to develop an expectation; they use their prior experience with similar types of tasks on the Web or just their assessment of the complexity of the task to help guide their expectation rating. One may naturally expect some tasks to be more difficult than others. For example, it should be easy to get a stock quote on Google, but it would probably be difficult to rebalance your entire portfolio to match a new projected retirement date.

By doing this, Albert and Dixon were able to take participants’ data and identify areas needing great improvement (expected to be easy, but wound up difficult), as well as other areas that posed an opportunity (expected to be difficult, was difficult).

Expectations can also be gauged on a broader level. For example, you could ask users how easy or difficult they expect the site to be in general and measure the discrepancy with post-study questions on the same topic.

### 3.3.3 Reducing bias later in the study

As with the introductory text, it’s important in both screening and starter questions not to “show your cards” or reveal too much information to the participant up front. This is a principle common to any kind of user experience study or market research survey. For example, let’s say that you’re planning to do a competitive study for Alpha Corp comparing three different products—Alpha Corp’s Web site, Beta Corp’s Web site, and Gamma Corp’s Web site. It’s probably not a good idea to ask all of the questions solely about the usage of Alpha Corp’s Web site, even if that’s the scope of the screening criteria, or later analysis. Instead, you may want to ask about the participants’ usage of all three products or Web sites so that they can’t guess who the sponsor is. Brand recognition can have a large impact on people’s reactions to products.

Another example may be that you plan to target a particular part of a Web site. Let’s say that you want to test how people use the Alpha Corp store locator, which is one small piece of the whole Web site, but you don’t want to lead the user to think unnaturally or excessively about the store locator before they begin the test. So you may want to add some “distractor” questions about other parts of the site. Distractor questions are questions that may not necessarily provide value, but make it harder for participants to guess the purpose or focus of the study up front and result in more accurate data for the central task.
3.4 CONSTRUCTING TASKS

Task construction is a vital part of an online study. The automated, self-service nature of the study means that there is no room for misinterpretation. You won’t be there to clarify tasks or interpret user intentions. The tasks need to be able to reliably speak for themselves.

The following sections outline some points to keep in mind when constructing tasks. Note a couple of things to consider when reading them.

- In these sections we assume that you want to provide a set of specific, targeted tasks to get the user to interact with specific areas of the site and record their task success, among other measures. For this reason, much of the discussion doesn’t apply to true intent and open Web studies, which tend to elicit less about empirical behavior and more about the user experience with a task or Web site in general.

- We also assume here that participants are giving some kind of answer to each task, whether it is a specific piece of information or data, a specific page identifier, or their own “self-reported” task success. For this reason, much of this discussion does not apply to studies where you’re using clickstream data as the only task completion metric.

3.4.1 Making the task easy to understand

It’s important to balance the need for specific tasks with the amount of mental effort required for participants. For example, if you’re testing a company directory’s search capability and the task is to “find a man with the last name...
of Flovegneiren whose office is on the third floor of the Park Street building, works in finance, and whose first name starts with a B,” the task begins to sound like a college entrance exam question rather than a task. (We know from experience, as one of our earlier online studies included tasks like this!) The result of this is time wasted interpreting a question, and potentially taxing the participant’s working memory, both of which can take a toll on task completion times, as well as the quality of results.

The end goal of the task should be clear as well. Using our Apollo program example, a task might be to “find out in what time period the Apollo program was started.” Here, the term “time period” is not clear enough. Some users may look for a specific year, whereas others may look for a decade or a presidential administration. A better task would be specific and pointed—perhaps to find the year of the first manned Apollo spacecraft launch.
Don’t mistake the term “easy to understand” for “easy to guess.” The task should not be so easy that a participant is likely to know it offhand without looking. If the task is one of common knowledge, for example, “Who won the 2008 U.S. presidential election,” then users will likely take shortcuts and pick the answer they know rather than find it on the site. In cases like this we recommend using some proxy question that still evaluates the same area of the design. For example, in the pets study, one of the tasks was to find the weight of a bag of a particular brand of ferret food that starts with a “Z.” This is likely not to be a common task; a small percentage of users would be looking to buy ferret food, let alone finding out the weight of the bags available. However, if we were to ask participants questions about a common cat or dog food brand, it’s very possible that users would already know some of these things. So here the ferret food example is just being used as a proxy task to have users navigate to a product page.

3.4.2 Writing tasks with task completion rates in mind

Assuming that task completion is one of the primary metrics for the study, you’ll need to know whether participants accomplished the task successfully. This is unlike a typical moderated study, where you may note someone’s degree of success. With online usability studies, there generally isn’t a concept of “partial” success. It’s usually a binary situation: the task was successful or unsuccessful for the participant. There may be different types of task success or failure, but there’s usually not an in-between category (i.e., there is no easy way to score something as “successful with minor problems” or “75% successful,” as with moderated usability testing where you’re watching them in person and able to make a subjective judgment). Depending on how you implement the study, this has implications for how you formulate and word the tasks.

Let’s use an example where you’re testing navigation on a bookseller Web site. If you want a task to be “Find the book by John Smith on bicycles,” you need to be able to confirm that the participant actually found the answer. Let’s assume that your goal is focused on users’ experience navigating to a book on the site when having just an author’s name (John Smith) and topic (bicycles). There are a few ways to approach this problem.

Option 1: The easiest way is to reword the task so that there can be a discrete answer. For example, “What is the ISBN number of the book by John Smith on bicycles?” That way, the user can either input the number or select it from a drop-down list of choices. In this case, finding the ISBN number in particular isn’t that important to you; it’s just the means by which participants can indicate their answer discretely. This option is very direct and our most commonly used method, but is sometimes not easy to construct when dealing with changing content on a live Web site (e.g., such as stock quotes) or a wizard-like task where there is no clear “answer” to find (e.g., opening an account).
Option 2: If you’re using a prototype of the Web site, you may have the ability to tag pages. For example, you could put a code or a letter of the alphabet on a variety of pages throughout the site and then have the user input the code of the page with the answer (see Figure 3.11). This option requires a little more work on the prototyping side, but frees you up to be more natural with the task wording, e.g., “Find the page that shows…” or “Find the photo that shows…” rather than asking for a discrete piece of information.

Similarly, rather than tagging pages of the Web site, you can create a “dummy” page that users may get when clicking on the appropriate link in a prototype (Figure 3.12).

These techniques can work well, but be careful; you’ll want to include a code on almost every page or at least every plausible page for a task so that participants don’t make educated guesses based on which pages have codes and which don’t. The codes should also be random, i.e., if you’re using the alphabet, the first task should not correspond to an answer page that is labeled “A,” the second task answer “B,” and so on.

Option 3: Similar to Option 2, if collecting clickstream data only, you may be able to word the task instructing the user to find a particular page as opposed to finding a specific piece of information. For example, instead of saying “Find out how much it costs for ground (regular) shipping of a gift card” and collecting the price as the answer, you might ask, “Find the page that tells you how much it costs for shipping gift cards.” In the latter case, you’re just using clickstream data to determine success rather than an answer that the participant gives or chooses.
FIGURE 3.12
Two different ways of approaching “dummy” pages. (Top) The page clicked to contains only the code. (Bottom) The link clicked brings up a pop-up with description and code.
Option 4: Rather than relying on clickstream data as in Option 3, another option is to have participants enter the URL or page title itself as the answer to a task. But make sure before you pursue this option that each URL or page title is a unique identifier. Some Web sites that use frames or make use of Rich Internet Application platforms such as AJAX or DHTML tend to yield a static URL that doesn’t change when users click to new pages or components.

“FIND AN ANSWER” OR “FIND A PAGE”?

Whether a task directs people to find a discrete answer or a particular page depends on what you’re looking to discover and the nature of the design. The “find a page” method is often used to test the navigational path in getting to a page, usually when the page itself isn’t that complex or important to test. The “find an answer” method is typically geared toward testing more interactive pages, where simply finding the page wouldn’t capture the entire experience. Here, participants need to go an extra step further to interact with the information on the page in order to answer the task.

3.4.3 Anticipating various paths to an answer

An important consideration with online usability studies is being aware of the various paths a participant can take to an end point. In our bookseller example, let’s say that your goal is to compare two designs of the primary navigation bar to one another. In this case, you might want to force the participant to use the site navigation rather than a search box. This is simply a way to minimize “noise” in data so that everyone is approaching the tasks the same way and you’re comparing “apples to apples” when looking across the data comparing the two designs. There are a few options for how to do this.

Option 1: If you’re using a prototype, you can take away the search box altogether for both designs. This was what we did with our photo clip-art study in which we compared different variations of an expand/collapse menu structure. Some comments showed that users were unsatisfied with the lack of a search field, which most likely led them to rate the experience lower (with self-reported subjective ratings), but this sentiment was rated equally low across both designs. This is okay if your only goal is to compare both designs’ navigation bars, as we did. In other words, only differences in ratings across designs matter rather than absolute ratings of any one design. As a result, if ratings for one design are significantly higher than the other, you can attribute that difference solely to the difference in the design of the navigation. If some used the search box (and especially more so on one design than the other, for whatever reason), it could have potentially muddied the results of the comparison.
Option 2: Disable the search box and instruct users at the beginning of the study not to try and use it. However, if participants don’t read the instructions (we know this is a likely scenario!) they may become frustrated that it’s not working. A similar technique is to provide a search input field, but always display a placeholder page or pop-up explaining that the search is not working in this prototype. Some commercial services allow the ability to show a message to a participant if he or she has done something specific, such as clicking the “Go” button next to the search box (see Figure 3.13). This can be a good option if you don’t have a prototype and are using a live Web site.

Option 3: Another option when testing a prototype is to set it up so that the user will arrive at different answers or unique pages depending on whether he or she searches or navigates. For the bookstore example, you can offer one ISBN number for participants whose path is through the search field and another ISBN number for participants whose path is through the site’s navigation. Alternatively, if you’re labeling pages with codes, you may label those two end pages with different codes. Either way, you actually may be able to determine during the analysis phase what percentage of participants used the search versus the navigation while attempting the task based on their answers (and then do your comparisons within the navigation-user groups). Of course, keep in mind that this may involve some fancy prototyping footwork to accomplish.

Option 4: You can word the tasks so that the participants would not be likely to use the search box for every task. Using our pets study example, one task was “You remember the ferret food you like to buy starts with a ‘Z’.”
Find out the weights of the two bags of this food available for purchase."
If the task were to specify that the food is called “Zupreem,” it is a likely item for users to search because it’s so unique. Creating the “starts with a Z” scenario may not be the most realistic or representative scenario for this user group, but it’s not a stretch either. The point we want to emphasize is that it serves as a proxy task to get users to navigate to a specific area without using the search. The experience with the actual task is not your goal in this case, but rather the navigational journey of getting to a product page.

One thing to note here is that you want to make sure there aren’t other answers that satisfy the same task wording. We were comfortable using “ferret food that starts with a ‘Z’” because we knew there weren’t other products on the site that satisfied that criterion. Be sure to do your homework first! And if you miss something, don’t worry—you should catch any odd result patterns in the pilot testing!

A small interjection here: we’re not implying that if you’re testing a Web site with a search engine you shouldn’t let participants use it. In the example of the book-seller, we were assuming a scenario in which you’re comparing two navigational designs. But let’s say that instead you were just assessing the general usability of one design and know that the most common scenarios are users coming to the site with a particular title and/or author in mind rather than a vague sense of what they’re looking for. In that case, you want participants to get to the answers fast, no matter how they do it. You’ll get more realistic and useful information by letting users get to the answer any way they choose.

3.4.4 Multiple-choice answers
As mentioned previously, you might be able to use clickstream data to determine whether the participant was successful in the task. In most cases, however, it’s easier to ask participants to find a page or a specific answer and enter it once they’ve found it. This can be a code, an actual piece of data, or something such as a page title or URL. Whenever possible, you should provide multiple-choice answers (via a drop-down list or radio buttons). Why? Well, a free-form text input box might pose unique analysis challenges. Consider the following examples.

- Different interpretations of answers: If you’re asking a dollar amount, one participant may enter “$11” whereas others may enter “Eleven,” “11 dollars,” or “11.00.”
- Misspelling: An author’s first name might be “Alison” but some people will probably spell it “Allison” (especially if they are unable to view the site simultaneously while they’re answering the question).
- Mistyping: Many people are not trained in typing and will look at the keyboard while typing rather than at the word itself. Others may type so fast that they switch letters around. You may wind up with common typing mistakes like “hte Doors.”
These problems are more common than you may think! It could be an analysis nightmare because there is no automatic way to score task success if you have a spreadsheet of data in all different, unanticipated formats. Therefore, you might have to look tediously through many entries to manually code a task as successful or not. Drop-down lists or radio buttons, however, have predetermined values that are returned with the same spelling and format for each participant, which simplifies the analysis greatly.

Ideally, participants would see the Web site, task, and a way to answer the task simultaneously, which is in fact how most of the vendor tools are built. But a few tools have constraints on their setup so that participants are asked to interact with the Web site first and then to answer the task on a subsequent page. Therefore, predefined answer choices make it much easier for participants because they are using recognition of the answers instead of having to recall them.

When using multiple-choice answers, you should be mindful of the number and types of choices you include in the list. Let’s say you’re trying to decide between using 5 or 10 choices. Obviously, offering 10 choices rather than 5 will decrease the chance of someone choosing the right answer randomly. It also decreases the likelihood that someone might use deduction to figure out the answer. But it depends on what the choices are. Consider the example shown in Figure 3.14 from the Apollo Program usability study.

In this example, the choices each contain two astronaut names. It might take a second for the participant to read through the choices. If there were 10 choices like this, it would take more mental effort on the part of the participant to review the choices and discern the differences. But let’s hypothetically say that the task was to pick one astronaut from a list. Then, having only 4 or 5 to choose from may be too easy to guess. In this case, you may want to increase the number of choices.

This example also points out some desirable features of the “distractor” options, that is, incorrect answers. In this case, the answer is Cernan and Schmitt, the Apollo 17 astronauts who walked on the moon. But all of the distractors are also the names of astronauts who walked on the moon, so they all have a “ring of authenticity” to them. Including options such as “Laurel and Hardy” might have been fun, but very easy for participants to eliminate.
Not only should the answers sound legitimate, but they should actually be plausible. For example, the participant’s task is to find the page of an ecommerce site where she can update her billing address—using the example of page letters, let’s say this is page H. You may be tempted to throw in random letters as the other choices in the answer dropdown—P, J, and R. But what if the billing address page looks exactly like the shipping address page, which is page Q? People may get to page Q, decide that it’s the answer, and find that it’s not in the list, so keep looking. The results of the study now have a slight bias toward success for that task where people otherwise might have gotten it wrong. Including plausibly wrong answers will give you additional insights when analyzing data, as it may confirm suspicions of possible usability issues you anticipated.

If you’re using a code for an answer such as a page number or letter of the alphabet, be careful of too few choices there as well. If the site or scope of the design is small, the participant could easily identify where all the choices are and decide which one is most correct. Here it’s recommended to use more choices, as shown in Figure 3.15.

3.4.5 Including a “none of the above” option

The example in Figure 3.15 also illustrates why it’s a good idea to include a “None of the Above” option. Including such an option is another way to discourage participants from answering by deduction. For this reason, you may not have to worry about including many answer choices if you include a “None of the Above” option. Take note, however, that some participants may use “None of the Above” as a way to give up on the task. Therefore we also recommend including this option in addition to a “Don’t Know” or “Give Up” option, which is discussed in the next section. Having both will allow you to differentiate between participants who were unable to find an answer and participants who found one that was not listed in the answer choices.
3.4.6 Including a “don’t know” or “give up” option

In most online usability studies we conduct, we include a “Don’t Know” or “Give Up” option. One example is shown in Figure 3.14, where “Give Up” is an option in the drop-down list. Some commercial services also provide ways to give up on a task, such as WebEffective, which offers a “Give Up” option next to it (Figure 3.16). Many of the tools also allow you to change the labels of these buttons if you want to use different verbiage.

There are different schools of thought around a Give Up option. Some may argue that providing this option prompts participants to put in less effort when going through the study. However, not providing the option will force people to guess an answer when they haven’t found one, which could introduce more noise in data and more frustration for the participant.

Our recommendation is to include the Give Up option. First, it will acknowledge to the participants that it’s okay to give up on a task and will encourage them to move on with the study rather than exiting altogether. Second, it will allow you to make a distinction between the percentages of tasks that failed as a result of giving up compared to participants thinking they’re in the right place when they’re not. This could be an important and insightful distinction.

IS “NONE OF THE ABOVE” AN INVITATION FOR MENTAL CHEATERS?

There is discussion in the market research world about whether including “None of the Above” as an answer for a survey question encourages mental cheaters to choose that option. There are arguments and study results on both sides of that discussion. Our answer to that debate is simply to include task times as one of the metrics in the study. That way, you can weed out data from anyone who spends less time answering a task than is possible for that task. Many commercial services also allow you to do this automatically by specifying a time threshold. We talk about this and other data cleanup techniques in Chapter 5.

FIGURE 3.16
Keynote’s WebEffective provides a “Give Up” option for every task. Copyright © 2009 Keynote Systems, Inc. All rights reserved.
The tool may also allow you to branch off where a person chooses the “Give Up” option. For example, if someone clicks “Give Up” you can interject an open-ended question asking why they gave up or perhaps some questions probing more pointed questions about their experience with the task. Be careful not to make the user feel like a failure when doing this. Instead of “Why did you give up?” ask something like, “What prompted you to move on from this task?”

### ARE PARTICIPANTS RELUCTANT TO ADMIT THAT THEY’RE GIVING UP?

One thing we’ve wondered is whether participants shy away from the idea of “giving up” on a task. As much as we remind them that the study is their evaluation of a Web site, some users inevitably feel as though it’s a challenge or test of some sort. Thus, giving up can give the perception of failure to participants. Although we don’t have data on it either way, we encourage you to try other phrases such as “Unable to find” or “Stop looking” and see if and how that impacts your data and the study participants’ perceptions.

#### 3.4.7 Randomizing task order and answer choices

As with traditional usability testing, it’s frequently beneficial to randomize the order of tasks to minimize the learning effects on any one task. The first task that a participant receives in an online usability study almost always takes a performance “hit,” especially in terms of task time. Randomizing the task order spreads this effect across all tasks. The exception is when some tasks are dependent on other tasks to be completed first or have a natural progression. Let’s use the example of buying a book on a bookstore Web site. Assume that one task is to locate a particular book, another task is to add it to the shopping cart, and another task is to check out. Obviously the participant can’t check out without finding and adding a book to the cart first, so there is a natural progression to the order of tasks. Randomizing these tasks would not only be confusing, but it wouldn’t be possible in this situation. Perhaps a less obvious example would be the TerraPass example study. TerraPass is a Web site that allows users to search for information about and purchase carbon offsets. Because the concept of “carbon offsets” is not readily understandable to many participants, it doesn’t make sense to have them jump directly into buying them. A couple of initial tasks would be appropriate, e.g., asking participants to search for information about TerraPass, and what it costs. This would be more representative of an actual flow of information seeking.

---

**TIP**

If you’re testing with an international audience, be aware of the cultural implications of giving up. Whereas Americans generally may not be shy about admitting that they can’t find something, some other cultures may be less prone to do so. As always, do your research first!
3.4.8 Using a subset of tasks

In some cases you might want to give each participant a smaller number of tasks selected from a larger list. This can be useful if you want to cover many aspects of the Web site without making the study too time-consuming for each participant. For example, assume you have eight tasks total but each participant is only given four of them. One participant may get Tasks 2, 5, 6, and 7, whereas another participant may get Tasks 3, 4, 5, and 8. If your tool is smart it will keep track of all the tasks and automatically balance the distribution across all participants. Otherwise, simple random chance will likely give you an acceptable distribution across a relatively large number of participants.

However, be sure that you anticipate getting enough participants performing each task or the results may not have enough statistical power. If each of the participants is only getting half of the tasks, you’ll need twice as many participants as you would if they were getting all of the tasks. Also take into consideration any major segmentation you may want to do on data, for example, looking at the results by novices versus experts, age groups, and so on. Tasks would need to be attempted fairly equally across each noteworthy group of participants.

3.4.9 Self-generated and self-selected tasks

As in a typical moderated usability study, you may want to have the participant self-generate a couple of tasks to perform along with the predefined tasks. This is useful because it allows participants to share with you what they’re looking to do with the Web site and perhaps identify something that you’ve overlooked.

If you use this technique, you should have participants generate their tasks at the beginning of the study before they see the Web site. If you let participants see or explore the site first, they will often click to a specific place first and then decide that the information they’ve found is the task they are interested in. Don’t allow them to put the cart before the horse!

Self-generated tasks are also at the heart of true intent studies. In a true intent study, a real user is going on a live Web site for a real purpose and is intercepted.
Users should be intercepted upon entering the site, upon exiting the site, or both. This essentially results in a contextual, live-site online usability study. Participants can be asked at the beginning what they’re trying to accomplish, and their reactions and self-reported success or failure are asked at the end of the interaction. Clickstream data can help supplement the story of their journey.

One way to capture a self-generated task is to have participants enter a task in a text box and then have the tool store that information and present it as one of the tasks during the study. The process of asking for input and saving it for later tasks and questions is referred to as “piping.”

There is usually no systematic way to score self-generated tasks as successful or not. Even if you were to have users provide answers to their own questions or if you capture clickstream data, most tasks are probably too specific to provide meaningful data across participants. However, you could ask participants an open-ended question afterward about how the task went for them or have them self-report whether they were successful. It may still be a useful piece of information to know what users are wanting or trying to do and whether they were satisfied with their experience. It also may just become a vehicle to get people using and thinking about the site in ways you didn’t anticipate. The end goal is the same—understanding how the site could be made easier and more understandable, as well as just a better user experience all around.

A related technique is to let participants self-select their tasks. This means having them choose the tasks appropriate for them from a list shown at the beginning of the study. This is another take on using subsets of a task list in which the participants choose the subset rather than the system choosing it for them.

If you use this method, be sure to include some guidance rather than just telling users to “choose the task you want to do.” In fact, it’s best to form it as a screening or starting question so that users don’t even know it’s a setup for tasks. Otherwise, you may wind up with some type of bias as a result of what tasks are chosen. Users may choose something that they know how to do already, do most of the time, think would be most easy, or even most challenging.

An example of such a technique might be to have participants choose activities they think are most important or interesting to them or that they do most frequently. If they’ve never been on the site before, perhaps it may be what they’d expect other users to do most often. There are several ways to implement this. Rather than just selecting a task, participants could rate the importance and/or frequency of doing that task and then they’re given the tasks that score in the top-2 box of that scale. Alternatively, you could have them rank order the tasks by some criteria. Figures 3.17 and 3.18 are some examples of this, using a digital video recorder (DVR) as the focus.
Be mindful about how you are using these criteria for participants to self-select tasks. For example, if you are using frequency to determine the tasks, this can bias participants to high task success if they are already very familiar with the areas of the Web site you’re testing. However, there are instances where it makes sense to have only frequent users test the product, particularly when...
it’s a realistic case that the user is very familiar with the functionality or has knowledge of a special domain that’s not easily understood by just any audience. As always this depends on the goals of the study and the nature of the product.

### 3.4.10 Self-reported task completion

Self-generated tasks are usually scored using self-reported task completion. Because the task and/or answer may be open or subjective rather than predefined, you let the participants decide for themselves whether they were successful with each of their tasks.

Figures 3.19 and 3.20 show some methods of collecting self-reported answers.

Self-reported task completion can be used in any study and with any set of tasks, self-generated or not. Aside from self-generated tasks, other cases for using this technique include the following.

- **When answers to questions or tasks may be changing.** This can be the case any time dynamic data are involved. Data may be changing moment to moment (e.g., the price of a share of stock) or more slowly (e.g., the price of a new bestseller). Sometimes you can get around this problem by providing ranges for predefined answers or by allowing a range of acceptable answers for free-form input (which, again, is a nightmare to analyze!)

- **When the important thing is that the users believe they were successful.** Self-reported task completion is a bit more controversial than other methods of collecting task completion, as in this case you may be using it to
determine what areas of the site have usability problems or whether a
competitive Web site or alternative design is better or worse. We’ve seen
from typical usability testing that participants often think they have found
the correct answer when they did not, or vice versa. Despite this, is the
user’s perception itself enough of a metric to judge the success of a Web
site’s user experience? It depends on the goals of your testing. In some
cases, it’s not as important that the users get an exact correct answer as it
is that they feel like they have found what they were looking for. Good
events may be any reasonably complex problem or issue (e.g., the
solutions to global warming), where there’s usually not a simple right or
wrong answer.

An example of the latter is provided by the example study comparing the Web
sites of the two major U.S. presidential candidates in the 2008 election: Barack
Obama and John McCain. In this study, participants were asked to do four tasks
on one of the two sites (e.g., find the candidate’s position on Social Security).
A politician’s position on something as complex as Social Security (a U.S.
government-funded program to provide some degree of financial support during
departure) is never going to be as simple as “Yes, I’m for it” or “No, I’m against
it.” The goal of the task is really to see if participants can learn enough about
the candidate’s position on Social Security to feel like they have a basic under-
standing of their views on the subject. In cases like this, you may want to pro-
vide a scale for participants to indicate their confidence that they accomplished
the task. In the Obama/McCain Web sites study, we provided three options for
responding to each task:

1. I definitely found the answer
2. Not sure if I found the answer or not
3. I definitely did NOT find the answer

Of course, self-report of task success can also introduce potential bias. In the
Obama/McCain study, for example, participants were assigned randomly to one
site or the other, without regard for their individual political leanings. It doesn’t
take much of a leap to see that someone who was pro-Obama might be more
generous in deciding that he had found the answer to a task on the Obama site
than on the McCain site. One way to account for that bias might have been to
ask participants a starter question about their political affiliation and/or prefer-
ence for a candidate and either sort participants into quotas so that participation
was equal across groups or segment the analysis by those groups.

If your goal is to thoroughly evaluate the usability of specific areas, first think
about the content and tasks you want to use. If users are able to know whether
they’ve found an answer, such as in the case of the presidential candidates’
positions on Social Security, then using self-reported answers should be suffi-
cient. Likewise, the example study using photo clip-art asks users to find an exact
picture. Although the answer choices are all picture titles, this could have been
implemented with a self-reported answer set. Users are likely to know whether
they’ve found the same exact picture as the one presented.
Designing the Study  CHAPTER 3

It is not a good idea to use self-reported answers if the subject matter is more complicated and more likely to cause users to have false impressions of success. Let’s take the example of one of the pets study task:

“Find the street of a PetSmart store that is in or nearest to Marietta, Georgia and offers grooming.”

This task is targeted at testing the store locator on the Web site. See Figure 3.21 as an example of this.

If the user fails to check off “full-service grooming” or doesn’t at least look for a store with grooming on the search results page, the task is incorrect. However, the participant may still feel as though he or she accomplished the task successfully. This is a case where self-reported success does not work, and it would be better to provide specific answers so that you can detect if users noticed and used all areas of the search functionality correctly.

Keep in mind some vendor tools commonly provide just the option to use two buttons, the most common being “success” or “abandon” (see Figure 3.22).

Sometimes these button names can be changed, but restricting the user to two choices may result in noisier data. If a user were to select “abandon,” there is no distinction between abandoning something because they are giving up or because they were unable to find the answer. Another way may be to either provide a drop-down list of multiple answers immediately available or change the “Success” button to something like “Answer” or “Done with Task” and then give them an answer list on the next page (see Figures 3.23 and 3.24). This example does not use an abandon button, but instead gives users an answer option to say “Give Up/Not Sure.”

FIGURE 3.21
PETCO store locator.

FIGURE 3.22
“Abandon Task” and “Task Complete” buttons used in Loop11.

FIGURE 3.23
The user clicks “Done with this Task” when they think they’ve found the answer (created in RelevantView).
Self-reported answers are perfect for more general online studies in which you’re less concerned with specific usability issues and are trying to assess the broader user experience of the site, its functionality, the brand, etc. In these true intent or open Web studies, the tasks can be even more general, such as “explore this Web site and decide when you have an understanding of what it’s about.” Some open Web studies can even give users a directed task, but lets them start at a search engine to find the answer anywhere they wish. The point here is that answers are not as important as the user’s experience when doing the task and their own assessment of how it went. In all of these cases, the study only uses one or a few general tasks as a way to get people to interact with the Web or a Web site. Then, participants can be asked a variety of questions about their experience using the site after the study.

3.5 POST-TASK QUESTIONS AND METRICS

It’s common in online usability studies to include a few questions after each task. Understanding what a user thinks about how a task went can provide insight into what specific areas of the product are working well or need attention. This section includes a discussion of some of the types of post-task ratings to collect, as well as other strategies to keep in mind when designing the study.

3.5.1 Self-reported data

There are several different types of post-task rating scales used in the field today, a few of which are covered here briefly. For a more detailed description of any of these ratings, refer to Tullis and Albert (2008).

- **Ease of use:** This is the simplest form of a rating scale asking participants how easy or difficult the task was to complete. It can be asked as a Likert scale, where a user strongly agrees or disagrees with a statement, such as:

  I found this task to be very easy:

  Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree
or it can be asked as a semantic differential where you have anchor labels, such as:

This task was:

“Very Difficult” ○ ○ ○ ○ ○ “Very Easy”

Figure 3.25 shows an example of this method.

- **After-scenario questionnaire (ASQ):** This method, developed by Jim Lewis (1991), asks a few questions after each task, each with a seven-point scale of agreement:
  - “I am satisfied with the ease of completing the tasks in this scenario.”
  - “I am satisfied with the amount of time it took to complete the tasks in this scenario.”
  - “I am satisfied with the support information (online help, messages, documentation) when completing the tasks.”

Figure 3.26 shows an example of the ASQ, as implemented in a UzerZoom study.

---

**FIGURE 3.25**

This study uses a simple semantic differential scale as a post-task rating, asking “This task was…”

**FIGURE 3.26**

The ASQ consists of three questions with Likert scales (created in UserZoom).
Expectation measures: As described earlier in this chapter, Albert and Dixon (2003) asked users to rate how easy/difficult they expected tasks to be before attempting them, and then again how easy/difficult they were after attempting tasks. This uses a seven-point scale, where 1 was “Very Difficult” and 7 was “Very Easy.” Plotting expectations against actual ease of use response gives insights of where to focus attention and opportunities for exceeding users’ expectations.

Each of these methods can be implemented in online usability studies. Expectation measures can be implemented by giving the participant a survey of expectations at the beginning of the study and then their ease of use rating again after each task. Finally, simple ease of use and ASQ questions can be implemented easily after each task.

We’ve found that although the ASQ and expectation ratings can provide insightful results, using the simple ease of use question does the trick just as well, if not better (Tedesco and Tullis, 2006).

You may want to ask users more specific rating scale questions about their experience for each task, such as how it was navigating through the interface. Just remember to keep it short, especially if you have a lot of tasks and questions to get through.

### 3.5.2 Open-ended responses

Although open-ended responses can be challenging to analyze, gathering participants’ feedback is essential. Because the study is unmoderated, you are left only to interpret data without any observational notes to guide you. Eliciting general feedback after each task can be as simple as allowing a user to enter comments or asking about their experience with a specific task. These general comments that are always available should be optional because people don’t always have feedback to give, and being forced to after every task or question can be frustrating and result in drop-off.

Figure 3.27 is an example of an open-ended response interface offered by Keynote’s WebEffective.

Another option for open-ended responses is to prompt participants to give feedback based on a certain event, such as when they find a task particularly difficult or choose to abandon or give up the task. Using conditional logic,
participants who rate a task as difficult can be assigned to answer an open-ended post-task question. Figure 3.28 is an example of this. Participants were given this question if they chose one of the bottom two box ratings of an ease-of-use scale elicited after the task.

The same conditional logic could be applied to unsuccessful task completion, but be careful about when to do this; participants should not be able to detect a pattern regarding when they have gotten a task “wrong.”

Finally, when providing the opportunity to enter post-task comments, take task times into account. If the tool you’re using includes open-ended response time as part of the time on task, this may increase task times significantly. It depends on other factors; if you offer open-ended questions equally across tasks and designs, task times would equally be affected and therefore it may not be an issue. However, if you are only allowing comments for select tasks, designs, or based on select ratings or performance, task times may be increased unevenly and can skew data.

3.6 POST-SESSION QUESTIONS AND METRICS

Post-session (or “post-study”) responses and metrics are usually collected at the end of the study, after all the tasks, in an effort to summarize the overall usability and/or user experience of the product. We’ve used them most successfully to compare one design alternative against others or one design against its competitors.

3.6.1 Overall rating scales

Some simple questions can be used just to sum up the user experience for participants. Similar to asking overall ease of use just after a task, you can ask overall ease of use regarding the entire experience on the Web site or section of the Web site. Figure 3.29 shows an example of this.

Of course, you are not getting the internal reliability of a full questionnaire on the topic, but this is a useful, simple barometer for how the experience went.
In addition to ease of use, there are other more detailed aspects of the user experience that you may want to collect data on, such as navigation, visual appeal, organization, and terminology. You can also get a feel for participants’ experiences with the site by asking questions such as whether they would return to the Web site or recommend it to a friend. These specific examples may be useful depending on the domain you’re testing and what your goals are. This is where you can get creative with those broader user experience questions. Perhaps you want to know if participants’ perception of the brand has changed after using the site or simply whether a new design is an improvement over an old one. If testing a Web site about food products, maybe you’ll want to ask questions about how looking at pictures of food makes them want to taste or buy it. Rating scale questions on these topics is a nice way to collect attitudinal and emotional reactions quantifiably, especially if you plan to use it for benchmarking or comparative purposes.

A couple of examples of these broader user experience questions are shown in Figures 3.30 and 3.31.
3.6.2 Overall assessment tools

In addition to rating scales asking about specific characteristics or features of the Web site, it’s relatively common to include a standard set of rating scales designed to assess the Web site as a whole. Several post-session questionnaires are available for free or as part of a service. Some of these include:

- System Usability Scale (SUS): 10 questions on a five-point Likert scale
- Computer System Usability Questionnaire (CSUQ): 19 questions on a seven-point Likert scale
- Questionnaire for User Interface Satisfaction (QUIS): 27 questions in the form of semantic differential questions with 10-point intervals
- Usefulness, Satisfaction, and Ease of Use Questionnaire (Lund, 2001): 30 questions on a seven-point Likert scale
- Web site Analysis and Measurement Inventory (WAMMI): 20 questions on a five-point Likert scale
- Net Promoter Score® (NPS) created by Satmetrix Systems (2008): A single question (“How likely is it that you would recommend [site or company name] to a friend or colleague?”) rated on an 11-point semantic differential scale from “Not at All Likely” to “Extremely Likely.” Just keep in mind that users’ perception of the company or brand could affect this likelihood rating.

Why would you want to include one of these standard questionnaires? Mainly because they were designed by experts in the field and, in some cases, there may be benchmark data available that you can compare the site’s results to (such as the WAMMI service). While all of these questionnaires can be useful, the SUS questionnaire has become quite popular for several reasons:

- It’s relatively short and simple (10 rating scales).
- It has been made freely available for commercial and noncommercial use as long as the source is credited.
- It has been demonstrated that it yields reliable results even with relatively small numbers of participants (Tullis and Stetson, 2004).

The SUS asks 10 questions all targeted at the same concept, and the polarity of the scale is flipped for every other question. For more information on analyzing SUS ratings, see Chapter 6.

Figure 3.32 shows an example of using SUS as part of a post-session questionnaire.
3.6.3 Open-ended questions

Finally, it is always useful to finish the session with a few open-ended questions about the experience. Instead of asking just an open-ended question on overall thoughts, participants will be more articulate with their thoughts if you ask more specific questions. Most online usability studies include one or more open-ended questions at the end of the study, such as the following.

- Were there any aspects of the Web site that you found particularly challenging or frustrating? If so, please describe.
- Were there any aspects of the Web site that you thought were particularly effective or intuitive? If so, please describe.
- Please describe anything you would do to improve this Web site.

Similar to the verbatim comments at the task level, these comments at the end of a study can be an incredibly rich source of insight into how to improve the Web site. We’ve found that generally about 60–80% of the participants who complete the tasks of a study provide at least some responses to the open-ended questions, particularly those asking about challenging aspects of the site or suggestions for improvement.

Another method commonly used in studies is to ask participants for their top likes and then a separate question for their top dislikes. Asking for the top few requires participants to think carefully about the experience enough to
provide more than one, and internally prioritize thoughts so as not to provide too many. Breaking likes and dislikes up into two questions also forces them to think about the two separately rather than just focusing on the more negative aspects of the experience.

If you plan to use “like” and “dislike” questions, be prepared to obtain feedback on all aspects of the Web site, including colors, look, and feel. These questions give participants free rein to comment. If you’re looking for more specific usability and behavioral feedback, using questions about challenging/frustrating and effective/intuitive experiences will yield more pointed results.

3.7 DEMOGRAPHIC QUESTIONS AND WRAP-UP

At the end of a session, it is common to finish with demographic questions. Then to wrap up, participants are brought to a thank-you page with some final information for them about their incentive and where they can go next.

3.7.1 Demographic questions

We typically ask demographic questions, such as age, gender, level of education, employment, and income, when there is a need to segment data and compare the sample of participants to either the general population or the targeted user population of the product. Although you may have worked with a panel that has this information on file, it’s a good idea to ask the questions again just to make sure that data are accurate and current.

We typically place demographic questions at the end of a survey or usability study. By the time participants have arrived at the demographic questions toward the end of the study, they are usually less apprehensive and have already invested a lot of time into the study. The only times we tend to place demographic questions at the beginning of the study are when they are needed for screening or branching purposes.

We recommend keeping the questions basic and trying to avoid questions of a sensitive nature if possible. Although age and gender are typical questions, many people consider them a bit sensitive so we recommend that they’re made optional. Having users type in their actual age often feels intrusive and personal to them, as can asking for a date of birth. The best method in this case is to allow people to select an age range. The appropriate age ranges might depend on market segments or other factors. If the study is being sent to the general public, try to provide as wide of a set of age ranges as possible. For example, don’t make the assumption that people over 65 are not using the Web. Likewise, if you feel that you can lump all people into one market segmentation category for the product of 55+, still make an effort to break down that 55+ category in the demographic question. Otherwise, some people may feel offended in taking the study if their category isn’t given equal “decomposition.”
An example of an age range question is shown in Figure 3.33.

Level of education, employment status, and individual or household income are other questions used commonly in surveys for segmentation purposes. Again, these types of questions are sensitive and should be made optional. As with age, income level should be presented in groups from which to choose.

Other especially sensitive questions are sometimes used in survey research, such as religious affiliation, race, or ethnicity. These should be avoided if possible. Once again, it’s not likely that these factors would impact behavioral research unless it’s related directly to the domain and usage of the product.

### 3.7.2 Wrap-up

When the study is complete and users submit answers to their demographic questions, it is customary to end with a “thank-you” page. It could be as simple as just thanking participants for their time, but it’s often useful to provide a little more information. This may include:

- Assurance that their data have been submitted
- Disclosure of the purpose of the study, if it was intentionally avoided at the beginning of the study (so as not to bias participants)
- Information about the incentive, such as when and how they will receive it, when a drawing will take place, and how they will be contacted in the event that they win
- Opportunity to enter their email and/or other contact information, either for incentive purposes (if not already collected) or to receive a summary of the results of the study
- A reminder of contact information if they have any questions or problems
- A link to close the window or redirect to another Web site (e.g., the currently available version of the Web site just tested)
- Ability to print the page, especially if there is contact information and drawing information

As always, be sure to use an appreciative tone with users and leave them feeling confident and hopefully even like they had fun!

Figure 3.34 is an example of a “Thank You” page.

### 3.8 SPECIAL TOPICS

This section provides a few additional topics related to constructing an online usability study, including progress indicators, pausing, and speed traps.
3.8.1 Progress indicators

Progress indicators give participants a sense of how far along they are within a session and the rate at which they're going. They are helpful to include, especially if there are many pages to the study. If the study is only one or two pages, as with a true intent or open Web study, it is not necessary to give indication of progress.

One common type of progress indicator implemented is a bar or graphic with incrementally more of it filled for each page complete, as in Figure 3.35. Sometimes these bars are accompanied by percentage values. If the study is 20 pages long, but each page is a quick task or question, seeing the visual indicator build up quickly assures participants that they are moving quickly through the study and allows them to gauge when they'll be finished.

Another way of indicating is to number questions and tasks, “1 of X” where X is the total number of questions or tasks. This also sets the users’ expectations of how long they will need to dedicate to the study.

Make sure that the tool used is showing progress for natural “chunks” of the study. For example, if the tool can only count each individual “element” as a question, but a single task is made up of four “elements”—task, answer, post-task rating, and post-task open-ended question—it will count as four questions. Therefore, if you have 10 tasks, it will be counted as 40 questions. Participants seeing “Question 3 of 40” can be turned off easily, when in reality it’s just 10 natural chunks to complete. In these cases, check if your tool has the ability to customize it so that you can instead say “Task X of 10” for each set of elements. If this isn’t possible, attend to this issue carefully during the pilot test and perhaps compare drop-off rates for progress bars turned both on and off.
3.8.2 Pausing

If possible, you may also want to provide the ability for users to pause the session if they’re unexpectedly interrupted. Not only will this allow participants to feel like they don’t have to rush or put their life on hold, but it will help the sample sizes and analysis. As shown in Chapter 5, because we usually filter data for excessively long task times, providing this ability may save otherwise legitimate data from being discarded.

3.8.3 Speed traps

Burke’s Panel Quality R&D (2008) has continuously monitored how “undesirable” respondents have impacted data as part of their panel quality research. They categorize undesirables by participants who misrepresent themselves in the screening process and “mental cheaters” who tend to race through surveys with little thought and effort. When looking at how undesirables’ data compared with “true” respondents on a set of survey data, they found that various attributes rated for a product were much different for undesirable participants than ratings for true participants.

In addition to putting measures in place during the screening process, it can be important to catch mental cheaters during the study using speed traps. Speed traps are ways of determining whether users are putting in a good-faith effort during the study. One tactic used by Burke and others in the field includes simply asking users to do or enter something very simple. For example, as part of the SUS questionnaire, participants might be asked to select the middle option of one of the rating scales. Speed traps work best when they are part of a group of answers or questions. An example is shown in Figure 3.36.

In one study we conducted, roughly 10% of the participants actually chose a radio button other than the middle choice! This rate differs from study to study.
as well as where speed traps are placed in the study, but nevertheless they’re still important to include.

Another speed trap is to ask the same question twice, with one worded positively and one worded negatively, for example, “This Web site was easy to use” and “This Web site was not easy to use.” These two questions should provide opposite ratings.

3.9 SUMMARY
The key to getting quality results in an online usability study is designing everything attentively—from instructions, screening and starter questions, to tasks, and follow-up questions. Remember the following points.

- The introduction should include things such as the purpose and sponsor and contact information to engage their trust, a realistic time estimate, an incentive and any other information to increase motivation, any technical requirements in clear terms, necessary informed consent and/or legal language, and straightforward instructions.
- Any screening questions should be included at the beginning of the study and asked in groups or all at once so as not to clue people into what parameters might screen them in or out. Consider including questions (or combinations of questions) designed to weed out those who are trying to misrepresent themselves. Finally, be graceful in the exit strategy so that people still feel appreciated and are likely to participate again.
- In any starter questions, such as Web site experience, Web savvy, or expectation ratings, stick to the principles of survey design. Make sure that the questions are interpreted as you intended. Try to avoid biasing participants by using distractor tasks rather than calling out any areas that you plan to study exclusively.
- Choose tasks that will cover a good cross section of the Web site or the area of the Web site you’re studying. Make tasks easy to interpret and understand. If collecting task success, use multiple-choice answers to make your analysis easier. Unless self-reported task completion suits the Web site and goals, provide a list of plausible answers that cannot be combed by process of elimination and that may even provide insight into wrong answers. Consider including a “Don’t Know/Give Up” option and a “None of the Above” option. When possible, randomize tasks to counteract ordering effects. Also consider when it’s appropriate to use self-selected and self-generated tasks.
- For more subjective data, include some self-reported post-task ratings or questions, perhaps with logic built in to ask open-ended questions in the event that a participant struggled with or didn’t like attempting a task. Consider using a post-session questionnaire, the most commonly used being the SUS. For more insight, provide a few open-ended questions, for example, those touching on intuitiveness, frustration, and/or likes, dislikes, and suggestions for improvement.
- Demographic questions are usually asked at the end, the most common of which being age group and gender. Be careful about when and how
you ask more sensitive questions such as employment, income level, and especially religion, race, and ethnicity.

- State wrap-up points kindly and appreciatively. Thank participants for their time, give them any necessary details about incentives, restate contact information, and collect their info if you want or need to contact them for study results or incentive purposes.

- Consider including some other aspects to the study, such as speed traps to eliminate mental cheaters, progress indicators, and a pausing feature if you’re collecting task times.