CHAPTER

UX Evaluation Introduction 12

Objectives

After reading this chapter, you will:

1. Understand the difference between formative and summative evaluation and the strengths and limitations of each
2. Understand the difference between analytic and empirical methods
3. Understand the difference between rigorous and rapid methods
4. Know the strengths and weaknesses of various data collection techniques, such as critical incident identification, thinking aloud, and questionnaires
5. Distinguish evaluation techniques oriented toward usability and emotional impact
6. Understand the concept of the evaluator effect and its impact on evaluation results

12.1 INTRODUCTION

12.1.1 You Are Here

We begin each process chapter with a "you are here" picture of the chapter topic in the context of the overall Wheel lifecycle template; see Figure 12-1. This chapter is an introduction that will lead us into the types and parts of UX evaluation of the following chapters.

12.1.2 Evaluate with a Prototype on Your Own Terms

Users will evaluate the interaction design sooner or later, so why not have them do it sooner—working with your team, using the proper techniques, and under the appropriate conditions—or you can wait until it is in the field, where you cannot control the outcome—visualize bad rumors about your product and huge costs to fix the problems because you have already committed the design to software.

12.1.3 Measurability of User Experience

But can you evaluate usability or user experience? This may come as a surprise, but neither usability nor user experience is directly measurable. In fact, most interesting phenomenon, such as teaching and learning, share the same

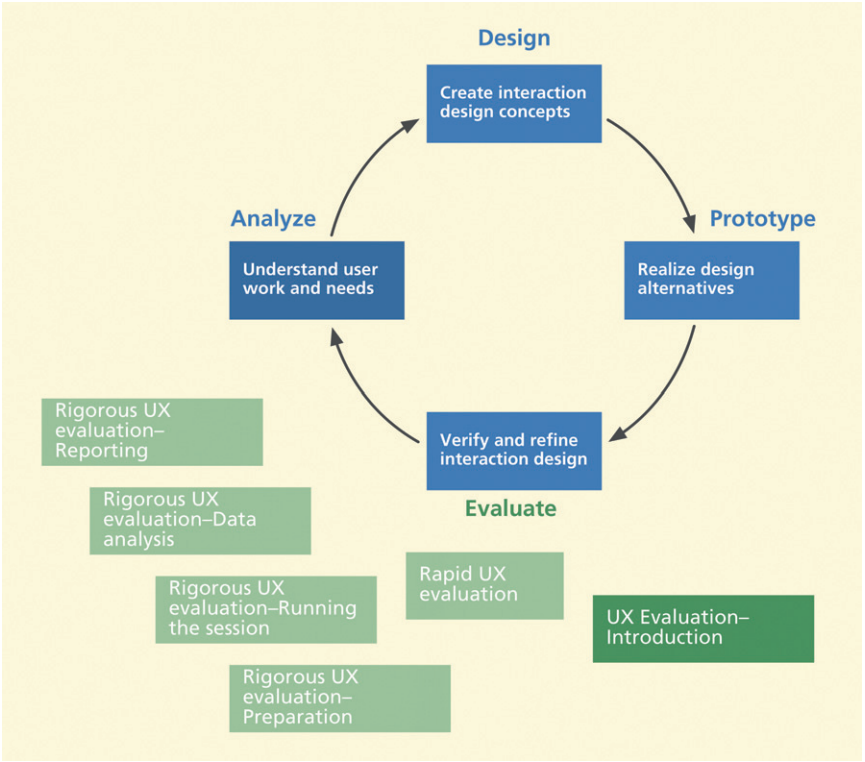

Figure 12-1
You are here, at the evaluation activity in the context of the overall Wheel lifecycle template.

difficulty. So we resort to measuring things we *can* measure and use those measurements as *indicators* of our more abstract and less measurable notions.

For example, we can understand usability effects such as productivity or ease of use by measuring observable user performance-based *indicators* such as time to task completion and error counts. You can design a feature so that the performance of a certain task in a usability lab will yield a desirable objective measurement of, say, time on task. In almost any work context this translates to good user performance.

Questionnaires also provide indicators of user satisfaction from their answers to questions we think are closely related to satisfaction. Similarly, emotional impact factors such as user satisfaction and joy of use also cannot be measured directly but only through indirect indicators.

12.1.4 User Testing? No!

Before we get into the different types of evaluation, let us first calibrate our perspective on what we are testing here. Ask yourself honestly: Do you use the term "user testing?" If you do, you are not alone: the term appears in many books

and papers on human–computer interaction (HCI) as it does in a large volume of online discussions and practitioner conversations.

You know what it means and we know what it means, but no user will like the idea of being tested and, thereby, possibly made to look ridiculous. No, we are not testing users, so let us not use those words. We might be testing or evaluating the design for usability or the full user experience it engenders in users, but *we are not testing the user.* We call it UX evaluation or even UX testing, but not "user testing!"

It might seem like a trivial PC issue, but it goes beyond being polite or "correct." When you are working with users, projecting the right attitude and making them comfortable in their role can make a big difference in how well they help you with evaluation. UX evaluation must be an ego-free process; you are improving designs, not judging users, designers, or developers.

We know of a case where users at a customer location were forced to play the user role for evaluation, but were so worried that it was a ruse to find candidates for layoffs and staff reductions that they were of no real value for the evaluation activities. If your user participants are employees of the customer organization, it is especially important to be sure they know you are not testing them. Your user participants should be made to feel they are part of a design process partnership.

12.2 FORMATIVE VS. SUMMATIVE EVALUATION

In simplest terms, formative evaluation helps you form the design and summative evaluation helps you sum up the design. A cute, but apropos, way to look at the difference: "When the cook tastes the soup, that's formative; when the guests taste the soup, that's summative" (Stake, 2004, p. 17).

The earliest reference to the terms formative evaluation and summative evaluation we know of stems from their use by Scriven (1967) in education and curriculum evaluation. Perhaps more well known is the follow-up usage by Dick and Carey (1978) in the area of instructional design. Williges (1984) and Carroll, Rosson, and Singley (1992) were among the first to use the terms in an HCI context.

Formative evaluation is primarily diagnostic; it is about collecting qualitative data to identify and fix UX problems and their causes in the design. *Summative evaluation* is about collecting quantitative data for assessing a level of quality due to a design, especially for assessing improvement in the user experience due to formative evaluation.

Qualitative Data

Qualitative data are non-numeric and descriptive data, usually describing a UX problem or issue observed or experienced during usage.

Quantitative Data

Quantitative data are numeric data, such as user performance metrics or opinion ratings.

Formal summative evaluation is typified by an empirical competitive benchmark study based on formal, rigorous experimental design aimed at comparing design hypothesis factors. Formal summative evaluation is a kind of controlled hypothesis testing with an m by n factorial design with y independent variables, the results of which are subjected to statistical tests for significance. Formal summative evaluation is an important HCI skill, but we do not cover it in this book.

Informal summative evaluation is used, as a partner of formative evaluation, for quantitatively summing up or assessing UX levels using metrics for user performance (such as the time on task), for example, as indicators of progress in UX improvement, usually in comparison with pre-established UX target levels (Chapter 10).

However, informal summative evaluation is done without experimental controls, with smaller numbers of user participants, and with only summary descriptive statistics (such as average values). We include informal summative evaluation in this book as a companion activity to formative evaluation.

12.2.1 Engineering Evaluation of UX: Formative Plus Informal Summative

> *Life is one big, long formative evaluation.*
>
> – Anonymous

Try as you might in the design phase, the first version of your interaction design is unlikely to be the best it can be in meeting your business goals of pleasing customers and your UX goals of pleasing users. Thus the reason for iteration and refinement cycles, to which evaluation is central.

You do not expect your first design to stand for long. Our friend and colleague, George Casaday calls it: "Waffle Wisdom" or "Pancake Philosophy"—like the first waffle or pancake, you expect from the start to throw away the first design, and maybe the next few. Formative evaluation is how you find out how to make the next ones better and better.

In UX engineering, formative UX evaluation includes any method that meets the definition of helping to form the design. Most, if not all, rapid UX evaluation methods (Chapter 13) have only a formative UX evaluation component and do not have a summative component. In lab-based UX testing sessions we also often use only formative evaluation, especially in early cycles of iteration when we are defining and refining the design and are not yet interested in performance numbers.

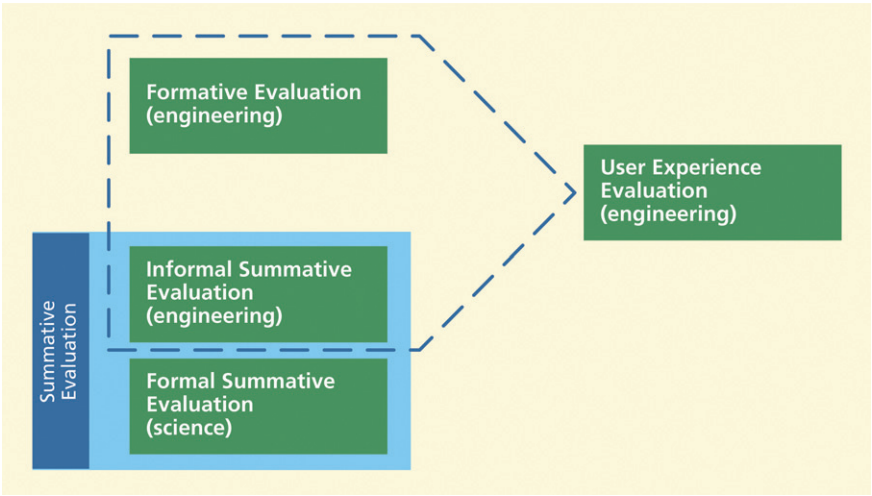

Figure 12-2
UX evaluation is a combination of formative and informal summative evaluation.

In rigorous UX evaluation we often add an informal summative evaluation component to formative evaluation, the combination being used to improve an interaction design and to assess how well it has been improved. We call this combination "UX engineering evaluation" or just "UX evaluation," as shown in Figure 12-2.

At the end of each iteration for a product version, the informal summative evaluation is used as a kind of acceptance test to compare with our UX targets and ensure that we meet our UX and business goals with the product design.

12.2.2 Engineering vs. Science

It's all very well in practice but it will never work in theory.

– French management saying

Sometimes empirical lab-based UX testing that includes quantitative metrics is the source of controversy with respect to "validity." Sometimes we hear "If you do not include formal summative evaluation, are you not missing an opportunity to add some science?" "Since your informal summative evaluation was not controlled testing, why should we not dismiss your results as too 'soft'?" "Your informal studies just are not good science. You cannot draw any conclusions."

These questions ignore the fundamental difference between formal and informal summative evaluation and the fact that they have completely different goals and methods. This may be due, in part, to the fact that the fields of HCI and UX were formed as a melting pot of people from widely varying backgrounds. From their own far-flung cultures in psychology, human factors

engineering, systems engineering, software engineering, marketing, and management they arrived at the docks of HCI with their baggage containing their own perspectives and mind-sets.

Thus, it is known that formal summative evaluations are judged on a number of rigorous criteria, such as validity, and that formal summative evaluation contributes to our *science* base. But informal summative evaluation may be less known as an important engineering tool in the HCI bag and that the *only* criterion for judging this kind of summative evaluation method is whether it works as part of an *engineering* process.

12.2.3 What Happens in Engineering Stays in Engineering

Because informal summative evaluation is engineering, it comes with some very strict limitations, particularly on sharing informal summative results.

Informal summative evaluation results are only for internal use as engineering tools to do an engineering job by the project team and cannot be shared outside the team. Because of the lack of statistical rigor, these results especially cannot be used to make any claims inside or outside the team. To make claims about UX levels achieved, for example, from informal summative results, would be a violation of professional ethics.

We read of a case where a CEO of a company got a UX report from a project team, but discounted the results because they were not statistically significant. This problem could have been avoided by following our simple rules and not distributing formative evaluation reports outside the team or by writing the report with careful caveats.

But what if you are required to produce a formative evaluation report for consumption beyond the team or what if you need results to convince the team to fix the problems you find in a formative evaluation? We address those questions and more in Chapter 17, evaluation reporting.

12.3 TYPES OF FORMATIVE AND INFORMAL SUMMATIVE EVALUATION METHODS

12.3.1 Dimensions for Classifying Formative UX Evaluation Methods

In practice, there are two orthogonal dimensions for classifying types of formative UX evaluation methods:

- empirical method vs. analytic method
- rigorous method vs. rapid method

12.3.2 Rigorous Method vs. Rapid Method

Formative UX evaluation methods can be either rigorous or rapid. We define rigorous UX evaluation methods to be those methods that maximize effectiveness and minimize the risk of errors regardless of speed or cost, meaning to refrain from shortcuts or abridgements.

Rigorous empirical UX evaluation methods entail a full process of preparation, data collection, data analysis, and reporting (Chapters 12 and 14 through 18). In practical terms, this kind of rigorous evaluation is usually conducted in the UX lab. Similarly, the same kind of evaluation can be conducted at the customer's location in the field.

Rigorous empirical methods such as lab-based evaluation, while certainly not perfect, are the yardstick by which other evaluation methods are compared. Rigorous and rapid methods exist mainly as quality vs. cost trade-offs.

- Choose a rigorous empirical method such as lab-based testing when you need effectiveness and thoroughness, but expect it to be more expensive and time-consuming.
- Choose the lab-based method to assess quantitative UX measures and metrics, such as time-on-task and error rates, as indications of how well the user does in a performance-oriented context.
- Choose lab-based testing if you need a controlled environment to limit distractions.
- Choose empirical testing in the field if you need more realistic usage conditions for ecological validity than you can establish in a lab.

Ecological Validity

Ecological validity refers to the realism with which a design of evaluation setup matches the user's real work context. It is about how accurately the design or evaluation reflects the relevant characteristics of the ecology of interaction, i.e., its context in the world or its environment.

However, UX evaluation methods can be faster and less expensive.

- Choose a rapid evaluation method for speed and cost savings, but expect it to be (possibly acceptably) less effective.
- Choose a rapid UX evaluation method for early stages of progress, when things are changing a lot, anyway, and investing in detailed evaluation is not warranted.
- Choose a rapid method, such as a design walkthrough, an informal demonstration of design concepts, as a platform for getting initial reactions and early feedback from the rest of the design team, customers, and potential users.

12.3.3 Analytic Method vs. Empirical Method

On a dimension orthogonal to rapid vs. rigorous, formative UX evaluation methods can be either empirical or analytic (Hix & Hartson, 1993; Hartson, Andre, & Williges, 2003). Empirical methods employ data observed in the performance of real user participants, usually data collected in lab-based testing.

Critical Incident

A critical incident is a UX evaluation event that occurs during user task performance or other user interaction, observed by the facilitator or other observers or sometimes expressed by the user participant, that indicates a possible UX problem. Critical incident identification is arguably the single most important source of qualitative data.

Think Aloud Technique

The think aloud technique is a qualitative data collection technique in which user participants verbally externalize their thoughts about their interaction experience, including their motives, rationale, and perceptions of UX problems. By this method, participants give the evaluator access to an understanding of their thinking about the task and the interaction design.

Analytical methods are based on looking at inherent attributes of the design rather than seeing the design in use. Many of the rapid UX evaluation methods (Chapter 13), such as design walkthroughs and UX inspection methods, are analytic methods.

Some methods in practice are a mix of analytical and empirical. For example, expert UX inspection can involve "simulated empirical" aspects in which the expert plays the role of the users, simultaneously performing tasks and "observing" UX problems.

Empirical methods are sometimes called "payoff methods" (Carroll, Singley, & Rosson, 1992; Scriven, 1967) because they are based on how a design or design change pays off in terms of real observable usage. Examples of the kind of data collected in empirical methods include quantitative user performance data, such as time on task and error rates, and qualitative user data derived from usage observation, such as UX problem data stemming from critical incident identification and think-aloud remarks by user participants. Analytical methods are sometimes called "intrinsic methods" because they are based on analyzing intrinsic characteristics of the design rather than seeing the design in use.

In describing the distinction between payoff and intrinsic approaches to evaluation, Scriven wrote an oft-quoted (Carroll, Singley, & Rosson, 1992; Gray & Salzman, 1998, p. 215) analogy featuring an axe (Scriven, 1967, p. 53): "If you want to evaluate a tool, say an axe, you might study the design of the bit, the weight distribution, the steel alloy used, the grade of hickory in the handle, etc., or you might just study the kind and speed of the cuts it makes in the hands of a good axeman," speaking of intrinsic and payoff evaluation, respectively. In Hartson, Andre, and Williges (2003) we added our own embellishments, which we paraphrase here.

Although this example served Scriven's purpose well, it also offers us a chance to make a point about the need to identify UX goals carefully before establishing evaluation criteria. Giving a UX perspective to the axe example, we note that user performance observation in payoff evaluation does not necessarily require an *expert* axeman (or axeperson). Expert usage might be one component of the vision in axe design, but it is not an exclusive requirement in payoff evaluation. UX goals depend on expected user classes of key work roles and the expected kind of usage.

For example, an axe design that gives optimum performance in the hands of an expert might be too dangerous for a novice user. For the weekend wood whacker, safety might be a UX goal that transcends firewood production, calling for a safer design that might necessarily sacrifice some efficiency. One hesitates to contemplate the metric for this case, possibly counting the number of 911

calls from a cellphone in the woods near Newport, Virginia, or the number of visits to the ER. Analogously, UX goals for a novice user of a software accounting system (e.g., TurboTax), for example, might place ease of use and data integrity (error avoidance) above sheer expert productivity.

Emotional impact factors can also be evaluated analytically. For example, a new axe in the hands of an expert might elicit an emotional response. Perhaps the axe head is made of repurposed steel from the World Trade Center—what patriotic and emotional impact that could afford! A beautiful, gleaming polished steel head, a gorgeously finished hickory wood handle, and a fine leather scabbard could elicit a strong admiration of the craftsmanship and aesthetics, as well as great pride of ownership.

Emotional impact factors can also be evaluated empirically. One need only observe the joy of use of a finely made, exquisitely sharpened axe. In a kind of think-aloud technique, the user exclaims with pleasure about the perfect balance as he or she hits the "sweet spot" with every fast-cutting stroke.

12.3.4 Where the Dimensions Intersect

Some example UX evaluation methods are shown in Figure 12-3 at the various intersections between the two dimensions empirical vs. analytic and rigorous vs. rapid.

We usually associate the rigorous empirical category with lab-based evaluation (Chapters 14 through 17), but empirical UX evaluation in a conference room or field setting can also be rigorous. The rapid evaluation methods (Chapter 13) are mostly analytic methods but at least one rapid empirical method (RITE) exists, designed to pick the low-hanging fruit at relatively low cost.

In addition, most practitioners in the field have their own versions of the lab-based method that might qualify as rapid because of severe abridgements but also still qualify as empirical because they involve data collection using participants. Rigorous analytic methods are beyond the scope of this book.

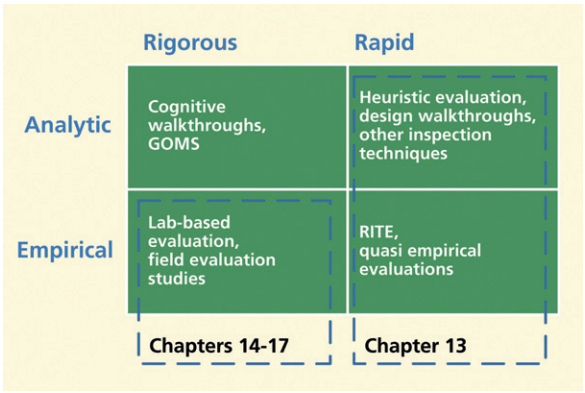

Figure 12-3
Sample UX evaluation methods at intersections between the dimensions of UX evaluation method types.

12.4 TYPES OF EVALUATION DATA

Fundamentally, UX data can be objective or subjective and it can be quantitative or qualitative. Because the two dimensions are orthogonal, you can see all four combinations,

objective and quantitative, subjective and quantitative, and so forth. When your rigorous evaluation is driven by benchmark tasks, the kinds of data collected in the process will mirror what is specified in UX targets and metrics.

12.4.1 Objective Data vs. Subjective Data

Objective UX data are data observed directly by either the evaluator or the participant. Subjective UX data represent opinions, judgments, and other subjective feedback usually from the user, concerning the user experience and satisfaction with the interaction design.

12.4.2 Quantitative Data vs. Qualitative Data

Quantitative data are numeric data, such as data obtained by user performance metrics or opinion ratings. Quantitative data are the basis of an informal summative evaluation component and help the team assess UX achievements and monitor convergence toward UX targets, usually in comparison with the specified levels set in the UX targets (Chapter 10). The two main kinds of quantitative data collected most often in formative evaluation are objective user performance data measured using benchmark tasks (Chapter 10) and subjective user-opinion data measured using questionnaires (coming up later).

Qualitative data are non-numeric and descriptive data, usually describing a UX problem or issue observed or experienced during usage. Qualitative data are usually collected via critical incident (also coming up later) and/or the think-aloud technique (see later) and are the key to identifying UX problems and their causes. Both objective and subjective data can be either qualitative or quantitative.

12.5 SOME DATA COLLECTION TECHNIQUES

12.5.1 Critical Incident Identification

The key objective of formative evaluation is to identify defects in the interaction design so that we can fix them. But during an evaluation session, you cannot always see the interaction design flaws directly. What we can observe directly or indirectly are the effects of those design flaws on the users. We refer to such effects on the users during interaction as critical incidents. Much of the attention of evaluators in evaluation sessions observing usage is spent looking for and identifying critical incidents.

Critical incidents

Despite numerous variations in procedures for gathering and analyzing critical incidents, researchers and practitioners agree about the definition of a critical incident. A critical incident is an event observed within task performance

that is a significant indicator of some factor defining the objective of the study (Anderssona & Nilsson, 1964).

In the UX literature (Castillo & Hartson, 2000; del Galdo, et al., 1986), critical incidents are indicators of "something notable" about usability or the user experience. Sometimes that notable indication is about something good in the user experience, but the way we usually use it is as an indicator of things that go wrong in the stream of interaction details, indicators of UX problems or features that should be considered for redesign.

The best kind of critical incident data are detailed, observed during usage, and associated closely with specific task performance. The biggest reason why lab-based UX testing is effective is that it captures exactly that kind of detailed usage data as it occurs.

Critical incidents are observed directly by the facilitator or other observers and are sometimes expressed by the user participant. Some evaluators wait for an obvious user error or task breakdown to record as a critical incident. But an experienced facilitator can observe a user hesitation, a participant comment in passing, a head shaking, a slight shrugging of the shoulders, or drumming of fingers on the table. A timely facilitator request for clarification might help determine if any of these subtle observations should be considered a symptom of a UX problem.

Critical incident data about a UX problem should contain as much detail as possible, including contextual information, such as:

- the user's general activity or task
- objects or artifacts involved
- the specific user intention and action that led immediately to the critical incident
- expectations of the user about what the system was supposed to do when the critical incident occurred
- what happened instead
- as much as possible about the mental and emotional state of the user
- indication of whether the user could recover from the critical incident and, if so, a description of how the user did so
- additional comments or suggested solutions to the problem

Relevance of critical incident data

Critical incident identification is arguably the single most important source of qualitative data in formative evaluation. These detailed data, perishable if not captured immediately and precisely as they arise during usage, are essential for isolating specific UX problems within the user interaction design.

History of critical incident data

The origins of the critical incident technique can be traced back at least to studies performed in the Aviation Psychology Program of the U.S. Army Air Forces in World War II to analyze and classify pilot error experiences in reading and interpreting aircraft instruments. The technique was first formally codified by the work of Fitts and Jones (1947). Flanagan (1954) synthesized the landmark critical incident technique.

Mostly used as a variation

When Flanagan designed the critical incident technique in 1954, he did not see it as a single rigid procedure. He was in favor of modifying this technique to meet different needs as long as original criteria were met. The variation occurring over the years, however, may have been more than Flanagan anticipated. Forty years after the introduction of Flanagan's critical incident technique, Shattuck and Woods (1994) reported a study that revealed that this technique has rarely been used as originally published. In fact, numerous variations of the method were found, each suited to a particular field of interest. In HCI, we have continued this tradition of adaptation by using our own version of the critical incident technique as a primary UX evaluation technique to identify UX problems and their causes.

Critical incident reporting tools

Human factors and human–computer interaction researchers have developed software tools to assist identifying and recording critical incident information. del Galdo et al. (1986) investigated the use of critical incidents as a mechanism to collect end-user reactions for simultaneous design and evaluation of both online and hard-copy documentation. As part of this work, del Galdo et al. (1986) designed a software tool to collect critical incidents from user subjects.

Who identifies critical incidents?

One factor in the variability of the critical incident technique is the issue of who makes the critical incident identification. In the original work by Fitts and Jones (1947), the user (an airplane pilot) reported the critical incidents after task performance was completed. Flanagan (1954) used trained observers to collect critical incident information while observing users performing tasks.

del Galdo et al. (1986) involved users in identifying their own critical incidents, reporting during task performance. The technique was also used as a self-reporting mechanism by Hartson and Castillo (1998, 2000) as the basis for

remote system or product usability evaluation. Further, Dzida, Wiethoff, and Arnold (1993) and Koenemann-Belliveau et al. (1994) adopted the stance that identifying critical incidents during task performance can be an individual process by either the user or an evaluator or a mutual process between the user and an evaluator.

Timing of critical incident data capture: The evaluator's awareness zone

While users are known to report major UX problems in alpha and beta testing (sending software out for comments on how well it worked), one reason these methods cannot be relied upon for thorough identification of UX problems to fix is the retrospective nature of that kind of data collection. Lab-based UX evaluation has the advantage of having the precious and volatile details right in front of you as they happen. The key to this kind of UX data is in the details, and details of these data are perishable; they must be captured immediately as they arise during usage.

As a result, do not lose this advantage; capture and document the details while they are fresh (and not just by letting the video recorder run). If you capture them as they happen, we call it concurrent data capture. If you capture data immediately after the task, we call it contemporaneous data capture. If you try to capture data after the task is well over, through someone trying to remember the details in an interview or survey after the session, this is retrospective data capture and many of the once-fresh details can be lost.

It is not as easy, however, as just capturing critical incident data immediately upon its occurrence. A critical incident is often not immediately recognized for what it is. In Figure 12-4, the evaluator's recognition of a critical incident will

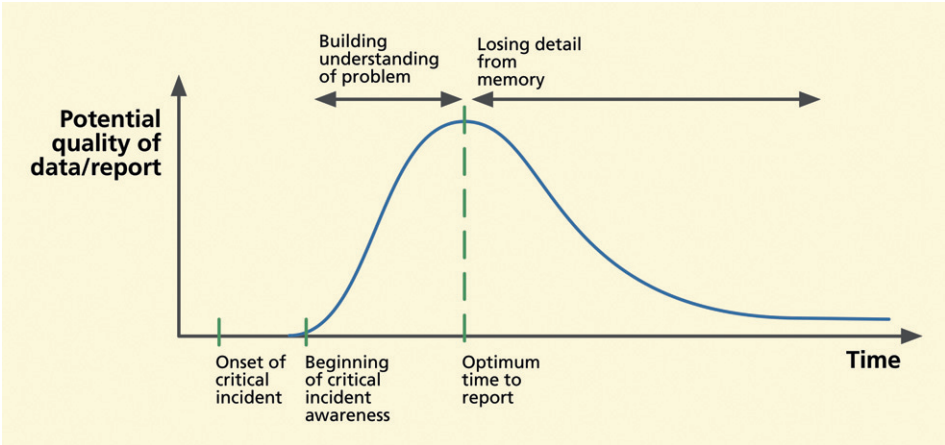

Figure 12-4
Critical incident description detail vs. time after critical incident.

necessarily occur sometime after it begins to occur. And following the point of initial awareness, after confirming that it is a critical incident, the evaluator requires some time and thought in a kind of "awareness zone" to develop an understanding of the problem, possibly through discussion with the participant.

The optimum time to report the problem, the time when the potential for a quality problem report is highest, is at the peak of this problem understanding, as seen in Figure 12-4. Before that point, the evaluator has not yet established a full understanding of the problem. After that optimum point, natural abstraction due to human memory limitations sets in and details drop off rapidly with time, accelerated by proactive interference from any intervening tasks.

12.5.2 The Think-Aloud Technique

Also called "think-aloud protocol" or "verbal protocol," the think-aloud technique is a qualitative data collection technique in which user participants, as the name implies, express verbally their thoughts about their interaction experience, including their motives, rationale, and perceptions of UX problems. By this method, participants let us in on their thinking, giving us access to a precious understanding of their perspective of the task and the interaction design, their expectations, strategies, biases, likes, and dislikes.

Why use the think-aloud technique?

General observational data are important during an evaluation session with a participant attempting to perform a task. You can see what parts of a task the user is having trouble with, you can see hesitations about using certain widgets, and so on. But the bulk of real UX problem data is hidden from observation, in the mind of the participant. What is really causing a hesitation and why does this participant perceive it as a problem or barrier? To get the best qualitative data, you have to tap into this hidden data, buried in the participant's mind, which is the goal of the think-aloud technique.

The think-aloud technique is simple to use, for both analyst and participant. It is useful for when a participant walks through a prototype or helps you with a UX inspection. Nielsen (1993, p. 195) says "thinking aloud may be the single most valuable usability engineering method." It is effective in accessing user intentions, what they are doing or are trying to do, and their motivations, the reasons why they are doing any particular actions. The think-aloud technique is also effective in assessing emotional impact because emotional impact is felt internally and the internal thoughts and feelings of the user are exactly what the think-aloud technique accesses for you.

The think-aloud technique can be used in both rigorous empirical methods (lab-based) and rapid empirical methods (quasi-empirical and RITE)—that is, any UX evaluation method that involves a user "participant." Variations of this simple technique are rooted in psychological and human factors experimentation well before it was used in usability engineering (Lewis, 1982).

What kind of participant works best?

Some participants can talk while working; get them if you can. The usual participant for think-aloud techniques is someone who matches the work role and user class definitions associated with the tasks you will use to drive the evaluation. This kind of participant will not be trained as a UX practitioner, but that usually will not deter them from offering opinions and theories about UX problems and causes in your design, which is what you want.

You must remember, however, that it is *your* job to accept their comments as inputs to your process and it is still up to you to filter and interpret all think-aloud data in the context of your design. Participants and think-aloud techniques are not a substitute for you doing your job.

So, if think-aloud participants are not typically trained UX practitioners, what about using participants who are? You can use trained UX practitioners as participants, if they are not stakeholders in your project. They will give a different perspective on your design, often reflecting a better and deeper understanding and analysis. But their analysis may not be accurate from a work-domain or task perspective, so you are still responsible for filtering and interpreting their comments.

Is thinking aloud natural for participants?

It depends on the participant. Some people find it natural to talk while they work. Some people are able and only too willing to share their thoughts while working or while doing anything. Others are naturally quiet or contemplative and must be prompted to verbalize their thoughts as they work.

Some people, in the cannot-walk-and-chew-gum category, have difficulty expressing themselves while performing a physical task—activities that require help from different parts of the brain. For these people, you should slow things down and let them "rest" and talk only between actions. Even the most loquacious of participants at times will need prompting, "what are you thinking about?" or "what do you think you should do next?"

Also, sometimes participants ask questions, such as "what if I click on this?," to which your reply should encourage them to think it through themselves, "what do you think will happen?" When an interaction sequence leads a participant to

surprise, confusion, or bewilderment—even fleetingly—ask them, "was that what you expected?" or "what *did* you expect?"

How to manage the think-aloud protocol?

The think-aloud technique is intended to draw out cognitive activity, not to confirm observable behavior. Therefore, your instructions to participants should emphasize telling you what they are thinking, not describing what they are doing. You want them to tell you why they are taking the actions that you can observe.

Once your participants get into thinking aloud, they may tend to keep the content at a chatty conversational level. You may need to encourage them to get past the "it is fine" stage and get down into real engagement and introspection. And sometimes you have to make an effort to keep the think-aloud comments flowing; some participants will not naturally maintain thinking aloud while they work and will have to be prodded gently.

Seeing a participant sit there silently struggling with an unknown problem tells us nothing about the problem or the design. Because we are trying to extract as much qualitative data as possible from each participant, elicitation of participant thoughts is a valuable facilitator skill. You might even consider a brief practice lesson on thinking aloud with each participant before you start the session itself.

Retrospective think-aloud techniques

If, as facilitator, you perceive that the think-aloud technique, when used concurrently with task performance, is interfering in some way with task performance or task measurements, you can wait until after task completion (hence the name "retrospective") and review a video recording of the session with the participant, asking for more complete "thinking aloud" during this review of his or her own task performance. In this kind of retrospective think-aloud technique, the participant is acting less as a task performer and more as an observer and outside "commentator," but with detailed inside information. The audio of this verbal review can be recorded on a second track in synchronism with the video, for even further later analysis, if necessary.

This approach has the advantage of capturing the maximum amount of think-aloud data but has the obvious downside of requiring a total of at least twice as much time as just the session itself. It also suffers from the time lag after the actual events. While better than a retrospective review even later, some details will already be fading from the participant's memory.

Co-discovery think-aloud techniques

Using a single participant is the dominant paradigm in usability and UX testing literature. Single users often represent typical usage and you want to be sure that single users can work their way through the tasks. However, you may also wish to try the increasingly common practice of using two or more participants in a team approach, a technique that originated with O'Malley, Draper, and Riley (1984). Kennedy (1989) named it "co-discovery" and that name has stuck.

While it can seem unnatural and inhibiting for a lone participant to be thinking aloud, essentially talking to oneself, there is more ease in talking in a natural conversation with another person (Wildman, 1995). A single individual participant can have trouble remembering to verbalize, but it is just natural with a partner. When the other person is also verbalizing in a problem-solving context, it amounts to a real and natural conversation, making this approach increasingly popular with practitioners and organizations.

Hackman and Biers (1992) found that multiple participants, while slightly more expensive, resulted in more time spent in verbalizing and, more importantly, participant teams spent more time verbalizing statements that had high value as feedback for designers.

Co-discovery is an especially good method for early low-fidelity prototypes; it gets you more viewpoints. And there is less need for the facilitator to interact, intervene, or give hints. The participants ask each other the questions and give themselves the guidance and prodding. When one participant gets stuck, the other can suggest things to try. When two people act as a real team, they are more willing to try new things and less intimidated than if they had been working solo.

Co-discovery pays off best when their thinking aloud becomes an interactive conversation between the participants, but this can produce qualitative data at a prodigious rate, at times more than twice as fast as from one participant. Two co-participants can bounce thoughts and comments back and forth. You may have to switch from zone defense, where each data collector does their best to catch what comes their way, to a person-to-person arrangement where each data collector is assigned to focus on comments by just one of the participants. This is where video capture makes for a good back-up to review selectively if you think you might have missed something.

There are many ways for a co-discovery session to play out. The scenarios often depend on participant personalities and who takes the lead. You should let them take turns at driving; both still have to pay attention. In cases where one participant has a dominant personality to the point where that person wants

to run things and perhaps thinks he or she knows it all, try to get the other participant to drive as much as possible, to give them some control. If one person seems to drift off and lose attention or interest, you may have to use the usual techniques for getting school children re-engaged in an activity, "Johnny, what do *you* think about that problem?"

Finally, as a very practical bonus from planning a co-discovery session, if one participant does not show up, you can still do the session. You avoid the cost of an idle slot in the lab and having to reschedule.

Does thinking aloud affect quantitative task performance metrics in lab-based evaluation?

It depends on the participant. Some participants can naturally chat about what they are doing as they work. For these participants, the concurrent think-aloud technique usually does not affect task performance when used with measured benchmark tasks.

This is especially true if the participant is just thinking aloud and not engaged with questions and answers by the facilitator. But for some participants, the think-aloud protocol does affect task performance. This is especially true for non-native speakers because their verbalizations just take longer.

12.5.3 Questionnaires

A questionnaire is the primary instrument for collecting subjective data from participants in all types of evaluations. It can be used to supplement objective (directly observable) data from lab-based or other data collection methods or as an evaluation method on its own. A questionnaire can contain probing questions about the total user experience. Although post-session questionnaires have been used primarily to assess user satisfaction, they can also contain effective questions oriented specifically toward evaluating broader emotional impact and usefulness of the design.

Questionnaires are a self-reporting data collection technique and, as Shih and Liu (2007) say, semantic differential questionnaires (see next section) are used most commonly because they are a product-independent method that can yield reliable quantitative subjective data. This kind of questionnaire is inexpensive to administer but requires skill to produce so that data are valid and reliable.

In the days of traditional usability, questionnaires were used mostly to assess self-reported user satisfaction. And they were "seen as a poor cousin to [measures of] efficiency" (Winograd & Flores, 1986), but Lund (2001, 2004), points out that subjective metrics, such as the kind one gets from questionnaire results, are often effective at getting at the core of the user experience and can

access "aspects of the user experience that are most closely tied to user behavior and purchase decisions."

Semantic differential scales

A semantic differential scale, or Likert scale (1932), is a range of values describing an attribute. Each value on the scale represents a different level of that attribute. The most extreme value in each direction on the scale is called an anchor. The scale is then divided, usually in equal divisions, with points between the anchors that divide up the difference between the meanings of the two anchors.

The number of discrete points we have on the scale between and including the anchors is the granularity of the scale, or the number of choices we allow users in expressing their own levels of the attribute. The most typical labeling of a point on a scale is a verbal label with an associated numeric value but it can also be pictorial.

For example, consider the following statement for which we wish to get an assessment of agreement by the user: "The checkout process on this Website was easy to use." A corresponding semantic differential scale for the "agreement" attribute to assess the user's level of agreement might have these anchors: strongly agree and strongly disagree. If the scale has five values, including the anchors, there are three points on the scale between the anchors. For example, the agreement scale might include strongly agree, agree, neutral, disagree, and strongly disagree with the associated values, respectively, of $+2$, $+1$, 0, -1, and -2.

The Questionnaire for User Interface Satisfaction (QUIS)

The QUIS, developed at the University of Maryland (Chin, Diehl, & Norman, 1988) is one of the earliest available user satisfaction questionnaires for use with usability evaluation. It was the most extensive and most thoroughly validated questionnaire at the time of its development for determining subjective interaction design usability.

The QUIS is organized around such general categories as *screen*, *terminology and system information*, *learning*, and *system capabilities*. Within each of these general categories are sets of questions about detailed features, with Likert scales from which a participant chooses a rating. It also elicits some demographic information, as well as general user comments about the interaction design being evaluated. Many practitioners supplement the QUIS with some of their own questions, specific to the interaction design being evaluated.

The original QUIS had 27 questions (Tullis & Stetson, 2004), but there have been many extensions and variations. Although developed originally for screen-based designs, the QUIS is resilient and can be extended easily, for example, by replacing the term "system" with "Website" and "screen" with "Web page."

Practitioners are free to use the results of a QUIS questionnaire in any reasonable way. In much of our use of this instrument, we calculated the average scores, averaged over all the participants and all the questions in a specified subset of the questionnaire. Each such subset was selected to correspond to the goal of a UX target, and the numeric value of this score averaged over the subset of questions was compared to the target performance values stated in the UX target table.

Although the QUIS is quite thorough, it can be administered in a relatively short time. For many years, a subset of the QUIS was our own choice as the questionnaire to use in both teaching and consulting.

The QUIS is still being updated and maintained and can be licensed[1] for a modest fee from the University of Maryland Office of Technology Liaison. In Table 12-1 we show a sample excerpted and adapted with permission from the QUIS with fairly general applicability, at least to desktop applications.

[1] http://lap.umd.edu/quis/

Table 12-1
An excerpt from QUIS, with permission

User Evaluation of Interactive Computer Systems

For each question, please circle the number that most appropriately reflects your impressions about this topic with respect to using this computer system or product.

1. Terminology relates to task domain	[distantly]	0 1 2 3 4 5 6 7 8 9 10	[closely]	NA
2. Instructions describing tasks	[confusing]	0 1 2 3 4 5 6 7 8 9 10	[clear]	NA
3. Instructions are consistent	[never]	0 1 2 3 4 5 6 7 8 9 10	[always]	NA
4. Operations relate to tasks	[distantly]	0 1 2 3 4 5 6 7 8 9 10	[closely]	NA
5. Informative feedback	[never]	0 1 2 3 4 5 6 7 8 9 10	[always]	NA
6. Display layouts simplify tasks	[never]	0 1 2 3 4 5 6 7 8 9 10	[always]	NA
7. Sequence of displays	[confusing]	0 1 2 3 4 5 6 7 8 9 10	[clear]	NA
8. Error messages are helpful	[never]	0 1 2 3 4 5 6 7 8 9 10	[always]	NA
9. Error correction	[confusing]	0 1 2 3 4 5 6 7 8 9 10	[clear]	NA
10. Learning the operation	[difficult]	0 1 2 3 4 5 6 7 8 9 10	[easy]	NA
11. Human memory limitations	[overwhelmed]	0 1 2 3 4 5 6 7 8 9 10	[are respected]	NA
12. Exploration of features	[discouraged]	0 1 2 3 4 5 6 7 8 9 10	[encouraged]	NA
13. Overall reactions	[terrible]	0 1 2 3 4 5 6 7 8 9 10	[wonderful]	NA
	[frustrating]	0 1 2 3 4 5 6 7 8 9 10	[satisfying]	NA
	[uninteresting]	0 1 2 3 4 5 6 7 8 9 10	[interesting]	NA
	[dull]	0 1 2 3 4 5 6 7 8 9 10	[stimulating]	NA
	[difficult]	0 1 2 3 4 5 6 7 8 9 10	[easy]	NA

The System Usability Scale (SUS)

The SUS was developed by John Brooke while at Digital Equipment Corporation (Brooke, 1996) in the United Kingdom. The SUS questionnaire contains 10 questions. As an interesting variation from the usual questionnaire, the SUS alternates positively worded questions with negatively worded questions to prevent quick answers without the responder really considering the questions.

The questions are presented as simple declarative statements, each with a five-point Likert scale anchored with "strongly disagree" and "strongly agree" and with values of 1 through 5. These 10 statements are (used with permission):

- I think that I would like to use this system frequently
- I found the system unnecessarily complex
- I thought the system was easy to use
- I think that I would need the support of a technical person to be able to use this system
- I found the various functions in this system were well integrated
- I thought there was too much inconsistency in this system
- I would imagine that most people would learn to use this system very quickly
- I found the system very cumbersome to use
- I felt very confident using the system
- I needed to learn a lot of things before I could get going with this system

The 10 items in the SUS were selected from a list of 50 possibilities, chosen for their perceived discriminating power.

The bottom line for the SUS is that it is robust, extensively used, widely adapted, and in the public domain. It has been a very popular questionnaire for complementing a usability testing session because it can be applied at any stage in the UX lifecycle and is intended for practical use in an industry context. The SUS is technology independent; can be used across a broad range of kinds of systems, products, and interaction styles; and is fast and easy for both analyst and participant. The single numeric score (see later) is easy to understand by everyone. Per Usability Net (2006), it is the most highly recommended of all the publically available questionnaires.

There is theoretical debate in the literature about the dimensionality of SUS scoring methods (Bangor, Kortum, & Miller, 2008; Borsci, Federici, & Lauriola, 2009; J. Lewis & Sauro, 2009). However, the practical bottom line for the SUS, regardless of these formal conclusions, is that the unidimensional approach to scoring of SUS (see later) has been working well for many

practitioners over the years and is seen as a distinct advantage. The single score that this questionnaire yields is understood easily by almost everyone.

The analysis of SUS scores begins with calculating the single numeric score for the instance of the questionnaire marked up by a participant. First, for any unanswered items, assign a middle rating value of 3 so that it does not affect the outcome on either side.

Next we calculate the adjusted score for positively worded items. Because we want the range to begin with 0 (so that the overall score can range from 0), we shift the scores for positively worded items down by subtracting 1, giving us a new range of 0 to 4.

To calculate the adjusted score for negatively worded items, we must compensate for the fact that these scales go in the opposite direction of positively worded scales. We do this by giving the negatively worded items an adjusted score of 5 minus the rating value given, also a giving us a range of 0 to 4.

Next, add up the adjusted item scores for all 10 questions, giving a range of 0 to 40. Finally, multiply by 2.5 to get a final SUS score in the range of 0 to 100.

What about interpreting the SUS score? Often a numerical score yielded by any evaluation instrument is difficult to interpret by anyone outside the evaluation team, including project managers and the rest of your project team. Given a single number out of context, it is difficult to know what it means about the user experience. The score provided by the SUS questionnaire, however, has the distinct advantage of being in the range of zero to 100.

By using an analogy with numerical grading schemes in schools based on a range of 0 to 100, Bangor, Kortum, and Miller (2008) found it practical and feasible to extend the school grading interpretation of numeric scores into letter grades, by the usual 90 to 100 being an "A", and so on (using whatever numeric to letter grade mapping you wish). Although this translation has no theoretical or empirical basis, this simple notion does seem to be an effective way to communicate the results, and using a one-number score normalized to a base of 100 allows you even to compare systems that are dissimilar.

Clearly, an evaluation grade of "A" means it was good and an evaluation of "D" or lower means the need for some improvement is indicated. At the end of the day, each project team will have to decide what the SUS scores mean to them.

The Usefulness, Satisfaction, and Ease of Use (USE) Questionnaire

With the goal of measuring the most important dimensions of usability for users across many different domains, Lund (2001, 2004) developed USE, a questionnaire for evaluating the user experience on three dimensions:

usefulness, satisfaction, and ease of use. USE is based on a seven-point Likert scale.

Through a process of factor analysis and partial correlation, the questions in Table 12-2 were chosen for inclusion in USE per Lund. As the questionnaire is still under development, this set of questions is a bit of a moving target.

The bottom line for USE is that it is widely applicable, for example, to systems, products, and Websites, and has been used successfully. It is available in the public domain and has good face validity for both users and practitioners, that is, it looks right intuitively and people agree that it should work.

Other questionnaires

Here are some other questionnaires that are beyond our scope but might be of interest to some readers.

Table 12-2
Questions in USE questionnaire

Usefulness	It helps me be more effective. It helps me be more productive. It is useful. It gives me more control over the activities in my life. It makes the things I want to accomplish easier to get done. It saves me time when I use it. It meets my needs. It does everything I would expect it to do.
Ease of use	It is easy to use. It is simple to use. It is user-friendly. It requires the fewest steps possible to accomplish what I want to do with it. It is flexible. Using it is effortless. I can use it without written instructions. I do not notice any inconsistencies as I use it. Both occasional and regular users would like it. I can recover from mistakes quickly and easily. I can use it successfully every time.
Ease of learning	I learned to use it quickly. I easily remember how to use it. It is easy to learn to use it. I quickly became skillful with it.
Satisfaction	I am satisfied with it. I would recommend it to a friend. It is fun to use. It works the way I want it to work. It is wonderful. I feel I need to have it. It is pleasant to use.

General-purpose usability questionnaires:

- Computer System Usability Questionnaire (CSUQ), developed by Jim Lewis (1995, 2002) at IBM, is well-regarded and available in the public domain.
- Software Usability Measurement Inventory (SUMI) is "a rigorously tested and proven method of measuring software quality from the end user's point of view" (Human Factor Research Group, 1990).[2] According to Usability Net,[3] SUMI is "a mature questionnaire whose standardization base and manual have been regularly updated." It is applicable to a range of application types from desk-top applications to large domain-complex applications.
- After Scenario Questionnaire (ASQ), developed by IBM, is available in the public domain (Bangor, Kortum, & Miller, 2008, p. 575).
- Post-Study System Usability Questionnaire (PSSUQ), developed by IBM, is available in the public domain (Bangor, Kortum, & Miller, 2008, p. 575).

Web evaluation questionnaires:

- Website Analysis and MeasureMent Inventory (WAMMI) is "a short but very reliable questionnaire that tells you what your visitors think about your web site" (Human Factor Research Group, 1996b).

Multimedia system evaluation questionnaires:

- Measuring the Usability of Multi-Media Systems (MUMMS) is a questionnaire "designed for evaluating quality of use of multimedia software products" (Human Factor Research Group, 1996a).

Hedonic quality evaluation questionnaires:

- The Lavie and Tractinsky (2004) questionnaire
- The Kim and Moon (1998) questionnaire with differential emotions scale

Modifying questionnaires for your evaluation

As an example of adapting a data collection technique, you can make up a questionnaire of your own or you can modify an existing questionnaire for your own use by:

- choosing a subset of the questions
- changing the wording in some of the questions
- adding questions of your own to address specific areas of concern
- using different scale values

---

[2]Human Factors Research Group (http://www.ucc.ie/hfrg/) questionnaires are available commercially as a service, on a per report basis or for purchase, including scoring and report-generating software.

[3]http://www.usabilitynet.org/tools/r_questionnaire.htm

On any questionnaire that does not already have its scale values centered on zero, you might consider making the scale something such as −2, −1, 0, 1, 2 to center it on the neutral value of zero. If the existing scale has an odd number of rating points, you can change it to an even number to force respondents to choose one side or the other of a middle value, but that is not essential here.

Finally, one of the downsides of any questionnaire based only on semantic differential scales is that it does not allow the participant to give indications of *why* any rating is given, which is important for understanding what design features work and which ones do not, and how to improve designs. Therefore, we recommend you consider supplementing key questions (or do it once at the end) with a free-form question, such as "If notable, please describe why you gave that rating."

Modifying the Questionnaire for User Interface Satisfaction. We have found an adaptation of the QUIS to work well. In this adaptation, we reduce the granularity of the scale from 12 choices (0 through 10 and NA) to 6 (–2, –1, 0, 1, 2, and NA) for each question, reducing the granularity of choices faced by the participant. We felt a midscale value of zero was an appropriately neutral value, while negative scale values corresponded to negative user opinions and positive scale values corresponded to positive user opinions. Some argue for an even number of numeric ratings to force users to make positive or negative choices. This is also an easy adaptation to the scale.

Modifying the System Usability Scale. In the course of their study of SUS, Bangor, Kortum, and Miller (2008) provided an additional useful item for the questionnaire that you can use as an overall quality question, based on an adjective description. Getting away from the "strongly disagree" and "strongly agree" anchors, this adjective rating statement is: "Overall, I would rate the user-friendliness of this product as worst imaginable, awful, poor, ok, good, excellent, and best imaginable."

Not caring for the term "user-friendliness," we would add the recommendation to change that phrase to something else that works well for you. In studies by Bangor, Kortum, and Miller (2008), ratings assigned to this one additional item correlated well with scores of the original 10 items in the questionnaire. So, for the ultimate in inexpensive evaluation, this one questionnaire item could be used as a soft estimator of SUS scores.

In application, most users of the SUS recommend a couple of minor modifications. The first is to substitute the term "awkward" for the term "cumbersome" in item 8. Apparently, in practice, there has been uncertainty, especially among participants who were not native English speakers, about the meaning of "cumbersome" in this context. The second modification is to

Semantic Differential Scale

A semantic differential, or Likert, scale is a range of values describing an attribute that is the focus of a question in a questionnaire. The extreme values of the attribute are called anchors and other discrete points on the scale divide up the difference between the meanings of the two anchors. Users choose values on the scale to give ratings in answering the questionnaire question.

substitute the term "product" for the term "system" in each item, if the questionnaire is being used to evaluate a commercial product.

Along these same lines, you should substitute "Website" for "system" when using the SUS to evaluate a Website. However, use caution when choosing the SUS as a measuring instrument for evaluating Websites. According to Kirakowski and Murphy (2009), the SUS is inappropriate for evaluating Websites because it tends to yield erroneously high ratings. They recommend using the WAMMI instead (mentioned previously). As one final caveat about using the SUS, Bangor, Kortum, and Miller (2008) warn that in an empirical study they found that SUS scores did not always correlate with their observations of success in task performance.

Warning: Modifying a questionnaire can damage its validity. At this point, the purist may be worried about validity. Ready-made questionnaires are usually created and tested carefully for statistical validity. A number of already developed and validated questionnaires are available for assessing usability, usefulness, and emotional impact.

For most things in this book, we encourage you to improvise and adapt and that includes questionnaires. However, you must do so armed with the knowledge that any modification, especially by one not expert in making questionnaires, carries the risk of undoing the questionnaire validity. The more modifications, the more the risk. The methods for, and issues concerning, questionnaire validation are beyond the scope of this book.

Because of this risk to validity, homemade questionnaires and unvalidated modifications to questionnaires are not allowed in summative evaluation but are often used in formative evaluation. This is not an invitation to be slipshod; we are just allowing ourselves to not have to go through validation for sensible modifications made responsibly. Damage resulting from unvalidated modifications is less consequential in formative evaluation. UX practitioners modify and adapt existing questionnaires to their own formative needs, usually without much risk of damaging validity.

12.5.4 Data Collection Techniques Especially for Evaluating Emotional Impact

Shih and Liu (2007), citing Dormann (2003), describe emotional impact in terms of its indicators: "Emotion is a multifaceted phenomenon which people deliver through feeling states, verbal and non-verbal languages, facial expressions, behaviors, and so on." Therefore, these are the things to "measure" or at least observe or ask about. Tullis and Albert devote an entire chapter (2008, Chapter 7) to the subject. For a "Usability Test Observation Form," a

comprehensive list of verbal and non-verbal behaviors to be noted during observation, see Tullis and Albert (2008, p. 170).

Indicators of emotional impact are usually either self-reported via verbal techniques, such as questionnaires, or physiological responses observed and measured in participants with non-verbal techniques.

Self-reported indicators of emotional impact

While extreme reactions to a bad user experience can be easy to observe and understand, we suspect that the majority of emotional impact involving aesthetics, emotional values, and simple joy of use may be perceived and felt by the user but not necessarily by the evaluator or other practitioner. To access these emotional reactions, we must tap into the user's subjective feelings; one effective way to do that is to have the user or participant do the reporting. Thus, verbal participant self-reporting techniques are a primary way that we collect emotional impact indicators.

Participants can report on emotional impact within their usage experience during usage via their direct commentary collected with the think-aloud technique. The think-aloud technique is especially effective in accessing the emotional impact within user experience because users can describe their own feelings and emotional reactions and can explain their causes in the usage experience.

Questionnaires, primarily those using semantic differential scales, are also an effective and frequently used technique for collecting self-reported retrospective emotional impact data by surveying user opinions about specific predefined aspects of user experience, especially emotional impact.

Other self-reporting techniques include written diaries or logs describing emotional impact encounters within usage experience. As a perhaps more spontaneous alternative to written reports, participants can report these encounters via voice recorders or phone messages.

Being subjective, quantitative, and product independent, questionnaires as a self-reporting technique have the advantages of being easy to use for both practitioners and users, inexpensive, applicable from earliest design sketches and mockups to fully operational systems, and high in face validity, which means that intuitively they seem as though they should work (Westerman, Gardner, & Sutherland, 2006).

However, self-reporting can be subject to bias because human users cannot always access the relevant parts of their own emotions. Obviously, self-reporting techniques depend on the participant's ability for conscious awareness of subjective emotional states and to articulate the same in a report.

Questionnaires as a verbal self-reporting technique for collecting emotional impact data (AttrakDiff and others)

Questionnaires about emotional impact allow you to pose to participants probing questions based on any of the emotional impact factors, such as joy of use, fun, and aesthetics, offering a way for users to express their feelings about this part of the user experience.

AttrakDiff, developed by Hassenzahl, Burmester, and Koller (2003), is an example of a questionnaire especially developed for getting at user perceptions of emotional impact. AttrakDiff (now AttrakDiff2), based on Likert (semantic differential) scales, is aimed at evaluating both pragmatic (usability plus usefulness) and hedonic[4] (emotional impact) quality in a product or system.

Reasons for using the AttrakDiff questionnaire for UX data collection include the following:

- AttrakDiff is freely available.
- AttrakDiff is short and easy to administer, and the verbal scale is easy to understand (Hassenzahl, Beu, & Burmester, 2001; Hassenzahl, et al., 2000).
- AttrakDiff is backed with research and statistical validation. Although only the German-language version of AttrakDiff was validated, there is no reason to believe that the English version will not also be effective.
- AttrakDiff has a track record of successful application.

With permission, we show it in full in Table 12-3 as it appears in Hassenzahl, Schöbel, and Trautman (2008, Table 1).

Across the many versions of AttrakDiff that have been used and studied, there are broad variations in the number of questionnaire items, the questions used, and the language for expressing the questions (Hassenzahl et al., 2000). Table 12-4 contains a variation of AttrakDiff developed by Schrepp, Held, and Laugwitz (2006), shown here with permission.

For a description of using AttrakDiff in an affective evaluation of a music television channel, see Chorianoipoulos and Spinellis (2004).

Once an AttrakDiff questionnaire has been administered to participants, it is time to calculate the average scores. Begin by adding up all the values given by the participant, excluding all unanswered questions. If you used a numeric scale of 1 to 7 between the anchors for each question the total will be in the range of 1 to 7 times the number of questions the participant answered.

[4] "Hedonic" is a term used mainly in the European literature that means about the same as emotional impact.

Table 12-3

AttrakDiff emotional impact questionnaire as listed by Hassenzahl, Schöbel, and Trautman (2008), with permission

Scale Item	Semantic Anchors	
Pragmatic Quality 1	Comprehensible	Incomprehensible
Pragmatic Quality 2	Supporting	Obstructing
Pragmatic Quality 3	Simple	Complex
Pragmatic Quality 4	Predictable	Unpredictable
Pragmatic Quality 5	Clear	Confusing
Pragmatic Quality 6	Trustworthy	Shady
Pragmatic Quality 7	Controllable	Uncontrollable
Hedonic Quality 1	Interesting	Boring
Hedonic Quality 2	Costly	Cheap
Hedonic Quality 3	Exciting	Dull
Hedonic Quality 4	Exclusive	Standard
Hedonic Quality 5	Impressive	Nondescript
Hedonic Quality 6	Original	Ordinary
Hedonic Quality 7	Innovative	Conservative
Appeal 1	Pleasant	Unpleasant
Appeal 2	Good	Bad
Appeal 3	Aesthetic	Unaesthetic
Appeal 4	Inviting	Rejecting
Appeal 5	Attractive	Unattractive
Appeal 6	Sympathetic	Unsympathetic
Appeal 7	Motivating	Discouraging
Appeal 8	Desirable	Undesirable

For example, because there are 22 questions in the sample in Table 12-3, the total summed-up score will be in the range of 22 to 154 if all questions were answered. If you used a scale from −3 to +3 centered on zero, the range for the sum of 22 question scores would be −66 to +66. The final result for the questionnaire is the average score per question.

Modifying AttrakDiff. In applying the AttrakDiff questionnaire in your own project, you can first make a choice among the different existing versions of AttrakDiff. You can then choose how many of those questions or items, and which ones, you wish to have in your version.

Table 12-4

A variation of the AttrakDiff emotional impact questionnaire, as listed in Appendix A1 of Schrepp, Held, and Laugwitz (2006), reordered to group related items together, with permission

Scale	Item	English Anchor 1	English Anchor 2
Pragmatic quality	PQ1	People centric	Technical
Pragmatic quality	PQ2	Simple	Complex
Pragmatic quality	PQ3	Practical	Impractical
Pragmatic quality	PQ4	Cumbersome	Facile
Pragmatic quality	PQ5	Predictable	Unpredictable
Pragmatic quality	PQ6	Confusing	Clear
Pragmatic quality	PQ7	Unmanageable	Manageable
Hedonic – identity	HQI1	Isolates	Connects
Hedonic – identity	HQI2	Professional	Unprofessional
Hedonic – identity	HQI3	Stylish	Lacking style
Hedonic – identity	HQI4	Poor quality	High quality
Hedonic – identity	HQI5	Excludes	Draws you in
Hedonic – identity	HQI6	Brings me closer to people	Separates me from people
Hedonic – identity	HQI7	Not presentable	Presentable
Hedonic – stimulation	HQS1	Original	Conventional
Hedonic – stimulation	HQS2	Unimaginative	Creative
Hedonic – stimulation	HQS3	Bold	Cautious
Hedonic – stimulation	HQS4	Innovative	Conservative
Hedonic – stimulation	HQS5	Dull	Absorbing
Hedonic – stimulation	HQS6	Harmless	Challenging
Hedonic – stimulation	HQS7	Novel	Conventional
Attractiveness	ATT1	Pleasant	Unpleasant
Attractiveness	ATT2	Ugly	Pretty
Attractiveness	ATT3	Appealing	Unappealing
Attractiveness	ATT4	Rejecting	Inviting
Attractiveness	ATT5	Good	Bad
Attractiveness	ATT6	Repulsive	Pleasing
Attractiveness	ATT7	Motivating	Discouraging

You then need to review the word choices and terminology used for each of the anchors and decide on the words that you think will be understood most easily and universally. For example, you might find "Pretty – Ugly" of the Schrepp et al. (2006) version a better set of anchors than "Aesthetic –

Unaesthetic" of the Hassenzahl version or you may wish to add "Interesting – Boring" to "Exciting – Dull" as suggested in Hassenzahl, Beu, and Burmester (2001).

Note also that the questions in AttrakDiff (or any questionnaire) represent strictly operational definitions of pragmatic and hedonic quality, and because you may have missed some aspects of these measures that are important to you, you can add your own questions to address issues you think are missing.

Alternatives to AttrakDiff. As an alternative to the AttrakDiff questionnaire, Hassenzahl, Beu, and Burmester (2001) have created simple questionnaire of their own for evaluating emotional impact, also based on semantic differential scales. Their scales have the following easy-to-apply anchors (from their Figure 1):

- outstanding vs. second rate
- exclusive vs. standard
- impressive vs. nondescript
- unique vs. ordinary
- innovative vs. conservative
- exciting vs. dull
- interesting vs. boring

Like AttrakDiff, each scale in this questionnaire has seven possible ratings, including these end points, and the words were originally in German.

Verbal emotion measurement instruments, such as questionnaires, can assess mixed emotions because questions and scales in a questionnaire or images in pictorial tools can be made up to represent sets of emotions (Desmet, 2003). PrEmo, developed by Desmet, uses seven animated pictorial representations of pleasant emotions and seven unpleasant ones. Desmet concludes that "PrEmo is a satisfactory, reliable emotion measurement instrument in terms of applying it across cultures."

There is a limitation, however. Verbal instruments tend to be language dependent and, sometimes, culture dependent. For example, the vocabulary for different dimensions of a questionnaire and their end points are difficult to translate precisely. Pictorial tools can be the exception, as the language of pictures is more universal. Pictograms of facial expressions can sometimes express emotions elicited more effectively than verbal expression, but the question of how to draw the various pictograms most effectively is still an unresolved research challenge.

An example of another emotional impact measuring instrument is the Self-Assessment Manikin (SAM) (Bradley & Lang, 1994). SAM contains nine symbols

indicating positive emotions and nine indicating negative emotions. Often used for Websites and print advertisements, the SAM is administered during or immediately after user interaction. One problem with application after usage is that emotions can be fleeting and perishable.

Observing physiological responses as indicators of emotional impact

In contrast to self-reporting techniques, UX practitioners can obtain emotional impact indicator data through monitoring of participant physiological responses to emotional impact encounters as usage occurs. Usage can be teeming with user behaviors, including facial expressions, such as ephemeral grimaces or smiles, and body language, such as tapping of fingers, fidgeting, or scratching one's head, that indicate emotional impact.

Physiological responses can be "captured" either by direct behavioral observation or by instrumented measurements. Behavioral observations include those of facial expressions, gestural behavior, and body posture.

The emotional "tells" of facial and bodily expressions can be fleeting and subliminal, easily missed in real-time observation. Therefore, to capture facial expressions data and other similar observational data reliably, practitioners usually need to make video recordings of participant usage behavior and do frame-by-frame analysis. Methods for interpreting facial expressions have been developed, including one called the Facial Action Coding System (Ekman & Friesen, 1975).

Kim et al. (2008) remind us that while we can measure physiological effects, it is difficult to connect the measurements with specific emotions and with causes within the interaction. Their solution is to supplement with traditional synchronized video-recording techniques to correlate measurements with usage occurrences and behavioral events. But this kind of video review has disadvantages: the reviewing process is usually tedious and time-consuming, you may need an analyst trained in identifying and interpreting these expressions often within a frame-by-frame analysis, and even a trained analyst cannot always make the right call.

Fortunately, software-assisted recognition of facial expressions and gestures in video images is beginning to be feasible for practical applications. Software tools are now becoming available to automate real-time recognition and interpretation of facial expressions. A system called "faceAPI"[5] from Seeing Machines is advertised to both track and understand faces. It comes as a software

[5] http://www.seeingmachines.com/product/faceapi/

module that you embed in your own product or application. An ordinary Webcam, focused on the user's face, feeds both faceAPI and any digital video-recording program with software-accessible time stamps and/or frame numbers.

Facial expressions do seem to be mostly culture independent, and you can capture expressions without interruption of the usage. However, there are limitations that generally preclude their use. The main limitation is that they are useful for only a limited set of basic emotions such as anger or happiness, but not mixed emotions. Dormann (2003) says it is, therefore, difficult to be precise about what kind of emotion is being observed.

In order to identify facial expressions, faceAPI must track the user's face during head movement that occurs in 3D with usage. The head-tracking feature outputs X, Y, Z position and head orientation coordinates for every video frame. The facial feature detection component of faceAPI tracks three points on each eyebrow and eight points around the lips.

The detection algorithm is "robust to occlusions, fast movements, large head rotations, lighting, facial deformation, skin color, beards, and glasses." This part of faceAPI outputs a real-time stream of facial feature data, time coordinated with the recorded video, that can be understood and interpreted via a suite of image-processing modules. The faceAPI system is a commercial product, but a free version is available to qualified users for non-commercial use.

Bio-metrics to detect physiological responses to emotional impact

The use of instrumented measurement of physiological responses in participants is called biometrics. Biometrics are about detection and measurement of autonomic or involuntary bodily changes triggered by nervous system responses to emotional impact within interaction events. Examples include changes in heart rate, respiration, perspiration, and eye pupil dilation. Changes in perspiration are measured by galvanic skin response measurements to detect changes in electrical conductivity.

Such nervous system changes can be correlated with emotional responses to interaction events. Pupillary dilation is an autonomous indication especially of interest, engagement, and excitement and is known to correlate with a number of emotional states (Tullis & Albert, 2008).

The downside of biometrics is the need for specialized monitoring equipment. If you can get some good measuring instruments and are trained to use them to get good measures, it does not get more "embodied" than this. But most equipment for measuring physiological changes is out of reach for the average UX practitioner.

It is possible to adapt a polygraph or lie detector, for example, to detect changes in pulse, respiration, and skin conductivity that could be correlated with emotional responses to interaction events. However, the operation of most of this equipment requires skills and experience in medical technology, and interpretation of raw data can require specialized training in psychology, all beyond our scope. Finally, the extent of culture independence of facial expressions and other physiological responses is not entirely known.

12.5.5 Data Collection Techniques to Evaluate Phenomenological Aspects of Interaction

Long-term studies required for phenomenological evaluation

Phenomenological aspects of interaction involve emotional impact, but emotional impact over time not emotional impact in snapshots of usage as you might be used to observing in other kinds of UX evaluation. The new perspective that the phenomenological view brings to user experience requires a new kind of evaluation (Thomas & Macredie, 2002).

Phenomenological usage is a longitudinal effect in which users invite the product into their lives, giving it a presence in daily activities. As an example of a product with presence on someone's life, we know someone who carries a digital voice recorder in his pocket everywhere he goes. He uses it to capture thoughts, notes, and reminders for just about everything. He keeps it at his bedside while sleeping and always has it in his car when driving. It is an essential in his lifestyle.

Thus, phenomenological usage is not about tasks but about human activities. Systems and products with phenomenological impact are understood through usage over time as users assimilate them into their lifestyles (Thomas & Macredie, 2002). Users build perceptions and judgment through exploration and learning as usage expands and emerges.

The timeline defining the user experience for this kind of usage starts even before first meeting the product, perhaps with the desire to own or use the product, researching the product and comparing similar products, visiting a store (physical or online), shopping for it, and beholding the packaging and product presentation. By the time long-term physiological studies are done, they really end up being case studies. The length of these studies does not necessarily mean large amounts of person-hours, but it can mean significant calendar time. Therefore, the technique will not fit with an agile method or any other approach based on a very short turnaround time.

It is clear that methods for studying and evaluating phenomenological aspects of interaction must be situated in the real activities of users to encounter a broad range of user experience occurring "in the wild." This means that you

cannot just schedule a session, bring in user participants, have them "perform," and take your data. Rather, this deeper importance of context usually means collecting data in the field rather than in the lab.

The best raw phenomenological data would come from constant attention to the user and usage, but it is seldom, if ever, possible to live with a participant 24/7 and be in all the places that a busy life takes a participant. Even if you could be with the participant all the time, you would find that most of the time you will observe just dead time when nothing interesting or useful is happening or when the participants are not even using the product. When events of interest do happen, they tend to be episodic in bursts, requiring special techniques to capture phenomenological data.

But, in fact, the only ones who can be there all the times and places where usage occurs are the participants. Therefore, most of the phenomenological data collection techniques are self-reporting techniques or at least have self-reporting components—the participants themselves report on their own activities, thoughts, problems, and kinds of usage. Self-reporting techniques are not as objective as direct observation, but they do offer practical solutions to the problem of accessing data that occur in your absence.

These longer term user experience studies are, in some ways, similar to contextual inquiry and even approach traditional ethnography in that they require "living with the users." The Petersen, Madsen, and Kjaer (2002) study of two families' usage of TV sets over 4 to 6 months is a good example of a phenomenological study in the context of HCI and UX.

The iPod is an example of a device that illustrates how usage can expand over time. At first it might be mostly a novelty to play with and to show friends. Then the user will add some applications, let us say the *iBird Explorer: An Interactive Field Guide to Birds of North America*.[6] Suddenly usage is extended out to the deck and perhaps eventually into the woods. Then the user wants to consolidate devices by exporting contact information (address book) from an old PDA.

Finally, of course, the user will start loading it up with all kinds of music and books on audio. This latter usage activity, which might come along after several months of product ownership, could become the most fun and the most enjoyable part of the whole usage experience.

Goals of phenomenological data collection techniques

Regardless of which technique is used for phenomenological data collection, the objective is to look for occurrences within long-term usage that are indicators of:

[6] http://www.ibird.com/

- ways people tend to use the product
- high points of joy in use, revealing what it is in the design that yields joy of use and opportunities to make it even better
- problems and difficulties people have in usage that interfere with a high-quality user experience
- usage people want but is not supported by the product
- how the basic mode of usage changes, evolves, or emerges over time
- how usage is adapted; new and unusual kinds of usage people come up with on their own

The idea is to be able to tell stories of usage and emotional impact over time.

Diaries in situated longitudinal studies

In one kind of self-reporting technique, each participant maintains a "diary," documenting problems, experiences, and phenomenological occurrences within long-term usage. There are many ways to facilitate this kind of data capture within self-reporting, including:

- paper and pencil notes
- online reporting, such as in a blog
- cellphone voice-mail messages
- pocket digital voice recorder

We believe that the use of voice-mail diaries for self-reporting on usage has importance that goes well beyond mobile phone studies. In another study (Petersen, Madsen, & Kjaer, 2002), phone reporting proved more successful than paper diaries because it could occur in the moment and had a much lower incremental effort for the participant. The key to this success is readiness at hand.

A mobile phone is, well, very mobile and can be kept ready to use at all times. Participants do not need to carry paper forms and a pen or pencil and can make the calls any time day or night and under conditions not conducive to writing reports by hand. Cellphones keep users in control during reporting; they can control the amount of time they devote to each report.

As Palen and Salzman (2002) learned, the mobile phone voice-mail method of data collection over time is also low in cost for analysts. Unlike paper reports, recorded voice reports are available immediately after their creation and

systematic transcription is fairly easy. They found that unstructured verbal data supplemented their other data very well and helped explain some of the observations or measurements they made.

Users often expressed subjective feelings, bolstering the phenomenological aspects of the study and relating phone usage to other aspects of their daily lives. These verbal reports, made at the crucial time following an incident, often mentioned issues that users forgot to bring up in later interviews, making voice-mail reports a rich source of issues to follow up on in subsequent in-person interviews.

If a mobile phone is not an option for self-reporting, a compact and portable handheld digital voice recorder is a viable alternative. If you can train the participants to carry it essentially at all times, a dedicated personal digital recorder is an effective and low-cost tool for self-reporting usage phenomena in a long-term study.

Evaluator triggered reporting for more representative data

Regardless of the reporting medium, there is still the question of when the self-reporting is to be done during long-term phenomenological evaluation. If you allow the participant to decide when to report, it could bias reporting toward times when it is convenient or times when things are going well with the product, or the participant might forget and you will lose opportunities to collect data.

To make the reporting a bit more randomly timed and according to your choice of frequency, thereby possibly being more likely to capture representative phenomenological activity, you can be proactive in requesting reports. Buchanan and Suri (2000) suggest that the participant be given a dedicated pager to carry at all times. You can then use the pager to signal randomly timed "events" to the participant "in the wild." As soon as possible after receiving the pager signal, the participant is to report on current or most recent product usage, including specific real-world usage context and any emotional impact being felt.

Periodic questionnaires over time

Periodically administered questionnaires are another self-reporting technique for collecting phenomenological data. Questionnaires can be used efficiently with a large number of participants and can yield both quantitative and qualitative data. This is a less costly method that can get answers to predefined questions, but it cannot be used easily to give you a window into usage in context to reveal growth and emergence of use over time. As a last resort, you can use a

series of questionnaires spaced over time and designed to elicit understanding of changes in usage over those time periods.

Direct observation and interviews in simulated real usage situations

The aforementioned techniques of self reporting, triggered reporting, and periodic questionnaires are ways of sampling long-term phenomenological usage activity. As another alternative, the analyst team can simulate real long-term usage within a series of direct observations and interviews. The idea is to meet with participant(s) periodically, each time setting up conditions to encourage episodes of phenomenological activity to occur during these observational periods. The primary techniques for data collection during these simulated real usage sessions are direct observation and interviews.

We described an example of using this technique in Chapter 15. Petersen, Madsen, and Kjaer (2002) conducted a longitudinal study of the use of a TV and video recorder by two families in their own homes. During the time of usage, analysts scheduled periodic interviews within which they posed numerous usage scenarios and had the participants do their best to enact the usage, while giving their feedback, especially about emotional impact. The idea is to set up conditions so you can capture the essence of real usage and reflect real usage in a tractable time-frame.

12.6 VARIATIONS IN FORMATIVE EVALUATION RESULTS

Before we conclude this chapter and move on to rapid and rigorous evaluation methods, we have to be sure that you do not entertain unrealistically high expectations for the reliability of formative evaluation results. The reality of formative evaluation is that, if you repeat an evaluation of a design, prototype, or system applying the same method but different evaluators, different participants, different tasks, or different conditions, you will get different results. Even if you use the same tasks, or the same evaluators, or the same participants, you will get different results. And you certainly get even more variation in results if you apply different methods. It is just not a repeatable process. When the variation is due to using different evaluators, it is called the "evaluator effect" (Hertzum & Jacobsen, 2003; Vermeeren, van Kesteren, & Bekker, 2003).

Reasons given by Hertzum and Jacobsen (2003) for the wide variation in results of "discount" and other inspection methods include:

- vague goals (varying evaluation focus)
- vague evaluation procedures (the methods do not pin down the procedures so each application is a variation and an adaptation)
- vague problem criteria (it is not clear how to decide when an issue represents a real problem)

The most important reason for this effect is due to the individual differences among people. Different people see usage and problems differently. Different people have different detection rates. They naturally see different UX problems in the same design. Also in most of the methods, issues found are not questioned for validity. This results in numerous false positives, and there is no approach for scrutinizing and weeding them out. Further, because of the vagueness of the methods, intra-evaluator variability can contribute as much as inter-evaluator variability. The same person can get different results in two successive evaluations of the same system.

As said earlier, much of this explanation of limited effectiveness applies equally well to lab-based testing, too. That is because many of the phenomena and principles are the same and the working concepts are not that different.

In our humble opinion, the biggest reason for the limitations of our current methods is that the problem—evaluating UX in large system designs—is very difficult. The challenge is enormous—picking away at a massive application such as MS Word or a huge Website with our Lilliputian UX tweezers. And this is true regardless of the UX method, including lab-based testing. Of course, for these massive and complex systems, everything else is also more difficult and more costly.

How can you ever hope to find your way through it all, let alone do a thorough job of UX evaluation? There are just so many issues and difficulties, so many places for UX problems to hide. It brings to mind the image of a person with a metal detector, searching over a large beach. There is no chance of finding all the detectable items, not even close, but often the person does find many things of value.

No one has the resources to look everywhere and test every possible feature on every possible screen or Web page in every possible task. You are just not going to find all the UX problems in all those places. One evaluator might find a problem in a place that other evaluators did not even look. Why are we surprised that each evaluator does not come up with the same comprehensive problem list? It would take a miracle.