CHAPTER

UX Evaluation Methods and Techniques

21

Highlights

- Types of UX evaluation data:
  - Quantitative versus qualitative data.
  - Objective versus subjective data.
- Formative versus summative evaluation.
- Informal versus formal summative evaluation versus engineering UX evaluation.
- Analytic versus empirical UX evaluation methods.
- Rigor versus rapidness in UX evaluation methods.
- Rapid UX evaluation methods.
- UX evaluation data collection techniques.
- Specialized UX evaluation methods and techniques.
- UX evaluation goals and constraints determine method choices.

21.1 INTRODUCTION

21.1.1 You Are Here

We begin each process chapter with a "you are here" picture of the chapter topic in the context of The Wheel, the overall UX design lifecycle template (Fig. 21-1). In this chapter, the first of the chapters about the Evaluate UX lifecycle activity, we begin by introducing UX evaluation methods and techniques.

This chapter is to introduce UX evaluation methods and techniques, associated terminology, distinctions among the methods and techniques, strengths of each, and how to choose the UX evaluation methods and techniques based on evaluation goals. The how-to details for applying each method and technique are coming later in Chapters 24 and 25 on UX evaluation data collection.

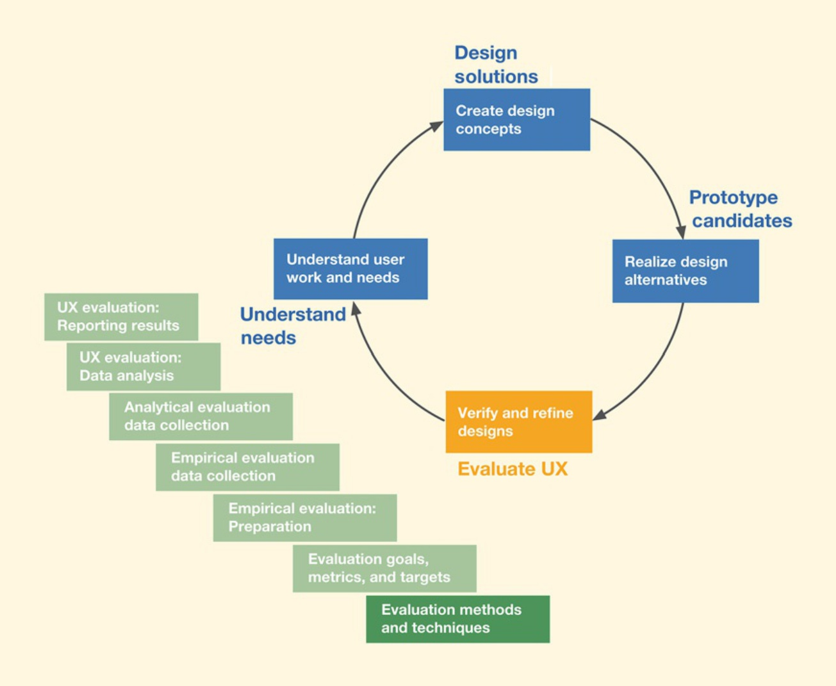
Design solutions
Create design concepts
Prototype candidates
Realize design alternatives
Verify and refine designs
Evaluate UX
Understand user work and needs
Understand needs
UX evaluation: Reporting results
UX evaluation: Data analysis
Analytical evaluation data collection
Empirical evaluation data collection
Empirical evaluation: Preparation
Evaluation goals, metrics, and targets
Evaluation methods and techniques

Fig. 21-1

You are here in the upfront chapter on UX evaluation methods and techniques for evaluation within the Evaluate UX lifecycle activity in the context of the overall Wheel lifecycle template.

Lab-based UX evaluation

An empirical UX evaluation method based on observing user participants performing tasks in a UX laboratory setting. Employs critical incident identification and think-aloud techniques for qualitative, and sometimes other quantitative, data collection (Section 21.2.4.1).

21.1.2 Methods versus Techniques

The concepts of methods and techniques were already established in Chapter 2. We review and interpret them here specifically in the context of UX evaluation.

There is not a clear-cut definition of the difference between a method and a technique in our practice, with the main difference being level. An evaluation method is a high-level overall way of doing UX evaluation and a technique is usually a lower-level way of doing specific steps within a method.

For example, lab-based empirical testing with users is an evaluation method. One of the techniques used to collect data about UX problems within that method is critical incident identification. Don't worry if these terms don't mean much to you now; we'll get to them soon.

In this chapter, we introduce some selected UX evaluation methods and techniques to get you familiar with the range of possibilities.

21.1.3 User Testing? No!

You know what "user testing" means, but it's not really an accurate term and no user will like the idea of being tested and, thereby, possibly made to look ridiculous. Users are participants who help us test or evaluate UX designs for usability or user experience, but *we are not testing the user.*

Traditionally, studies in psychology and human factors have referred to "subjects" as the people who perform the tasks while others observe and measure. In UX, we wish to invite these volunteers to join our team and help us evaluate designs. Because we want them to participate, we use the terms "participant" or "user participant" instead of "subject."

21.1.4 Types of UX Evaluation Data

UX evaluation data can be objective or subjective and it can be quantitative or qualitative. Practically, the two dimensions are orthogonal, so both objective and subjective data can be either qualitative or quantitative. For example, questionnaire results (Section 24.3.2) are usually both subjective and quantitative.

21.1.4.1 Quantitative versus qualitative data

Quantitative data are numeric data, usually from measurements, used to assess a level of achievement. The two most common kinds of quantitative data collected most often in formative evaluation are objective user performance data measured using benchmark tasks and subjective user opinion data measured using questionnaires. Quantitative data are the basis of the 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 22).

Qualitative data are nonnumeric descriptive data used to find and fix UX problems. Qualitative data from UX evaluation are usually descriptions of UX problems or issues observed or experienced during usage. Qualitative data, the key to identifying UX problems and their causes, are usually collected via critical incident identification, the think-aloud technique, and UX inspections methods.

21.1.4.2 Objective versus subjective data

Objective UX data are data observed directly. Objective data arise from observations by either the UX evaluator or the participant. Objective data are always associated with empirical methods.

Subjective UX data represent opinions, judgments, and other feedback. Subjective data originate from opinions of either UX evaluator or participant, concerning

Objective UX evaluation data

Qualitative or quantitative data acquired through direct empirical observation, usually of user performance (Section 21.1.4.2).

Formal summative evaluation

A formal, statistically rigorous summative (quantitative) empirical UX evaluation that produces statistically significant results (Section 21.1.5.1).

Critical incident

An event that occurs during user task performance or other user interaction that indicate possible UX problem(s). Critical incident identification, an empirical UX evaluation data collection technique based on the participant and/or evaluator detecting and analyzing critical incidents, is arguably the single most important qualitative data collection technique (Section 24.2.1).

the user experience and satisfaction with the design. Analytic UX evaluation methods (Chapter 25) yield only qualitative subjective data (UX problem identification based on an expert opinion of the UX inspector). Questionnaires (Section 24.3.2) yield data that are quantitative and subjective (data on numeric scales based on opinions of users).

Think-aloud technique

A qualitative empirical data collection technique in which participants verbally express thoughts about the interaction experience, including their motives, rationale, and perceptions of UX problems, especially to identify UX problems (Section 24.2.3).

21.1.5 Formative Evaluation versus Summative Evaluation

The distinction between formative evaluation and summative evaluation is based on a long-standing dichotomy:

- *Formative UX evaluation is diagnostic UX evaluation using qualitative data collection with the objective to form a design, that is, for finding and fixing UX problems and thereby refining the design.*
- *Summative UX evaluation is defined to be UX evaluation with the objective to sum up or assess the success of a UX design.*

Inspection (UX)

An analytical evaluation method in which a UX expert evaluates an interaction design by looking at it or trying it out, sometimes in the context of a set of abstracted design guidelines. Expert evaluators are both participant surrogates and observers, asking themselves questions about what would cause users problems and giving an expert opinion predicting UX problems (Section 25.4).

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, Singley, and Rosson (1992) were among the first to use the terms in an HCI context.

Formative evaluation is primarily diagnostic, with the aim of identifying and fixing UX problems and their causes in the design. Summative evaluation is primarily rating or scoring; it is about collecting quantitative data for assessing a level of quality due to a design.

Analytic UX evaluation

Evaluation methods based on examining inherent attributes of the design rather than seeing the design in use (Section 21.2.2 and Chapter 25).

21.1.5.1 Formal summative evaluation

Summative UX evaluation includes both formal and informal methods. *A formal summative (quantitative) UX evaluation method is an empirical method that produces statistically significant results.* The term "formal" is used because the process is statistically rigorous.

In science, there is no substitute for formal summative studies, inferential statistics, and statistically significant results to find the "truth" in answers to science and research questions. But most of the work we do in UX evaluation is more engineering than science, where getting at "truth" is a more practical and less exact business. In many ways, engineering is about judgment based on hunches and intuition that are, in turn, based on skill and experience.

Formal summative evaluation is based on an experimental design for controlled comparative hypothesis testing using an *m* by *n* factorial design with *y* independent variables, the results of which are subjected to statistical tests for significance. This takes special training and skills, so don't promise summative evaluation if you can't deliver. In addition, it's expensive and time consuming to do a proper summative evaluation. In sum, formal summative evaluation is an important HCI research skill, but in our view it is not part of UX practice.

As an example of a design change that is probably not (by itself) measurably better in terms of usability but is arguably better, consider a particular button label. If the whole team agrees that the old button label was vague and confusing and the new button label is clear and easily understood, then the team probably should make that design change.

Formal summative evaluation is outside the scope of the rest of this book.

For more discussion of why we don't consider formal summative evaluation as part of UX practice, see Section 28.2.

Participant

A participant, or user participant, is a user, potential, or user surrogate who helps evaluate UX designs for usability and user experience. These are the people who perform tasks and give feedback while we observe and measure. Because we wish to invite these volunteers to join our team and help us evaluate designs (i.e., we want them to participate), we use the term "participant" instead of "subject" (Section 21.1.3).

21.1.5.2 Informal summative evaluation

An *informal summative UX evaluation method is a quantitative summative UX evaluation method that is not statistically rigorous and does not produce statistically significant results.* Informal summative evaluation is used in support of formative evaluation, as an engineering technique to help assess how well you are achieving good usability and UX.

Informal summative evaluation is done without experimental controls, with smaller numbers of user participants, and with only summary descriptive statistics (such as average values). At the end of each iteration for a product version, the informal summative evaluation can be used as a kind of acceptance test to compare with our UX targets (Chapter 22) and help ensure that we meet our UX and business goals with the product design.

Table 21-1 highlights the differences between formal and informal summative UX evaluation methods.

21.1.5.3 Engineering UX evaluation: Formative plus informal summative

As an engineering method, UX evaluation can include formative evaluation plus an optional informal summative component (Fig. 21-2). The summative part can't be used to *prove* anything, but it is a valuable guide to the UX design process. Evaluation methods such as design reviews, heuristic methods, and other UX

Table 21-1

Some differences between formal and informal summative UX evaluation methods

Formal Summative UX Evaluation	Informal Summative UX Evaluation
Science	Engineering
Randomly chosen subjects/participants	Deliberately nonrandom participant selection to get most formative information
Concerned with having large enough sample size (number of subjects)	Deliberately uses relatively small number of participants
Uses rigorous and powerful statistical techniques	Deliberately simple, low-power statistical techniques (e.g., simple mean and, sometimes, standard deviation)
Results can be used to make claims about "truth" in a scientific sense	Results cannot be used to make claims, but are used to make engineering judgments
Relatively expensive and time consuming to perform	Relatively inexpensive and rapid to perform
Rigorous constraints on methods and procedures	Methods and procedures open to innovation and adaptation
Tends to yield "truth" about very specific scientific questions (A vs. B)	Can yield insight about broader range of questions regarding levels of UX achieved and the need for further improvement
Not used within a UX design process	Intended to be used within a UX design process in support of formative methods

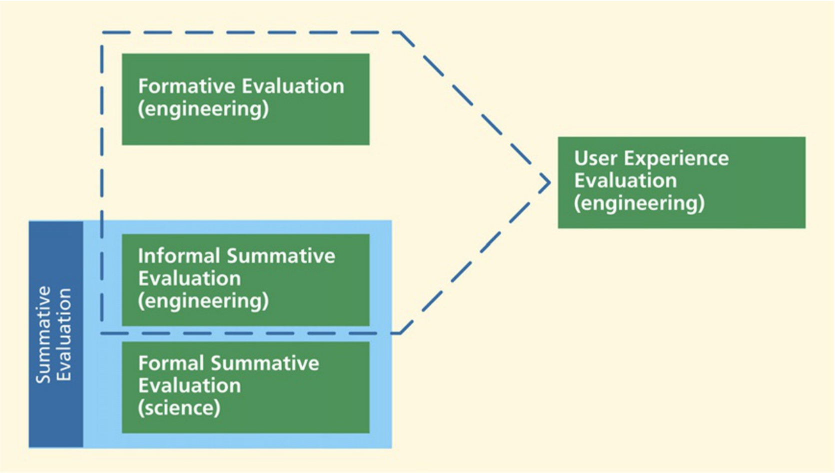
Formative Evaluation (engineering)
User Experience Evaluation (engineering)
Summative Evaluation
Informal Summative Evaluation (engineering)
Formal Summative Evaluation (science)

Fig. 21-2

Engineering UX evaluation is a combination of formative evaluation and informal summative evaluation.

inspection methods are good examples of purely formative evaluation methods (without a summative component).

Empirical methods such as testing with user participants can also be limited to a formative evaluation component, especially in early stages, when we are defining and refining the design and are not yet interested in performance indicators.

21.1.6 Our Goal-Oriented Approach

We will describe our approach in which UX evaluation goals determine the choices of methods and techniques needed to achieve the goals. But first we need to establish some terminology.

21.2 UX EVALUATION METHODS

For formative UX evaluation, you can use either empirical or analytic UX evaluation methods (Hartson, Andre, & Williges, 2003).

21.2.1 Empirical UX Evaluation Methods

Empirical methods, by definition, depend on data observed in the performance of real user participants and data coming directly from user participants. These data include critical incident data observed in empirical evaluation and comments from users while "thinking aloud" and/or in their questionnaire responses.

Empirical methods can be performed in a UX laboratory setting, at a conference table, or in the field. Empirical testing can produce both quantitative and qualitative data from the same measuring instrument, such as user task performance.

The UX lab is a more or less controlled environment, which is a plus in terms of limiting distractions, but testing in the real work context of a field setting can be more effective at ensuring realistic task conditions for ecological validity.

21.2.2 Analytic UX Evaluation Methods

Analytic methods are based on examining inherent attributes of the design rather than seeing the design in use. Except for numerical ratings and similar data, analytic methods yield qualitative subjective data. Although analytic UX evaluation methods (Chapter 25) can be applied rigorously and correspondingly more slowly, they were developed as faster and less expensive methods to produce approximations to or predictors of empirical results.

Empirical UX evaluation

A family of UX evaluation methods that depends on data observed in the performance of real user participants and data coming directly from user participants (Section 21.2.1).

Think-aloud technique

A qualitative empirical data collection technique in which participants verbally express thoughts about the interaction experience, including their motives, rationale, and perceptions of UX problems, especially to identify UX problems (Section 24.2.3).

Measuring instrument

The means of generating values for the particular UX measure, the vehicle through which values are measured for a UX character to be evaluated. Examples include benchmark tasks and questionnaires (Sections 22.6 and 22.7).

Analytic methods include design reviews, design walkthroughs, and inspection methods, such as heuristic evaluation (HE).

21.2.3 Comparison

Empirical methods are sometimes called "payoff methods" (Carroll et al., 1992; Scriven, 1967) because they are based on how a design or design change pays off in real observable usage. Analytic methods are sometimes called "intrinsic methods" because they are based on analyzing intrinsic characteristics of the design.

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

In describing the distinction between payoff and intrinsic approaches to evaluation, Scriven wrote an oft-quoted (Carroll et al., 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." He was speaking of intrinsic and payoff evaluation, respectively. In Hartson et al. (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 a *good* axeman (or axeperson). UX goals (and, therefore, evaluation 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 inexperienced user, safety might be a UX goal that transcends firewood production, calling for a safer design that might necessarily sacrifice some efficiency.

21.2.4 Some Specific Empirical UX Evaluation Methods

21.2.4.1 Lab-based evaluation

Lab-based UX evaluation is an empirical UX evaluation method based on observing user participants performing tasks in a UX laboratory setting. It employs critical incident identification and think-aloud techniques for qualitative, and sometimes quantitative, data collection.

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, that is, its context in the world or its environment (Sections 16.3, 22.6.4.4).

Subjective UX evaluation data

Data based on opinion or judgment, of evaluator or user (Section 21.1.4.2).

Lab-based empirical UX evaluation relies on observing user participants performing representative tasks. Qualitative data lead to identifying UX problems to fix and quantitative data, if collected, lead to assessment of how well users perform with a given design.

In Chapter 22, we discuss quantitative lab-based evaluation goals and metrics. Chapter 23 is about preparing for empirical design. Chapter 24 covers empirical evaluation data collection methods and techniques.

21.2.4.2 RITE

The approach called Rapid Iterative Test and Evaluation (RITE) (Medlock, Wixon, McGee, & Welsh, 2005; Medlock, Wixon, Terrano, Romero, & Fulton, 2002), a fast user-based UX evaluation approach designed to pick the low-hanging fruit at relatively low cost, is one of the best rapid empirical evaluation methods. The key to fast iteration with RITE is fixing the problems as soon as they are identified. This fast turnaround of qualitative evaluation results and problem fixing makes RITE one of the most agile empirical methods. Problem reporting occurs while the team is still there, so they are already informed and immersed in the process. More details about RITE are in Section 24.6.1.

21.2.4.3 Quasiempirical evaluation

Quasiempirical UX evaluation methods are hybrid approaches arising when UX professionals develop their own methods with shortcuts, explained in more detail in Section 24.6.2.

Quasiempirical UX evaluation

Hybrid approaches arising when UX professionals develop their own methods with shortcuts. They are empirical because they involve data collection using participants or participant surrogates. But they are "quasi" because they are informal and flexible with respect to process and protocol and the UX evaluator can play a significant analytic role (Section 24.6.2).

21.2.5 Weaknesses of UX Evaluation Methods

21.2.5.1 Measurability of user experience: A problem on the empirical quantitative side

Quantitative evaluation of an attribute such as usability or UX implies some kind of measurement. But can you measure usability or user experience? This may come as a surprise, but neither usability nor user experience is directly measurable. In fact, most interesting phenomena, such as teaching and learning, share the same 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 complete a task and counts of errors encountered by users within task performance.

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

21.2.5.2 Reliability of UX evaluation methods: A problem on the qualitative side

In simple terms, the reliability of a UX evaluation method means *repeatability*, and it's a problem with both empirical and analytic methods (Hartson et al., 2003). It means that, if you use the same formative evaluation method with several different user participants (for empirical methods) or several different UX inspectors (for analytic methods), you won't get the same list of UX problems each time. In fact, the differences can be fairly large. We have to live with imperfect reliability. For more about UX evaluation reliability, see Section 28.3.

The good news is that even UX evaluation methods and techniques with low reliability can still be very effective—i.e., the methods still find UX problems that need fixing and often they find the most important problems (Hartson et al., 2003). Low reliability is not always a serious drawback; much of each iteration of formative evaluation in UX practice is about learning as much about the design as you can at the lowest cost and then moving on.

Design review

A slightly more comprehensive UX evaluation technique than design walkthroughs, usually done with click-through wireframe prototypes to demonstrate workflow and navigation. Often the primary evaluation method for task-level UX designs in the fast iteration of the late funnel (Section 25.2.2).

So, while it would be nice to have perfect reliability, as a practical matter if the method is reasonably effective, the process still works. Each time you do formative evaluation, you will get some list of UX problems. If you apply more rigor in the evaluation method, your list will be more complete and accurate. If you fix all those problems and do formative evaluation again, you'll get another list (of some of the remaining UX problems). Eventually, you can find and fix most of the UX problems, especially the important ones. This kind of approximation to the ideal is what engineering is all about.

21.2.6 Some Specific Analytic UX Evaluation Methods

21.2.6.1 Early design reviews and design walkthroughs

Design walkthrough

Informal technique for getting initial reactions to design concepts, usually employing only scenarios, storyboards, screen sketches, and/or some wireframes. No real interaction capability, so UX designer has to do the "driving" (Section 25.2.1).

Early design reviews and design walkthroughs are demos of the design by the UX team to get early reactions and feedback from team members and other stakeholders, including users and people in the client organization. We classify these as analytic methods because they are based on descriptions of how the design works rather than real usage by users.

Earliest presentations might employ scenarios and storyboards for evaluating the ecological view or conceptual design and screen sketches for task-level evaluation—nothing interactive. These media will rapidly evolve into click-through wireframe prototypes. You, the UX team, have to do the "driving" to demonstrate interaction and navigation; it's too early for anyone else in a user role to engage in real interaction.

The leader walks the group through key workflow patterns that the design is intended to support. In the early funnel part of the lifecycle process, this will involve an overview, the flow model, and the conceptual design. In the late funnel part, this will be centered on mostly the interaction design of one feature at a time (focusing on a small set of tasks). As the team follows the scenarios, looking systematically at parts of the design and discussing the merits and potential problems, the leader tells stories about users and usage, user intentions and actions, and expected outcomes.

Wireframe prototype

A prototype composed of wireframes, which are line-drawing representations of UX designs, especially the interaction design of screens (Section 20.4).

21.2.6.2 Expert UX inspection

The expert UX inspection is a rapid analytic evaluation method. Expert UX inspectors use their professional experience and knowledge of UX design guidelines to spot UX problems during an in-depth inspection of the design. They also often simulate real usage by playing the part of users and carrying out key tasks in search of problems.

Although the UX inspector may be an expert in UX, he or she may not be an expert at all in the system in question or its associated work domain. In these cases, the UX inspector can leverage this unfamiliarity to find problems for novice users, or the UX inspector can team up with a subject matter expert.

Inspection (UX)

An analytical evaluation method in which a UX expert evaluates an interaction design by looking at it or trying it out, sometimes in the context of a set of abstracted design guidelines. Expert evaluators are both participant surrogates and observers, asking themselves questions about what would cause users problems and giving an expert opinion predicting UX problems (Section 25.4).

Subject matter expert (SME)

Someone with a deep understanding of a specific work domain and the range of work practices within that domain (Section 7.4.4.1).

21.2.6.3 Heuristic evaluation (HE)

The heuristic evaluation (HE) method (Nielsen, 1992; Nielsen & Molich, 1990) is the best known and most popular of the inspection methods. In the HE method, inspectors are guided by an empirically derived list of about 20 "heuristics" or rules that govern good UX design. The UX professionals on the team do an expert UX inspection, asking how well each of these rules is followed in the design. The HE method has the advantages of being inexpensive, intuitive, and easy to motivate practitioners to do, and is especially effective for use early in the UX process.

Heuristic evaluation

An analytic evaluation method based on expert UX inspection guided by a set of heuristics, general high-level UX design guidelines (Section 25.5).

21.3 RIGOR VERSUS RAPIDNESS IN UX EVALUATION METHODS AND TECHNIQUES

The relationship between method rigor and rapidness (Section 3.2.7) is multifaceted:

- There is a tradeoff between the rigor with which you apply any method and the rapidness that can be achieved.
- All methods can span a range of rigor (and, therefore, rapidness).
- High rigor is not always a goal.
- Some methods were invented to favor rapidness over rigor.

21.3.1 There Is a Tradeoff between Rapidness and Achievable Rigor

In general, applying an evaluation method (or any method) with more rigor can achieve more complete and more accurate results, but will take more time and be more costly. Similarly, by taking shortcuts you can usually increase speed and reduce the cost of almost any UX evaluation method, but at the price of reduced rigor. See Section 3.2.7 for more discussion of rapidness versus rigor.

21.3.2 All Methods Can Span a Range of Rigor and Speed

Each UX evaluation method has its own range of potential rigor. For example, you can perform a lab-based empirical method in a highly rigorous way that will maximize effectiveness and minimize the risk of errors by refraining from shortcuts and by retaining all the data.

When high rigor is not required, you can also perform lab-based empirical evaluation rapidly and with many shortcuts. By filtering and abstracting the evaluation data down to the most important points, you can gain efficiency (higher speed and lower cost).

Similarly, an analytic method can be performed rapidly and at a low level of rigor or it can be performed with a high level of rigor, paying careful attention to the sources, completeness, and purity of the data.

21.3.3 High Rigor Is not Always a Goal

In many design situations, such as early project stages where things are changing rapidly, rigor isn't a priority. It's more important to be agile to iterate and learn quickly.

21.3.4 Some Methods were Invented to Favor Rapidness Over Rigor

Not all methods cover the same range of potential rigor, so there are choices to be made to match your need for rigor. While certainly not perfect, empirical UX evaluation methods performed at a high level of rigor have long been considered the standard of comparison with respect to effectiveness of methods.

Some other UX evaluation methods, including analytic methods (Chapter 25), were invented specifically to be faster and more cost-effective substitutes for the fully rigorous empirical methods. Analytic methods are designed to be shortcut methods for approximating what really counts, UX problems that could be found empirically.

So, there are two ways you can view the rigor of a UX evaluation method:

- The rigor with which any given method is applied.
- The range of rigor inherent in the method itself.

Because design reviews, walkthroughs, and UX inspections can be performed rapidly, they are the frequent choice in late-funnel task-level evaluation. There are also some empirical methods specifically designed to be rapid, including RITE and quasiempirical methods.

RITE

Rapid Iterative Test and Evaluation (Medlock et al., 2005; Medlock et al., 2002), a fast user-based UX evaluation approach designed to pick low-hanging fruit at relatively low cost. Based on fast iteration and fixing the problems as soon as they are identified. One of the best rapid empirical evaluation methods (Section 21.2.4.2).

Quasiempirical UX evaluation

Hybrid approaches arising when UX professionals develop their own methods with shortcuts. They are empirical because they involve data collection using participants or participant surrogates. But they are "quasi" because they are informal and flexible with respect to process and protocol and the UX evaluator can play a significant analytic role (Section 24.6.2).

21.4 UX EVALUATION DATA COLLECTION TECHNIQUES

21.4.1 Quantitative Data Collection Techniques

21.4.1.1 Objective data: User performance measures

Some quantitative data collection techniques employ user performance measures taken during empirical UX testing with user participants. Users perform benchmark tasks and UX evaluators take objective measures, such as the time to complete a task.

21.4.1.2 Subjective data: User questionnaires

Other quantitative data collection techniques employ questionnaires or user surveys to gather subjective data about how users view the design (Section 24.3.2). Questionnaires can be used as an evaluation method on their own or to supplement your objective UX evaluation data with subjective data directly from the user. Questionnaires are simple to use, for both analyst and participant, and can be used with or without a lab. Questionnaires are good for evaluating specific targeted aspects of the user experience, including perceived usability, usefulness, and emotional impact.

21.4.1.3 Warning: Modifying a questionnaire can damage its validity

The validity of a questionnaire is a statistical characteristic more of concern to summative studies. 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.

However, if you want or need to modify an existing questionnaire to suit specific needs, don't worry that modifying an existing, already validated, questionnaire might affect its validity; questionnaire validity is rarely a practical concern in UX practice.

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.

Participant

A participant, or user participant, is a user, potential, or user surrogate who helps evaluate UX designs for usability and user experience. These are the people who perform tasks and give feedback while we observe and measure. Because we wish to invite these volunteers to join our team and help us evaluate designs (i.e., we want them to participate), we use the term "participant" instead of "subject" (Section 21.1.3).

Benchmark task

A task description devised for a participant to perform during UX evaluation so that UX measures such as time on task and error rates can be obtained and compared to a baseline value across the performances of multiple participants (Section 22.6).

Usefulness

A component of user experience based on utility, system functionality that gives users the ability to accomplish the goals of work (or play) through using the system or product (Section 1.4.3).

21.4.2 Qualitative Data Collection Techniques

Qualitative data collection techniques are used to capture data for UX problem identification. Critical incident identification, think aloud, and codiscovery, described next, are among the most popular qualitative data collection techniques.

21.4.2.1 Critical incident identification

Critical incident identification is a qualitative UX data collection technique that involves the UX team observing user participants performing tasks and detecting "critical incidents," or occurrences where users encounter UX problems. Problems thus identified are traced to their causes in the UX design and put on a list to be fixed in subsequent iterations (Section 26.4.9).

21.4.2.2 User think-aloud techniques

The think-aloud technique is usually applied together with critical incident identification as a second way of spotting UX problems during user task performance. In this technique, users are encouraged to express their thoughts verbally as they perform tasks and exercise the UX design, thus revealing otherwise possible hidden qualitative data about UX problems. The think-aloud technique is perhaps the most useful of all UX evaluation techniques as it gets at the user's state of mind precisely at the time of use.

21.4.2.3 Codiscovery

You can use two or more participants in a team approach to the think-aloud technique (O'Malley, Draper, & Riley, 1984), an approach that Kennedy (1989) called "codiscovery" (Section 24.2.3.3). Multiple participants sometimes offer the natural ease of talking in conversation with another person (Wildman, 1995), leading to data from multiple viewpoints.

21.5 SPECIALIZED UX EVALUATION METHODS

In addition to the "standard" UX evaluation methods and techniques of the previous sections, there are a number of specialized methods and techniques. We briefly describe a few here.

21.5.1 Alpha and Beta Testing and Field Surveys

Alpha and beta testing are useful postdeployment evaluation methods. After almost all development is complete, manufacturers of software applications sometimes send out alpha and beta (prerelease) versions of the application software to select users, experts, customers, and professional reviewers as a preview. In exchange for the early preview, users try it out and give feedback on the experience. Often little or no guidance is given for the review process beyond just survey questions such as "tell us what you think is good and bad and what needs fixing, what additional features you would like to see, etc."

An alpha version of a product is an earlier, less polished version, usually with a smaller and more trusted "audience." Beta is as close to the final product as they can make it and is sent out to a larger community. Most companies develop a beta trial mailing list of a community of early adopters and expert users, mostly known to be friendly to the company and its products and helpful in their comments.

Alpha and beta testing are easy and inexpensive ways to get high-level feedback, and they are based on real usage. But alpha and beta testing barely qualify as formative evaluation because:

- You don't get the kind of detailed UX problem data you get from a mainstream formative evaluation process.
- It is usually too late to change the design in any significant way if problems are identified.

Alpha and beta testing are very much individualized to a given development organization and environment. Full descriptions of how to do alpha and beta testing are beyond our scope.

Like alpha and beta testing, user field survey information is retrospective and, while it can be good for getting at user satisfaction, it does not capture the details of use within the usage experience.

Anything is better than nothing, but please don't let these after-the-fact methods be the only formative evaluation used within the product lifecycle in your given organization.

21.5.2 Remote UX Evaluation

Remote UX evaluation methods (Dray & Siegel, 2004; Hartson & Castillo, 1998) are good for evaluating systems after they have been deployed in the field. Methods include:

- Simulating lab-based UX testing using the Internet as a long extension cord to the user (e.g., UserVue by TechSmith).
- Online surveys for getting after-the-fact feedback.
- Software instrumentation of clickstream and usage event information.
- Software plug-ins to capture user self-reporting of UX issues.

The latter approach (Hartson & Castillo, 1998) uses self-reporting of UX problems by users as the problems occur during their normal usage, allowing you to get at the perishable details of the usage experience, especially in real-life daily work usage. As always, the best feedback for design improvement is feedback deriving from Carter's (2007) "inquiry within experience," or formative data given concurrent with usage rather than retrospective recollection.

21.5.3 Automatic UX Evaluation

Lab-based and UX inspection methods are labor intensive and, therefore, limited in scope (small number of users exercising small portions of large systems). But large and complex systems with large numbers of users offer the potential for a vast volume of usage data. Think of "observing" a hundred thousand users using Microsoft Word. Automatic methods have been devised to take advantage of this boundless pool of data, collecting and analyzing usage data without need for UX specialists to deal with each individual action. Sometimes multiple versions of the product are released to different sets of users and resulting data are compared to ascertain which version is better. This type of evaluation is often called A-B testing, where A and B are two variations of the design.

The result is a massive amount of data about keystrokes, clickstreams, and pause/idle times. But all data are at the low level of user actions, without any information about tasks, user intentions, cognitive processes, etc. There are no direct indications of when the user is having a UX problem somewhere in the midst of that torrent of user action data. Basing redesign on click counts and low-level user navigation within a large software application could well lead to low-level optimization of a system with a bad high-level design. A full description of how to do automatic usability evaluation is beyond our scope.

21.6 ADAPTING AND APPROPRIATING UX EVALUATION METHODS AND TECHNIQUES

It's not enough to choose the "standard" UX evaluation methods and techniques that fit your goals and constraints. You will find it very natural to tailor and tune the methods and techniques to the nuances of your team and project until you have appropriated them as your own for each different design situation. This aspect of choosing methods and techniques is discussed at length in Section 2.5.

For UX evaluation, as perhaps for most UX work, our motto echoes that old military imperative: improvise, adapt, and overcome! Be flexible and customize your methods and techniques, creating variations to fit your evaluation goals and needs. This includes adapting any method by leaving out steps, adding new steps, and changing the details of a step.