CHAPTER

24

Empirical UX Evaluation: Data Collection Methods and Techniques

Highlights

- Empirical ways of generating and collecting data within the needs pyramid.
- Empirical methods and techniques for generating and collecting UX evaluation data.
- Empirical methods and techniques for generating and collecting qualitative UX data:
  - Critical incident identification.
  - User think-aloud technique.
- Empirical methods and techniques for generating and collecting quantitative UX data.
- Methods and techniques for collecting data about emotional impact and meaningfulness.
- Procedures for empirical data collection:
  - Preliminaries with participants.
  - Session protocol.
  - Data collection.
- Rapid empirical methods for generating and collecting qualitative UX evaluation data.

24.1 INTRODUCTION

24.1.1 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. 24-1). In this chapter, we elaborate on how to collect data for empirical UX evaluation.

24.1.2 Empirical Ways of Generating and Collecting Data Within the Needs Pyramid

As a reminder, the UX needs pyramid has these layers (bottom to top):

- Ecological.

- Interaction.
- Emotional.

Scope (of Delivery)

Describes how the target system or product is "chunked" (broken into what size pieces) in each iteration or sprint for delivery to the client and users for feedback and to the software engineering team for agile implementation (Section 3.3).

Almost all the empirical data collection methods and techniques in this chapter can be used to conduct UX evaluation within any layer of the needs pyramid, at any scope, and any level of rigor.

Foremost, essentially everything in the UX evaluation chapters (Chapters 21–27) applies to the interaction level. Many topics in this chapter are generally relevant to evaluating the emotional level, too, and those instances

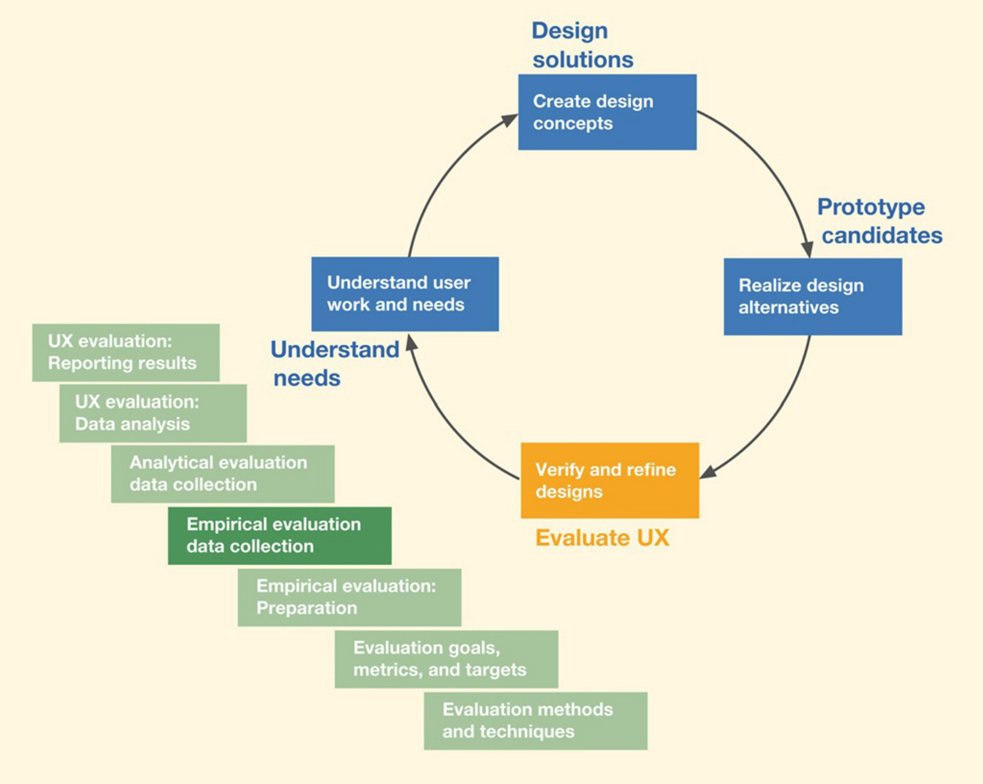
Design solutions
Create design concepts
Prototype candidates
Realize design alternatives
Understand user work and needs
Understand needs
Verify and refine designs
Evaluate UX
UX evaluation: Reporting results
UX evaluation: Data analysis
Analytical evaluation data collection
Empirical evaluation data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

Fig. 24-1

You are here in the chapter on data collection within the Evaluate UX lifecycle activity in the context of the overall Wheel lifecycle process.

are pointed out as such. In addition, there is an entire section (Section 24.4) devoted to specific ways to evaluate the emotional level. That leaves the ecological level, which is the subject of the next section.

As usual, the early funnel is best for evaluating in the ecological layer, where you have a large scope to embrace the product or system ecology and its conceptual design. The late funnel is best for evaluating within the interaction layer, where you address user actions in task-level design.

Ecology

In the setting of UX design, the ecology is the entire set of surrounding parts of the world, including networks, other users, devices, and information structures, with which a user, product, or system interacts (Section 16.2.1).

24.1.2.1 Empirical methods and techniques for generating and collecting UX evaluation data in the ecological layer

Ecological concepts are sometimes difficult to evaluate. Because ecology is about how the broader system works, and the various devices and contexts it encompasses, the issues with its design tend to be high-level in nature and generally abstract. Therefore, evaluating the conceptual design of an ecology requires measuring instruments that bring into focus the themes and structures of how the overarching system works.

There are two main goals for evaluating the design of an ecology:

- Ascertain whether users *understand* how the ecology is structured:
  - Evaluate whether the designers' mental model is communicated clearly through the conceptual design.
  - Evaluate whether users are able to form a clear mental model as a result.
- Ensure the users are able to get work done in the new ecology. Evaluate whether the conceptual design of the ecology is *appropriate* for their work context.

Evaluate user understanding. An effective way to evaluate the first goal is to first introduce the system's ecology to users and allow them to get familiar by performing benchmark tasks on individual designs. This usually involves prototypes of the UX designs on all devices that constitute the ecology. After they get familiar with the various devices and their capabilities, address the first goal by asking them to articulate how the overall system works. A specific task asking them to explain how the system works to a new colleague who joined their team is a good way to get at their understanding.

Specifically, you can ask questions such as "How would you describe this system to someone who has never seen it before? What is the underlying "model" for this system? Is that model appropriate? Where does it deviate? Does it meet your expectations? Why and how? These questions get to the root of determining the user's mental model for the system.

Another way to evaluate understanding is by following benchmark tasks on a device with questions on expectations of what the consequences will be on another device. This probes their mental model on how the two devices work together. For example, asking what they expect to see on the phone or watch when they save a record on the desktop gets to their understanding on what data they expect to be available on what platforms.

Evaluate appropriateness to work context. The second goal of evaluating appropriateness of the design to a given work practice is a more task-centric issue—about how well users can get work done in the ecology. For this, you need benchmark tasks that require the user to switch among various devices in the ecology. This second goal also requires tasks pertaining to participating in the ecology (e.g., sign up or create an account). Tasks that start on one device, are interrupted, and then resume on another device help evaluate user expectations on what context is maintained across devices. Questionnaires are another evaluation instrument that probes the users about capabilities they expect on each of the devices in the ecology.

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).

24.2 EMPIRICAL METHODS AND TECHNIQUES FOR GENERATING AND COLLECTING QUALITATIVE UX DATA

Qualitative data collection is by far the most important kind of evaluation data across all UX practice. The defining formative UX evaluation goal is to identify UX problems and their causes so the design can be improved. In empirical testing with participants, this goal is achieved through qualitative UX data collected primarily through observation and recording of critical incidents (next section) and use of the think-aloud technique.

Qualitative UX Evaluation Data

Nonnumeric descriptive data taken, for example, while observing user task performance, used to find and fix UX problems (Section 21.1.4.1).

24.2.1 Critical Incident Identification

Along with the think-aloud technique, critical incident identification is arguably the real workhorse of empirical data collection techniques. Critical incident identification is an empirical data collection technique based on the participant or evaluator detecting and analyzing occurrences of events that indicate a (usually) bad user experience.

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).

24.2.1.1 What is a critical incident?

Despite many 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 occurring within usage that reveals a barrier,* *problem, or difficulty encountered by the user, or simply something the user did not like* (Castillo and Hartson, 2000; del Galdo, Williges, Williges, and Wixon, 1986).

For more on the history and background of the critical incident identification technique, see Section 28.7.

24.2.1.2 Mostly used as a variation

In today's practice, there is no single critical incident identification technique, but each UX evaluator adapts a variation most suitable to the needs at hand. For more about how the application of this technique varies, see Section 28.7.2.

24.2.1.3 Who identifies critical incidents?

For simplicity, in our interpretation of the critical incident data collection technique, the UX evaluator is the one who identifies and records critical incident data. For more about who identifies critical incidents, see Section 28.7.3.

24.2.2 Critical Incident Data Capture

The way that you collect and document critical incident data as you go will have much to say about the accuracy and efficiency of your subsequent data analysis. Write concise but detailed critical incident and UX problem descriptions as clearly, precisely, and completely as you can in real time. Terse notes will be more difficult to interpret later.

The best kind of critical incident data are:

- Detailed.
- Observed during usage.
- Captured immediately.
- Associated closely with specific task performance.

Critical incident data are perishable. The biggest reason why empirical UX evaluation is effective is that it allows the capture of detailed usage-related data as it occurs. These detailed data, perishable if not captured immediately and precisely as they arise during usage, are essential for isolating specific problems with the user UX design.

Sometimes critical incident data are subtle. Experienced UX evaluators will know how to see critical incidents in subtle user behavior—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 request for clarification might help determine if any of these subtle observations should be considered a symptom of a UX problem.

For a discussion about the timing of critical incident data capture and the evaluator's awareness zone, see Section 28.7.4.

24.2.2.1 What's in critical incident data?

Critical incident data about a UX problem should contain as much detail as possible, including usage research 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.

24.2.2.2 Avoid video recording

In past years, some UX labs used video recording routinely to capture all user and screen actions and facilitator and participant comments and, thereby, captured the raw data necessary to identify critical incidents. However, most videotaping setups were cumbersome, complicated, expensive, and often unreliable. And reviewing the video was time consuming and expensive. Today's UX practice calls for more lightweight and nimble techniques for data collection.

24.2.2.3 Manual note taking for critical incident data collection

Instead, manual note taking is the most basic critical incident capture technique and is the most useful and efficient approach. Evaluators take comprehensive, real-time raw critical incident notes with a laptop or with pencil and paper. When thoughts come faster than they can write, they might make audio notes to themselves on a handheld digital voice recorder—anything to capture raw data while it is still fresh.

24.2.2.4 Follow up on hunches

If you get an intuitive feeling during UX evaluation that something is wrong with the design that is not coming out explicitly in the data, you should not let it go but you should follow up on it.

24.2.3 The Think-Aloud Data Collection Technique

Also called "verbal protocol" in the early human factors literature, the think-aloud technique is a qualitative data collection technique in which user participants, as the name implies, verbally express 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 UX design, their expectations, strategies, biases, likes, and dislikes. Variations of this simple technique are rooted in psychological and human factors experimentation well before it was used in usability engineering (Lewis, 1982).

24.2.3.1 Why use the think-aloud technique?

The think-aloud technique is simple to use for both analyst and participant. It is most useful during empirical evaluation of user task performance, but it is also useful 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.

Observational data are important during an evaluation session with a participant attempting to perform a task. But often quite a bit of the 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? The goal of the think-aloud technique is to tap into this data hidden in the participant's mind.

24.2.3.2 How to manage the participant in the think-aloud technique

Although there are some points to watch for, in its simplest form this technique could not be easier. It simply entails having participants think out loud and share their thoughts verbally while they perform tasks or otherwise interact with a product or system you want to evaluate. Here's what to do with the participants involved:

- At the beginning, explain the concept of thinking aloud.
- Explain that this means you will expect them to talk while they work and think, sharing their thoughts by verbalizing them to you.

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).

Emotional Impact

An affective component of user experience that influences user feelings. Includes such effects as enjoyment, pleasure, fun, satisfaction, aesthetics, coolness, engagement, and novelty and can involve deeper emotional factors such as self-expression, self-identity, a feeling of contribution to the world, and pride of ownership (Section 1.4.4).

- Ask participants to tell you what they are thinking and not describing what they are doing.
- You might start with a little exercise or practice session to get warmed up and to get participants acclimated to thinking aloud.
- Among the thoughts you should encourage participants to express are descriptions of their intentions, what they are doing or are trying to do, and their motivations, the reasons why they are doing any particular actions.
- Encourage them to get past the chatty stage and get down to real engagement and introspection.
- You especially want them to speak out when they get confused, frustrated, or blocked.
- Actively elicit participant thoughts if they are not forthcoming.

Depending on the individual, thinking aloud usually comes quite naturally; it does not take much practice. Occasionally you might have to encourage or remind the participant to keep up the flow of thinking aloud.

24.2.3.3 Codiscovery think-aloud techniques

You may wish to try using two or more participants in a team approach, a technique that originated with O'Malley, Draper, and Riley (1984) and was named "codiscovery" by Kennedy (1989).

While it can seem unnatural and inhibiting to 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.

Hackman and Biers (1992) found that using 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.

24.2.3.4 Does thinking aloud affect quantitative task performance metrics in empirical 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.

24.3 EMPIRICAL METHODS AND TECHNIQUES FOR GENERATING AND COLLECTING QUANTITATIVE UX DATA

As we have said, quantitative measures are not used much anymore in UX evaluation.

24.3.1 Objective Quantitative Data for User Performance Measurement

If you do need quantitative data, though, the most popular quantitative data collection techniques involve measurement of user performance of benchmark tasks, usually at the same time you collect qualitative data.

For example, an evaluator may measure the time it takes the participant to perform a task, count the number of errors a participant makes while performing a task, count the number of tasks a participant can perform within a given time period, and so on, depending on the measures established in your UX targets (Chapter 22).

24.3.1.1 Timing task performance

By far the simplest way to measure time on task is by manually using a stopwatch. It is really the only sensible way for low-fidelity prototypes, such as click-through wireframe prototypes.

For the rare times when precise timing measurements are required, it is possible to embed software timers to instrument the software internally.

24.3.1.2 Counting user errors

The simplest way to count user errors during task performance is to use a manual event counter such as a handheld “clicker” for counting people coming through a gate for an event. Manual counters are perfect for low-fidelity, especially paper, prototypes.

The key to counting errors correctly is in knowing what constitutes an error. Not everything that goes wrong, not even everything a user does wrong, during task performance should be counted as a user error. So what are we looking for? A user error is usually considered to have occurred when the participant takes *any action that does not lead to progress in performing the desired task within the boundaries of the intended design (and not, for example, gaps in a prototype’s functionality).*

24.3.1.3 What generally does not count as a user error?

Typically, we don't count accessing online help or other documentation as an error. As a practical matter, we also want to exclude any random act of curiosity or exploration that might be interjected by the user (e.g., "I know this is not right, but I am curious what will happen if I click this"). Also a different successful path "invented" by the user is not really an error, but probably should be noted as an important observation.

And we don't usually include "oops" errors, what Norman (1990, p. 105) calls "slips." These are errors that users make by accident when, in fact, they know better. For example, the user knows the right button to click but clicks the wrong one, perhaps through a slip of the hand, a brain burp, or being too hasty. Finally, we don't usually include typing errors, unless their cause could somehow be traced to a problem in the design or unless the application is about typing (Section 22.6.2.5).

Subjective UX Evaluation Data

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

24.3.2 Subjective Quantitative Data Collection: Questionnaires

A questionnaire is a fast and easy way to collect subjective UX data, either as a supplement to any other rapid UX evaluation method or as a method on its own.

Questionnaires with good track records, such as the Questionnaire for User Interface Satisfaction (QUIS), the System Usability Scale (SUS), or Usefulness, Satisfaction, and Ease of Use (USE), are all easy and inexpensive to use and can yield varying degrees of UX data. Perhaps the AttrakDiff questionnaire might be the best choice for a rapid standalone method, as it is designed to address both pragmatic (usability and usefulness) and emotional impact issues. All these questionnaires are discussed further at the end of this section.

Emotional Impact

An affective component of user experience that influences user feelings. Includes such effects as enjoyment, pleasure, fun, satisfaction, aesthetics, coolness, engagement, and novelty and can involve deeper emotional factors such as self-expression, self-identity, a feeling of contribution to the world, and pride of ownership (Section 1.4.4).

24.3.2.1 Questionnaires as supplements to lab-based sessions

Postsession questionnaires can be used to supplement what you have discovered objectively in the session. Most questionnaire responses are written, but you might also consider asking survey questions orally to gather postsession information. The direct verbal exchange allows you to pursue issues of interest with impromptu follow-up questions.

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).

24.3.2.2 Questionnaires as an evaluation method on their own

A questionnaire can also be used as the primary UX data collection instrument when used as an evaluation method on its own. A questionnaire can contain probing questions about the total user experience. Although 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 create so that data are valid and reliable.

24.3.2.3 Semantic differential scales

A semantic differential scale, or Likert scale (1932), is a range of semantic 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. It is helpful to also include verbal (or pictorial) labels associated with each numeric value.

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 might have these labels: strongly agree, agree, neutral, disagree, and strongly disagree, with the associated numeric values, respectively, of +2, +1, 0, −1, and −2.

24.3.2.4 The Questionnaire for User Interface Satisfaction (QUIS)

The QUIS, developed at the University of Maryland (Chin, Diehl, and Norman, 1988), is one of the earliest available questionnaires for evaluating user satisfaction. 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 and Stetson, 2004), but there have been many extensions and variations. Although developed originally for screen-based designs, QUIS is resilient and can be extended easily, for example, by replacing the term "system" with "website" and "screen" with "webpage."

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.

Last we heard, 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 24-1, we show a sample excerpted and adapted with permission from the QUIS with fairly general applicability, at least to desktop applications. The columns represent the UX attribute being evaluated and the semantic anchors.

Table 24-1

An excerpt adapted from QUIS, with permission

1. Terminology relates to task domain	Distantly—closely
2. Instructions describing tasks	Confusing—clear
3. Instructions are consistent	Never—always
4. Operations relate to tasks	Distantly—closely
5. Informative feedback	Never—always
6. Display layouts simplify tasks	Never—always
7. Sequence of displays	Confusing—clear
8. Error messages are helpful	Never—always
9. Error correction	Confusing—clear
10. Learning the operation	Difficult—easy

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

Table 24-1 An excerpt adapted from QUIS, with permission —cont'd

11. Human memory limitations	Overwhelmed—are respected
12. Exploration of features	Discouraged—encouraged
13. Overall reactions	Terrible—wonderful
Overall reactions	Frustrating—satisfying
Overall reactions	Uninteresting—interesting
Overall reactions	Dull—stimulating
Overall reactions	Difficult—easy

24.3.2.5 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 1–5. 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 objective UX data 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

Objective UX Evaluation Data

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

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. According to Usability Net (2006), it was the most highly recommended of all the publicly available questionnaires.

24.3.2.6 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.

According to Lund, questions were chosen for inclusion in USE through a process of factor analysis and partial correlation.

USE has been applied successfully to systems, products, and websites. 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.

Here is an abbreviated version of the USE questionnaire questions:

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.

24.3.2.7 Other questionnaires

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

General-purpose usability questionnaires:

- Computer System Usability Questionnaire (CSUQ), developed by James Lewis (Lewis, 1995, 2002) at IBM, is well regarded and available in the public domain.
- Software Usability Measurement Inventory (SUMI)[2] is "a rigorously tested and proven method of measuring software quality from the end user's point of view." 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 desktop applications to large domain-complex applications.
- After Scenario Questionnaire (ASQ), developed by IBM, is available in the public domain (Bangor, Kortum, and Miller, 2008, p. 575).
- Post-Study System Usability Questionnaire (PSSUQ), developed by IBM, is available in the public domain (Bangor et al., 2008, p. 575).

[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.

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 website" Human Factor Research Group (2010).

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 (1996).

Evaluation questionnaires that address emotional impact:

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

24.3.2.8 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.

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 don't and how to improve designs. Therefore, we recommend you consider adding to each question a free-form space, labeled as "If notable, please describe why you gave that rating."

24.3.2.9 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 (0–10 and NA) to 6 (−2, −1, 0, 1, 2, and NA) for each question, reducing the number 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.

24.3.2.10 Modifying the System Usability Scale

In the course of their study of SUS, Bangor et al. (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, or best imaginable."

Not caring for the term "user-friendliness," we would add the recommendation to change that phrase to something else such as "usability" or "UX quality." In studies by Bangor et al. (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.

If you are concerned about questionnaire validity, see Section 21.4.1.3.

24.3.3 Methods and Techniques for Generating and Collecting Emotional Impact and Meaningfulness Data

In this section we describe a series of techniques for collecting data about emotional impact and meaningfulness, given in order of increasing specialization. You may never need to use some of the more painstaking of these techniques in your practice of UX evaluation, but they are here for completeness.

24.3.4 The Most Important Technique: Direct Observation

Before looking to the upcoming more specialized techniques for emotional impact and meaningfulness, your best practical approach is to pick up on indications of emotional impact and meaningfulness during your regular use of critical incident identification and think-aloud techniques.

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

You will identify emotional impact through its indicators: "verbal and nonverbal languages, facial expressions, behaviors, and so on" (Shih and Liu, 2007, citing Dormann, 2003). For a "Usability Test Observation Form," a comprehensive list of verbal and nonverbal behaviors to be noted during observation, see Tullis and Albert (2008, p. 170).

One difficulty in any observation or measurement of physiological reactions to usage events is that often there is no way to connect the physiological response to a particular emotion and its cause within the interaction.

For more on direct observation of physiological responses as indicators of emotional impact, see Section 28.8.1.

24.3.5 Verbal Self-Reporting Techniques for Collecting Emotional Impact Data

Beyond observing indicators of emotional impact and meaningfulness in your usual qualitative data collection, the more commonly used and less complex techniques involve indicators that are self-reported via verbal techniques such as the think-aloud technique or questionnaires.

24.3.5.1 Using the think-aloud technique to evaluate emotional impact

We have already talked about using the think-aloud technique for capturing the participant's view of interaction, critical incidents, and UX problems. The think-aloud technique can also be excellent as a window into the mind of the user with respect to emotional feelings as they occur. Because user think-aloud is a kind of self-reporting technique, we add some detail here.

Depending on the nature of the interaction, emotional impact indicators may be infrequent in the flow of task performance user actions, and you may see them mainly as a by-product of your hunt for other UX problem indicators. So, when you do encounter an emotional impact indicator during observation in task performance, you certainly should make a note of it. You can also make emotional impact factors the primary focus during the think-aloud technique:

- When you explain the concept of thinking aloud, be sure participants understand that you want to include emotional feelings resulting from interaction and usage.

- Explain that this means you will expect them to share their emotions and feelings while they work and think by talking about them to you.
- As you did when you used the think-aloud technique to capture qualitative UX data, you may wish to begin with a little exercise to be sure participants are on the same page about the technique.
- As before, you will mainly capture think-aloud data by written or typed notes.
- Also, as before, you may have to remind participants occasionally to keep the thinking aloud flowing.

During the flow of interaction:

- You can direct participants to focus their thinking aloud on comments about joy of use, aesthetics, fun, and so on.
- You should observe and note the more obvious manifestations of emotional impact, such as expressions like "I love this" and "this is really cool" and "wow" expressions, annoyances, or irritation.
- You should be sensitive to detecting when emotional impact goes flat, when there is no real joy of use; ask participants about it in terms of causes and how it can be improved.

Finally, a caution about cultural dependency. Most emotions themselves are pretty much the same across cultures, and nonverbal expressions of emotion, such as facial expressions and gestures, are fairly universal. But cultural and social factors can govern an individual's willingness to communicate about emotions. Different cultures may also have different vocabularies and different perspectives on the meaning of emotions and the appropriateness of sharing and revealing them to others.

24.3.5.2 Questionnaires as a self-reporting technique for collecting emotional impact data

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.

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 as well as being inexpensive, applicable from the 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, and Sutherland, 2006).

Subjective UX Evaluation Data

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

24.3.5.3 The AttrakDiff questionnaire as a verbal self-reporting technique for collecting emotional impact data

AttrakDiff (now AttrakDiff2), developed by Hassenzahl, Burmester, and Koller (2003), is an example of a questionnaire based on Likert (semantic differential) scales especially developed for getting at user perceptions of emotional impact.

Reasons for considering the AttrakDiff questionnaire include:

- AttrakDiff is freely available.
- AttrakDiff is short and easy to administer, and the verbal scale is easy to understand (Hassenzahl, Beu, and Burmester, 2001; Hassenzahl, Platz, Burmester, and Lehner, 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.

Example: AttrakDiff Questionnaire and a Variation of Same

With permission, here we show the full AttrakDiff questionnaire as taken from Hassenzahl, Schöbel, and Trautman (2008, Table 1), with semantic anchors for each scale item:

- 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.

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). Schrepp, Held, and Laugwitz (2006) developed a variation, reordered to group related items together.

24.3.5.4 Scoring ATTRAKDIFF questionnaires

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

For example, because there are 22 questions, as shown by the sample AttrakDiff questionnaire items above, the total summed-up score will be in the range of 22–154 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.

24.3.5.5 Alternatives to AttrakDiff

Hassenzahl, Beu, and Burmester

As an alternative to the AttrakDiff questionnaire, Hassenzahl et al. (2001) have created a 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 Fig. 1):

- Outstanding versus second rate.
- Exclusive versus standard.
- Impressive versus nondescript.
- Unique versus ordinary.
- Innovative versus conservative.
- Exciting versus dull.
- Interesting versus boring.

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

PrEmo

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 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 is 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.

Self-Assessment Manikin

An example of another emotional impact measuring instrument is the Self-Assessment Manikin (SAM) (Bradley and 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.

24.3.6 Direct Detection of Physiological Responses as Indicators of Emotional Impact

Beyond direct observation or self-reporting, there is a genre of techniques, called biometrics, for detecting emotional impact by directly measuring physiological responses in participants. These nonverbal techniques, which usually entail deploying probes and instrumentation, are beyond what almost any UX design project needs or wants and are almost exclusively in the realm of UX research. For a discussion of physiological measurements, see Section 28.8.2.

24.3.7 Generating and Collecting Meaningfulness Evaluation Data

Meaningfulness, or long-term emotional impact, occurs when users invite the product into their lives, giving it a presence in daily activities.

As an example of a product with a presence in 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 part of his lifestyle. He feels lost without it.

If it is a goal to understand the meaningfulness of a product, its adoption into human lifestyles, and how it impacts user lives over long periods of time, you should plan ways to study these phenomena situated in the real activities of users over time from the earliest thinking about the product to adoption into their lifestyles. Most choices for data collection techniques will include some kind of self-reporting by users, because you will not be able to be with your participants all the time.

As you collect data, you will be looking for indicators of all the different ways your users involve the product in their lives; the high points of joy in use; how the basic mode of usage changes, evolves, or emerges over time; and especially how usage is adapted to emerge as new and unusual kinds of usage. As we said earlier, you want to be able to tell stories of usage and emotional impact over time.

24.3.7.1 Long-term studies to evaluate meaningfulness

As users experience a product over time, they build perceptions and judgment through exploration and learning as usage expands and emerges (Thomas and Macredie, 2002). Thus, meaningfulness is not about tasks but about human activities. So, naturally, meaningfulness must be studied longitudinally and not just in the snapshots of usage that you might be used to observing in other kinds of UX evaluation.

The timeline defining meaningfulness within the user experience 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 meaningfulness 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 short turnaround time.

It is clear that methods for studying and evaluating the meaningfulness aspects of usage must be situated in the real activities of users to encounter a broad range of user experiences 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 iPad 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*.[4] Suddenly usage is extended out to the deck and perhaps eventually into the woods.

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 most enjoyable part of the whole usage experience.

24.3.7.2 Goals of meaningfulness data collection techniques

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

- 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 their original impressions and expectations of the product evolve over time, and why.
- How usage is adapted; new and unusual kinds of usage people come up with on their own.
- How important the product has become in the life of the user.

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

24.3.7.3 Direct observation and interviews in simulated real usage situations

Before we get into the techniques of self reporting, triggered reporting, and periodic questionnaires as ways of sampling meaningfulness within usage activity, let's look at a more direct approach. 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 showing emotional impact to occur during these

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

observational periods. The primary techniques for data collection during these simulated real usage sessions are direct observation and interviews. You will need to create conditions to encourage episodes of long-term usage activity to occur during these observational periods. 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.

As an example of using this technique, 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, periodic interviews were scheduled in the analysts' office, except in cases where users had difficulty in getting there and, then, the interviews were conducted in the users' homes. Within these interviews, the evaluators 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 that you can capture the essence of real usage and reflect real usage in a tractable time frame.

Here are some tips for success with this approach:

- Establish the interview schedule to take into account learning through usage by implementing a sequence of sessions longitudinally over time.
- As in usage research, it is necessary to observe user activities in addition to asking about them; as we know, the way people talk about what they do is often not the same as what they actually do.
- If you must make video recordings, be cautious and discreet in more private settings (such as the participant's home) usually found in this kind of usage context.

24.3.7.4 The importance of self-reporting

The best raw meaningfulness data would come from constant attention to the user and usage, but it is not 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, requiring special techniques to capture meaningfulness data.

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

24.3.7.5 Periodic questionnaires to sample meaningfulness

Another way you could choose to sample data about meaningfulness is by periodic questionnaires over time. You can use a series of such questionnaires to elicit understanding of major changes in usage over those time periods.

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 easily used to give you a window into the more revealing details of usage in context to reveal growth and emergence of use over time.

24.3.7.6 Diary-based self-reporting by users

We encourage you to improvise a self-reporting technique yourself, but you should definitely consider a diary-based technique in which each participant maintains a "diary," documenting problems, experiences, and occurrences of meaningfulness within usage. Diaries can be kept via paper and pencil notes, online reports, cell-phone messages, or voice recorders. Diaries are an effective and efficient technique for meaningfulness data collection, but analysis of the data can take time and effort.

For a diary-based technique to be effective, participants must be primed in advance:

- Give your users a list of the kinds of things to report, including problems, experiences, and occurrences of meaningfulness within long-term usage.
- Give them some practice exercises in identifying relevant situations and reporting on them.
- Get them to internalize the need to post a report whenever they confront a usage problem, use a new feature, or encounter anything interesting or fun within 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 voicemail messages.
- Pocket digital voice recorder.

24.3.7.7 Voicemail to capture user reports

Because of its flexibility and convenience and the ability of a user to make a phone call at almost any time, the use of voicemail is a meaningfulness data collection technique to consider for self-reporting on usage.

In one study (Petersen et al., 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 can be kept ready to use at all times. Participants don't 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.

To encourage participants to use voicemail for reporting, consider paying them a per-call monetary compensation (in addition to whatever payment you give them for participating in the study). In the Palen and Salzman (2002) study, they found that a per-call payment encouraged participants to make calls. There is a possibility, however, that this incentive might bias participants into making some unnecessary calls, but that did not seem to happen in this study.

As Palen and Salzman (2002) learned, the mobile phone voicemail 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 from voicemails supplemented their other data very well and helped explain some of the observations or measurements they made.

These verbal reports, made at the crucial time following an incident, often mentioned issues that users forgot to bring up in later interviews, making voicemail 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.

24.3.7.8 Evaluator-triggered reporting to control timing

Regardless of the reporting medium, there is still the question of when the self-reporting is to be done during meaningfulness 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.

You might choose to trigger reporting to control the timing of self-reporting to make it a bit more randomly timed and according to your choice of frequency. This approach to timing could result in a more random sampling of the occurrence of meaningfulness within long-term usage. Buchenau 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.

For other methods to evaluate emotional impact and meaningfulness, see Section 28.8.

24.4 PROCEDURES FOR EMPIRICAL DATA COLLECTION SESSIONS

24.4.1 Preliminaries with Participants

24.4.1.1 Introduce yourself and the lab: Be sure participants know what to expect

If you have a separate reception room in your UX facility, this is where you meet your participants before getting down to business with evaluation. Greet and welcome each participant and thank him or her for helping. Bring them in and show them around.

Informed Consent

A formal and signed permission given to UX professionals by usage research and evaluation participants to use the data gathered within the UX lifecycle activities, usually with certain stipulated limits (Section 23.7.3).

Introduce them to the setup and show them the lab. If you have one-way glass, explain it and how it will be used and show them the other side—what happens "behind the curtain." Openly declare any video recording you will do, which should have been explained in the consent form, too. Make participants feel that they are partners in this endeavor.

Tell your participants all about the design being evaluated and about the process in which they are participating. For example, you might say "We have early screen designs for our product in the form of a low-fidelity prototype of a new system for …" Tell them how they can help and what you want them to do.

Do your best to relieve anxiety and satisfy curiosity. Be sure that your participants have all their questions about the process answered before you proceed into evaluation. Make it very clear that they are helping you evaluate and you are not evaluating them in any way.

24.4.1.2 Paperwork

While still in the reception room or as soon as the user has entered the participant room:

- Have each participant read the general instructions (Section 23.7.4.1) and explain anything verbally, as needed.
- Have the participant read the institutional review board consent form (Section 23.7.3.2) and explain the consent form verbally as well.
- Have the participant sign the consent form (two copies); it must be signed "without duress." You keep one signed copy and give the participant the other signed copy; your copy must be retained for at least three years (the period may vary by organization).
- Have the participant sign a nondisclosure form (Section 23.7.4.2), if needed.
- Have the participant fill out any demographic survey you have prepared (to ensure they meet the requirements of your intended work activity role and corresponding user class characteristics).

24.4.2 Session Protocol and Your Relationship with Participants

Session protocol is about the mechanical details of session setup, your relationship with participants, and how you handle them throughout each session.

24.4.2.1 Your attitude toward UX problems

Before you actually do evaluation, it is easy to agree that this UX testing is a positive thing and we are all working together to improve the design. However, once you start hearing about problems participants are having with the design, it can trigger unhelpful reactions in your ego, instincts, and pride and you might feel inclined to be defensive about the design. Resist that temptation and proceed in your testing with a positive attitude; it will pay off.

Codiscovery

A qualitative data collection technique employing two or more participants interacting in a team approach to evaluation, usually with a think-aloud data collection technique. Two people can verbalize more naturally, yielding multiple viewpoints expressed within conversational interplay (Sections 21.4.2.3 and 24.2.3.3).

24.4.2.2 Cultivating a partnership with participants

Take the time to build rapport with your participants. More important to the success of your UX evaluation sessions than facilities and equipment is the relationship you establish with participants as partners in helping you evaluate and improve the product design. Once in the participant room, the facilitator should take a little time to "socialize" with the participant. If you have taken the participant on a "tour" of your facilities, that will have been a good start.

If you are using codiscovery techniques, allow some time for codiscovery partners to get to know each other and do a little bonding, perhaps while you are setting things up. Starting the session as total strangers can make them feel awkward and can interfere with their performance.

24.4.3 Prepare Yourself for Evaluating with Low-Fidelity Prototypes

If your wireframe prototype deck is printed on paper, lay it out, and have it ready to go (see example in Fig. 24-2). If your wireframe prototype deck is on a laptop, have it ready to operate (by you or by the user participant). We'll describe the process in terms of a paper wireframe prototype deck but it works the same way on a laptop.

Before each participant enters, the "executor" should "boot up" the prototype by putting the initial "screen" (wireframe) on the table. Arrange everything necessary for running the prototype, including the deck of subsequent wireframes.

Have the whole evaluation team ready to assume their roles and be ready to carry them out in the session.

- Evaluation *facilitator*, to keep the session moving, to interact with participants, and to possibly take notes on critical incidents (pick a person who has leadership abilities and "people" skills).
- Mark each page of the wireframe deck with an identifier of its role in the expected interaction sequence; it also helps to mark each object within a wireframe with the identifier of the destination wireframe should the user make an action on that object.

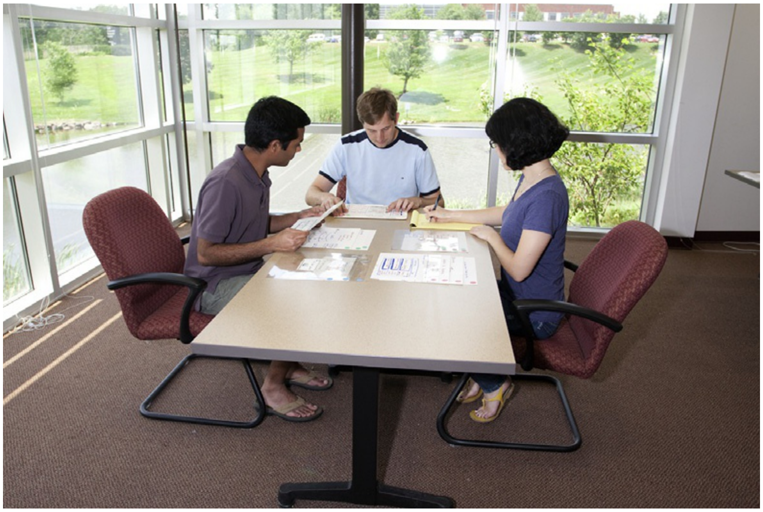

Fig. 24-2

Typical setup with participant at the end of a table for evaluation with the prototype deck printed on paper.

- Prototype *executor*, to move through the wireframes in response to user actions (pick a person who knows the design well).
- User performance *timer*, to time participants performing tasks and/or count errors (to collect quantitative data)—the timer person may want to practice with a stopwatch a bit before getting into a real session.
- Critical incident note takers (for spotting and recording critical incidents and UX problems).

Review your own in-session protocol. Some of the "rules" we suggest include:

- Team members *must not coach* participants as they perform tasks.
- The executor *must not anticipate user actions* and especially must not give the correct computer response for a wrong user action. The person playing computer must respond only to what the user actually does!
- The person playing computer may not speak, make gestures, etc.
- You may not change the design on the fly, unless that is a declared part of your process.

24.4.4 The Data Collection Session

24.4.4.1 The session begins

If you are using benchmark tasks, after the preliminaries and when you both are ready to start, have the participant read the first benchmark or other task description and ask if there are any questions. If you are taking timing data, don't include the benchmark task reading time as part of the task.

Once the evaluation session is under way, interesting things can happen quickly. Data you need to collect may start arriving in a flood. It can be overwhelming, but, by being prepared, you can make it easy and fun, especially if you know what kinds of data to collect.

24.4.4.2 Interacting with participants during the session

Now is the time to administer the benchmark tasks and do critical incidents identification, use the user think-aloud technique, and apply what you have learned here about the data collection techniques. Gather up your lists of critical incidents, UX problems, and user performance measures, as appropriate.

The facilitator is responsible for ensuring that the session runs smoothly and efficiently. It is generally the job of the facilitator to listen and not talk. But at key junctures you might elicit important data, if it does not interfere with task timing or if you are focusing on qualitative data. You can ask brief questions, such as "What are you trying to do?" "What did you expect to happen when you clicked on the such-and-such icon?" "What made you think that approach would work?"

If you are focusing on qualitative data, the evaluator may also ask leading questions, such as "How would you like to perform that task?" "What would make that icon easier to recognize?" If you are using the "think-aloud" technique for qualitative data gathering, encourage the participant by prompting occasionally: "Remember to tell us what you are thinking as you go."

If participants show signs of stress or fatigue, give them a break. Let them leave the participant room, walk around, and/or have some refreshments. Don't be too uptight about the session schedule.

24.4.4.3 To help the participant or not to help the participant?

When using participants for the critical incident technique and/or the think-aloud technique, you may be asked by the participant for a hint to help decide what to do next. Sometimes when participants are not making progress, they can benefit from a hint to get them back on track, but direct help almost always works against the goals of the session. You want to see whether the *participant* can determine how to perform the task, so don't tell them how to do it. It's often best to lead them to answer their own questions.

For example, don't answer questions such as "Is this right?" or "What if I click on this?" directly, but by asking your own questions, directing them to think it through for themselves, asking back what they think will happen.

24.4.4.4 Keeping your participant at ease

Remind yourself and your whole team that you should never, never laugh at anything during a UX evaluation session. You may be in a control room and think you have a soundproof setup but laughter has a way of piercing the glass. Because participants cannot see people behind the glass, it is easy for participants to assume that someone is laughing at them.

If participants become visibly flustered, frustrated, "zoned out," or blame themselves continually for problems in task performance, they may be suffering from stress and you should intervene. Take a short break and reassure and calm them. If participants become so discouraged that they want to quit the entire session, there is little you can or should do but thank them, pay them, and let them go.

24.4.5 Wrapping Up an Evaluation Session

24.4.5.1 Postsession probing via interviews and questionnaires

Immediately after each session, ask probing questions to clear up any confusion you have about critical incidents or UX problems. Conduct postsession interviews and administrator questionnaires to capture user thoughts and feelings while they are fresh.

Facilitators often start with some kind of standard *structured interview*, asking a series of preplanned questions aimed at probing the participant's thoughts about the product and the user experience. A typical postsession interview might include, for example, the following general questions. "What did you like best about the interface?" "What did you like least?" "How would you change so-and-so?" An interesting question to ask is "What are the three most important pieces of information that you must know to make the best use of this interface?"

For example, in one design, some of the results of a database query were presented graphically to the user as a data plot, the data points of which were displayed as small circles. Because most users did not at first realize that they could get more information about a particular data point if they clicked on the corresponding circle, one very important piece of information users needed to know about the design was that they should treat a circle as an icon and that they could manipulate it accordingly. Find out if your users got this.

24.4.5.2 Reset for the next participant

After running an evaluation session with one participant, you should organize the wireframe prototype to be ready for the next participant.

For web-based evaluation, clear out the browser history and browser cache, delete temporary files, remove any saved passwords, and so on. For a software prototype, save and back up any data you want to keep. Then reset the prototype state and remove any artifacts introduced in the previous session.

Finally, give the participant(s) their pay, gifts, and/or premiums, thank them, and send them on their way.

24.5 RAPID EMPIRICAL METHODS FOR GENERATING AND COLLECTING QUALITATIVE UX EVALUATION DATA

Empirical methods can be time-consuming and expensive. So some methods have evolved that are still empirical but are specifically designed with shortcuts to speed them up.

24.5.1 The Rapid Iterative Testing and Evaluation (RITE) UX Evaluation Method

24.5.1.1 Introduction

The approach called RITE (Medlock, Wixon, Terrano, Romero, and Fulton, 2002; Medlock, Wixon, McGee, and Welsh, 2005), a fast user-based testing approach, is representative of the category of rapid empirical UX evaluation methods and is one of the best.

RITE employs a fast collaborative (team members and participants) test-and-fix cycle designed to pick the low-hanging fruit at relatively low cost. The whole team is involved in arriving at the results.

The defining feature of RITE is fast turnaround—fixing key UX problems as soon as they are identified. Immediately after the product is evaluated, the whole project team, including the participants, analyzes the problems and decides on which changes to make. Changes are then implemented straightaway. If warranted, another immediate iteration of testing and fixing might ensue.

Because changes are included in all testing that occurs after that point, further testing can determine the effectiveness of the changes—whether the problem is, in fact, fixed and whether the fix introduces any new problems. Fixing a problem immediately also gives access to any aspects of the product that could not be tested earlier because they were blocked by that problem.

In his inimitable Wixonian wisdom, our friend Dennis reminds us that, "In practice, the goal is to produce, in the quickest time, a successful product that meets specifications with the fewest resources while minimizing risk" (Wixon, 2003).

24.5.1.2 How to do it: The RITE UX evaluation method

This description of the RITE UX evaluation method is based mainly on Medlock et al. (2002, 2005).

The project team starts by selecting a UX practitioner, whom we call the facilitator, to direct the testing session. The UX facilitator and the team prepare by:

- Identifying the characteristics needed in participants.
- Deciding on which tasks they will have the participants perform.
- Agreeing on critical tasks, the set of tasks that every user must be able to perform.
- Constructing a test script based on those tasks.
- Deciding how to collect qualitative user behavior data.
- Recruiting participants (Section 23.6) and scheduling them to come into the lab.

The UX facilitator and the team conduct the evaluation session with one to three participants, one at a time:

- Gathering the entire project team and any other relevant project stakeholders, either in the observation room of a UX lab or around a table in a conference room.
- Bringing in the participant playing the role of user.
- Introducing everyone and setting the stage, explaining the process and expected outcomes.

- Making sure that everyone knows the participant is helping evaluate the system and the team is not in any way evaluating the participant.
- Having the participant perform a small number of selected tasks while all project stakeholders observe silently.
- Having the participants think aloud as they work.
- Working together with the participants to find UX problems and ways the design should be improved.
- Taking thorough notes on problem indicators, such as task blocking and user errors.
- Focusing session notes on finding usability problems and noting their severity.

The UX facilitator and other UX practitioners:

- Identify from session notes the main UX problems observed and their causes in the design.
- Give everyone on the team the list of UX problems and causes.

The UX practitioner and the team, including the participants, address problems:

- Identifying problems with obvious causes and obvious solutions, such as those involving wording or labeling, to be fixed first.
- Determining which other problems can also reasonably be fixed.
- Determining which problems need more discussion.
- Determining which problems require more data (from more participants) to be sure they are real problems.
- Sorting out which problems they cannot afford to fix right now.
- Deciding on feasible solutions for the problems to be addressed.
- Implementing fixes for problems with obvious causes and obvious solutions.
- Starting to implement other fixes and bringing them into the current prototype as soon as feasible.

The UX practitioner and the team immediately conduct follow-up evaluation by:

- Bringing in new participants.
- Having them perform the tasks associated with the fixed problems, using the modified design.
- Working with the participants to see if the fixes worked and to be sure the fixes did not introduce any new UX problems.

The entire process just described is repeated until you run out of resources or the team decides it is done (all major problems found and addressed).

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).

24.5.1.3 Variations in RITE data collection

The flexibility of RITE allows consideration of alternative data collection techniques. For example, instead of testing with user participants, the team could employ a UX inspection method, heuristic evaluation, or otherwise for data collection while retaining the fast analysis and fixing parts of the cycle.

24.5.2 Quasiempirical UX Evaluation

Quasiempirical UX evaluation methods are 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.

24.5.2.1 Introduction to quasiempirical UX evaluation

Quasiempirical UX evaluation methods:

- Are still empirical, using volunteer participants (or an evaluator emulating a user).
- Are defined by the freedom given to the practitioner to innovate, to make it up as they go while being flexible about goals and approaches.
- Are very informal with respect to protocol—evaluators are encouraged to interrupt and intervene at opportune moments to elicit more thinking aloud and to ask for explanations and specifics.
- Do not involve any quantitative data.
- Can take place anywhere—UX lab, conference room, office, cafeteria, in the field.
- Are often punctuated with impromptu changes of pace, changes of direction, and changes of focus.
- Are characterized by jumping on issues as they arise and milking them to get the most information about problems, their effects on users, and potential solutions.
- Are not based on predefined "benchmark tasks," but a session can be task-driven, drawing on usage scenarios, essential use cases, step-by-step task interaction models, or other task data or task models.
- Can be driven by exploration of features, screens, widgets, or whatever suits.

24.5.2.2 Preparing for a quasiempirical evaluation session

Begin by ensuring that you have a set of representative, frequently used, and mission-critical tasks for your participants to explore. Have some exploratory questions ready.

Assign your UX evaluation team roles effectively, including participant, facilitator, and data collectors. If useful, try two evaluators for codiscovery. Further prepare for your quasiempirical session the same way you would for a full empirical session, only less formally and less thoroughly, to match the more rapid and more opportunistic nature of the quasiempirical approach.

24.5.2.3 Conduct a quasiempirical session, collecting data

As you, the facilitator, sit with each participant:

- Cultivate a partnership with the participant; you get the best results from working closely in collaboration.
- Make extensive use of the think-aloud data collection technique; encourage the participant by prompting occasionally: "Remember to tell us what you are thinking as you go."
- Encourage the participant to explore the system for a few minutes and get familiarized with it.
- Use some of the tasks that you have at hand, from the preparation step given earlier, more or less as props to support the action and the conversation; you are not interested in user performance times or other quantitative data.
- Work together with the participant to find UX problems and ways the design should be improved; take thorough notes—they are the sole raw data from the process.
- Let the user choose some tasks to do.
- Be ready to follow threads that arise rather than just following prescripted activities.
- Listen as much as you can to the participant; most of the time it is your job to listen, not talk.
- It is also your job to lead the session, which means saying the right thing at the right time to keep it on track and to switch tracks when useful.

At any time during the session, you can interact with the participant with questions such as:

- Ask participants to describe initial reactions as they interact with this system.
- Ask what parts of the design are not clear and why.
- Inquire about how the system compares with others they have used in the past.
- Ask if they have any suggestions for changing the designs.
- To place them in the context of their own work, ask them how they would use this system in their daily work; in other words, ask them to walk you through some tasks they would perform using this system in a typical workday.

Exercise 24-1: Empirical UX Evaluation Data Collection for Your System

Goal: To get a little practice in empirical data collection for a very simple formative UX evaluation using a wireframe prototype deck.

Activities: This is perhaps the most fun and most rewarding of all the exercises when you finally get to see some users in action with your UX design.
New team formation: This is described in terms of multiple teams in a classroom setting. For other setups, make appropriate adjustments.

- After all the teams are gathered and sitting around a table, make the switch of participants with another team. You send the two people in the participant role from your team to another team. Curb the potential confusion here by doing the swap in an orderly circular fashion among the teams.
- You will now have new participants from a different team who are unfamiliar with your design. These new participants are now permanently on your team for the rest of these exercises, including data collection, analysis, and reporting.
- As an alternative, if you do not have multiple teams, try recruiting a couple of coworkers or friends as participants.
- Sitting together in your newly formed teams, get out your UX target table form, your benchmark task descriptions, and your questionnaires.
- Dismiss your two participants (the new team members you just got) to the hallway or other waiting area.

Data collection:

- "Boot up" your prototype.
- Call in your first participant into the "lab," greet the participant, and explain the evaluation session.
- Have this first participant perform your first benchmark task for your UX targets. Have the participant read the first benchmark task aloud.
- Ask the participant to perform that task while thinking aloud.
- The executor moves prototype parts in response to participant actions.
- The facilitator directs the session and keeps it moving.
- Timer(s) writes down or enters timing and error count data as indicated in UX targets as the user performs the task (do not count participant's reading aloud of task in task timing).
- Everyone else available should be used to take notes on critical incidents and UX problems.
- Remember the rules about not coaching or anticipating user actions. And the computer may not speak!
- Have the participant read the second task aloud and ask any questions that might exist.
- Perform it.
- Be ready to collect data according to your UX targets.

- Have this first participant perform your second benchmark task while thinking aloud.
- How much data to collect? You need to collect a dozen or more critical incidents in this overall exercise (i.e., from both participants doing both benchmark tasks). If you do not get at least a half dozen from each participant, continue with that participant doing exploratory use of your prototype until you get enough critical incidents. For example, have them browse through each screen, looking at each object (button, menu, etc.), commenting on and giving their opinion about the quality of the user experience relating to various features.
- Have this participant complete your questionnaire and then give them their "reward."
- Keep your first participant as a new member of the rest of the team to help with observations.
- Bring in the second participant and perform the same session again.

Deliverables: All your data.
Schedule: Complete by end of class (about an hour and a half, if you are efficient).