CHAPTER

Empirical UX Evaluation: UX Goals, Metrics, and Targets

22

Highlights

- Concepts of UX goals, metrics, and targets.
- UX target tables:
  - Work role and user classes.
  - UX goals.
  - UX measures.
  - Measuring instruments.
  - UX metrics.
  - Level setting.
- Practical tips and cautions.
- Rapid goals, metrics, and targets.

22.1 INTRODUCTION

22.1.1 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. 22-1). In this chapter, we establish operational targets for user experience to assess the level of success in your designs so that you know when you can move on to the next iteration.

UX goals, metrics, and targets help build scaffolding to support planning for evaluation that will successfully reveal problems with user performance and emotional satisfaction. If used, UX goals, metrics, and targets are set up early as part of preparation for evaluation, and can serve to guide much of the process from analysis through evaluation.

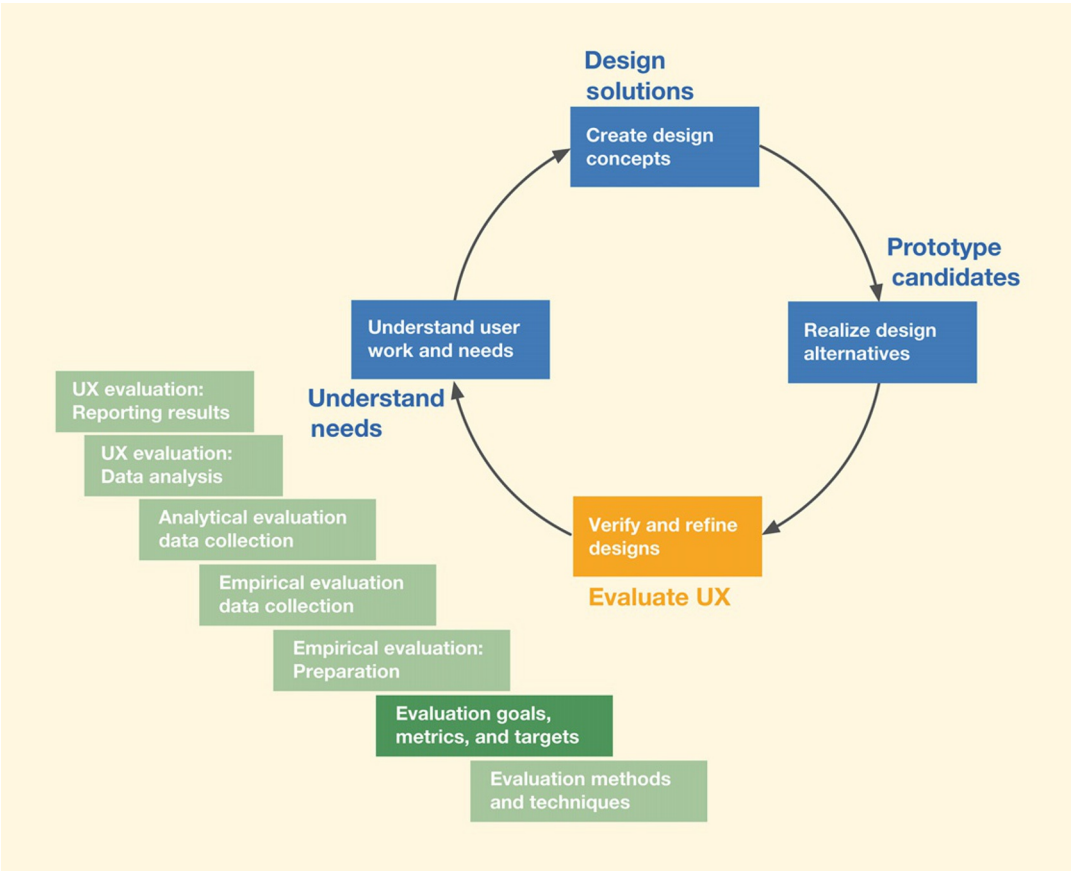

Fig. 22-1

You are here in the chapter on UX evaluation goals, metrics, and targets within the Evaluate UX lifecycle activity in the context of the overall Wheel lifecycle process.

22.1.2 Project Context for UX Metrics and Targets

In the early stages, evaluation usually focuses on qualitative data for finding UX problems. In these early evaluations, the absence of quantitative data precludes the use of UX metrics and targets. But you can start to establish them at any point if you intend to use them in later evaluations.

However, you might wish to forego UX metrics and targets altogether. In most practical contexts, specifying UX metrics and targets and following up

with the correspondingly rigorous evaluation may be too expensive. This level of completeness is only possible in a few organizations where there are significant established UX resources. In many projects, one round of evaluation is all you get. Also, as designers, we can know which parts of the design need further investigation just by looking at the results of the first round of evaluation. In such cases, quantitative UX metrics and targets may not be useful but benchmark tasks are still useful as vehicles for driving evaluation.

Regardless, the trend in the UX field is moving away from a focus on quantitative user performance measures and more toward rapid qualitative evaluation of usability, user satisfaction, and enjoyment. Nonetheless, we include the full treatment of UX goals, metrics, and targets here and quantitative data collection and analysis in the later UX evaluation chapters for completeness. This is because some readers and practitioners still want or need coverage of the topic.

In any case, we find that specifying UX goals, metrics, and targets is often overlooked, either because of lack of knowledge or because of lack of time. Sometimes this can be unfortunate because it can diminish the potential of what can be accomplished with the resources you will be putting into user experience evaluation. This chapter can help you avoid that pitfall.

Fortunately, creating UX metrics and targets, after a little practice, does not take much time. You will then have specific quantified UX goals against which to test rather than just waiting to see what happens when you put users in front of your UX design. Because UX metrics and targets provide feasible objectives for formative evaluation efforts, the results can help you pinpoint where to focus on redesign most profitably.

And, finally, UX goals, metrics, and targets offer a way to help manage the lifecycle by defining a quantifiable end to what can otherwise seem like endless iteration. Of course, designers and managers can run out of time, money, and patience before they meet their UX targets—sometimes after just one round of evaluation—but at least then they know where things stand.

For a bit more discussion about the historical roots of UX metrics and targets, see Section 28.7.

22.2 UX TARGET TABLES

Through years of working with real-world UX professionals and doing our own user experience evaluations, we have refined the concept of a UX target table, in the form shown in Table 22-1, from the original conception of a usability

Table 22-1

Our UX target table, as evolved from the Whiteside, Bennett, and Holtzblatt (1988) usability specification table

Work Role: User Class	UX Goal	UX Measure	Measuring Instrument	UX Metric	Baseline Level	Target Level	Observed Results

User class

A description of the relevant characteristics of the user population who can take on a particular work role. User class descriptions can include such characteristics as demographics, skills, knowledge, experience, and special needs—for example, because of physical limitations (Section 9.3.4).

specification table as presented by Whiteside et al. (1988). A spreadsheet is an obvious way to implement these tables.

For convenience, one row in the table is called a "UX target." The first column is for the work role and related user class to which this UX target applies. The next two columns are for the related UX goal and the associated UX measure. These all go together because each UX measure is aimed at supporting a UX goal and is specified with respect to a work role and user class combination. Next, you will see where you get the information for these three columns.

As a running example to illustrate the use of each column in the UX target table, we will progressively set some UX targets for the new Ticket Kiosk System.

22.3 WORK ROLE AND USER CLASSES

Because UX targets are aimed at specific work roles, we label each UX target by work role. Recall that different work roles in the user models perform different task sets.

So the key task sets for a given work role will have associated usage scenarios or other task sequence representations, which will inform benchmark task descriptions we create as measuring instruments to go with UX targets. Within a given work role, different user classes will generally be expected to perform to different standards, that is, at different target levels.

Example: A Work Role and User Class for the Ticket Kiosk System

For the Ticket Kiosk System, let's begin by focusing on the user work role of the ticket buyer. As we saw earlier, user class definitions for a work role can be based on, among other things, level of expertise, disabilities and limitations, and other demographics. For this work role, user classes could include a casual town resident user from Middleburg and a student user from Middleburg University. In this example, we feature the casual town user, as shown in Table 22-2.

Table 22-2

Choosing a work role and a user class for a UX target

Work Role: User Class	UX Goal	UX Measure	Measuring Instrument	UX Metric	Baseline Level	Target Level	Observed Results
Ticket buyer: Casual new user, for occasional personal use							

22.4 UX GOALS

UX goals are high-level objectives for a UX design, stated in terms of user experience objectives. UX goals can be driven by business goals and reflect real use of a product and identify what is important to an organization, its customers, and its users. They are expressed as desired effects to be experienced in usage by users of features in the design and they translate into a set of UX measures to be assessed in evaluation.

You can extract UX goals from user concerns captured in work activity notes, the flow model, social models, and work objectives, some of which will be market driven, reflecting competitive imperatives for the product. User experience goals can be stated for all users in general, in terms of a specific work role or user class, or for specific kinds of tasks.

Examples of user experience goals include ease of use for all users, ease of remembering for intermittent users, power performance for experts, avoiding errors for safety-critical systems, high customer satisfaction, walk-up-and-use learnability for new users, and so on.

Work activity note

A brief, clear, concise, and elemental (relating to exactly one concept, idea, fact, or topic) statement used to document a single point about the work practice as synthesized from raw usage research data (Section 8.1.2).

Example: User Experience Goals for the Ticket Kiosk System

From our usage research data, we can define the primary high-level UX goals for the ticket buyer to include:

- Fast and easy walk-up-and-use user experience, with absolutely no user training.
- Fast learning so new user performance (after limited experience) is on par with that of an experienced user.
- High customer satisfaction leading to high rate of repeat customers.

Some other possibilities:

- High learnability for more advanced tasks.
- Draw, engagement, attraction.
- Low error rate for completing transactions correctly, especially in the interaction for payment.

Table 22-3

Choosing a work role and a user class for a UX target

Work Role: User Class	UX Goal	UX Measure	Measuring Instrument	UX Metric	Baseline Level	Target Level	Observed Results
Ticket buyer: Casual new user, for occasional personal use	Walk-up ease of use for new user						

Translating the goal of "fast-and-easy walk-up-and-use user experience" into a UX target table entry for the UX goal is straightforward. This goal refers to the ability of a typical occasional user to do at least the basic tasks on the first try, certainly without training or manuals. We see the beginnings of a UX target in Table 22-3.

Exercise 22-1: Identify UX Evaluation Goals for Your System

Goal: A little experience in stating user experience goals.

Activities: Review the work activity affinity diagram (WAAD) and user concerns in the social model for the system of your choice, noting user or customer concerns relating to UX goals.

Deliverables: A short list of UX goals for one user class of the system of your choice.

Schedule: A half hour or so (it should be easy by now).

22.5 UX MEASURES

Objective UX evaluation data

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

Within a UX target, *the UX measure is the general user experience characteristic to be measured with respect to usage of your UX design.* The choice of UX measure implies something about which types of measuring instruments and UX metrics are appropriate.

UX targets are based on quantitative data—both objective data, such as observable user performance, and subjective data, such as user opinion and satisfaction.

Some common UX measures that can be paired with quantitative metrics include:

- Objective UX measures (directly measurable by evaluators):
  - Initial performance.
  - Long-term performance (longitudinal, experienced, steady state).
  - Learnability.

- Retainability.
- Advanced feature usage.

- Subjective UX measures (based on user opinions):
  - First impression (initial opinion, initial satisfaction).
  - Long-term (longitudinal) user satisfaction.
  - Emotional impact.
  - Meaningfulness to user.

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).

Initial performance refers to a user's performance during the very first use (somewhere between the first few minutes and the first few hours, depending on the complexity of the system). Initial performance is a key UX measure because any user of a system must, at some point, use it for the first time.

Long-term performance typically refers to performance during more constant use over a longer period of time (fairly regular use over several weeks, perhaps). Long-term usage usually implies a steady-state learning plateau by the user; the user has become familiar with the system and is no longer constantly in a learning state.

Learnability and retainability refer, respectively, to how quickly and easily users can learn to use a system and how well they retain what they have learned over some period of time.

Advanced feature usage is a UX measure that helps determine the user experience of more complicated functions of a system. The user's initial opinion of the system can be captured by a first impression UX measure, whereas long-term user satisfaction refers, as the term implies, to the user's opinion after using the system for some greater period of time, after some allowance for learning.

Initial performance and first impression are appropriate UX measures for virtually every UX design. Other UX measures often play support roles to address more specialized UX needs. Conflicts among UX measures are not unheard of. For example, you may need both good learnability and good expert performance. In the design, those requirements can work against each other. This, however, just reflects a normal kind of design tradeoff. UX targets based on the two different UX measures imply user performance requirements pulling in two different directions, forcing the designers to stretch the design and face the tradeoff honestly.

Example: UX Measures for the Ticket Kiosk System

For the walk-up ease-of-use goal of our casual new user, let us start simply with just two UX measures: initial performance and first impression. Each UX measure will appear in a separate UX target in the UX target table, with the work role and user class repeated, as in Table 22-4.

Table 22-4

Choosing initial performance and first impression as UX measures

Work Role: User Class	UX Goal	UX Measure	Measuring Instrument	UX Metric	Baseline Level	Target Level	Observed Results
Ticket buyer: Casual new user, for occasional personal use	Walk-up ease of use for new user	Initial user performance					
Ticket buyer: Casual new user, for occasional personal use	Initial customer satisfaction	First impression					

22.6 MEASURING INSTRUMENTS: BENCHMARK TASKS

Subjective UX evaluation data

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

Within a UX target, *a measuring instrument is the means of generating values for the particular UX measure*, it's the vehicle through which values are measured for the UX measure.

Although you can get creative in choosing your measuring instruments, objective measures are commonly associated with a benchmark task—for example, a time-on-task measure as timed on a stopwatch, or an error rate measure made by counting user errors—and subjective measures are commonly associated with a user questionnaire—for example, the average user rating-scale scores for a specific set of questions.

For example, we will see that the objective "initial user performance" UX measure in the UX target table for the Ticket Kiosk System is associated with a benchmark task and the "first impression" UX measure is associated with a questionnaire. Both subjective and objective measures and data can be important for establishing and evaluating user experience coming from a design.

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).

22.6.1 What Is a Benchmark Task?

As a measuring instrument for an objective UX measure, a benchmark task is a representative task that you will have user participants attempt to accomplish in evaluation while you observe their performance and behavior. As such, a benchmark task is a "standardized" task that can be used to compare performance among different users and across different design versions.

22.6.2 Selecting Benchmark Tasks

Here are some guidelines for choosing kinds of benchmark tasks.

22.6.2.1 Address designer questions with benchmark tasks and UX targets

As designers work on UX designs, questions arise constantly. Sometimes the design team simply cannot decide an issue for themselves and they defer it to UX testing ("let the users decide").

Or maybe you do agree on the design for a feature, but you are very curious about how it will play out with real users. If you have kept a list of such design questions as they came up in design activities, they now play a role in setting benchmark tasks to get feedback from users.

22.6.2.2 Create benchmark tasks for a representative spectrum of user tasks

Choose realistic tasks intended to be used by each user class of a work role across the system. To get the best coverage for your evaluation investment, your choices should represent the cross section of real tasks with respect to frequency of performance and criticality to goals of the users of the envisioned product. Benchmark tasks are also selected to evaluate new features, "edge cases," and business-critical or mission-critical tasks. While some of these tasks may not be performed frequently, getting them wrong could cause serious consequences.

22.6.2.3 Start with short and easy tasks and then increase difficulty progressively

In most cases, it's best to start with relatively easy tasks to get users accustomed to the design and feeling comfortable in their role as evaluators. After building user confidence and engagement, especially with the tasks for the "initial performance" UX measure, you can introduce more features, more breadth, variety, complexity, and higher levels of difficulty.

Example: Initial Benchmark Task Choice for the Ticket Kiosk System

For our ticket kiosk system, maybe start with finding a movie that is currently playing. Then follow with searching for and reserving tickets for a movie that is to be showing 20 days from now and then go to more complex tasks such as purchasing concert tickets with seat and ticket selection.

22.6.2.4 Include some navigation where appropriate

In real usage, because users usually have to navigate to get to where they will do the operations specific to performing a task, you want to include the need for this navigation even in your earliest benchmark tasks. It tests their

knowledge of the fact that they do need to go elsewhere, where they need to go, and how to get there.

22.6.2.5 Avoid large amounts of typing (unless typing skill is being evaluated)

Avoid anything such as extensive typing in your benchmark task descriptions that can cause large user performance variation not related to user experience in the design.

22.6.2.6 Match the benchmark task to the UX measure

Obviously, if the UX measure is "initial user performance," the task should be among those a first-time user realistically would face. If the UX measure is about advanced feature usage, then, of course, the task should involve use of that feature to match this requirement. If the UX measure is "long-term usage," then the benchmark task should be faced by the user after considerable practice with the system. For a UX measure of "learnability," a set of benchmark tasks of increasing complexity might be appropriate.

22.6.2.7 Adapt scenarios or other task sequence representations already developed for design

Design scenarios clearly represent important tasks to evaluate because they have already been selected as key tasks in the design. However, you *must* remember to remove information about how to perform the tasks, which is usually abundant in a scenario. See guideline "Tell the user *what* task to do, but not *how* to do it" in the next section for more discussion.

22.6.2.8 Use tasks in realistic combinations to evaluate task flow

To measure user performance related to task flow, use combinations of tasks and activities such as those that will occur together frequently. In these cases, you should set UX targets for such combinations because difficulties related to user experience that appear during performance of the combined tasks can be different than for the same tasks performed separately. For example, in the Ticket Kiosk System, you may wish to measure user performance on the task thread of searching for an event and then buying tickets for that event.

Example: Benchmark Task for the Ticket Kiosk System

As another example, a benchmark task might require users to buy four tickets for a concert under a total of $200 while showing tickets in this price range for the upcoming few days as sold out. This would force users to perform the task of

searching through other future concert days, looking for the first available day with tickets in this price range.

22.6.2.9 Pick tasks where you think or know the design has weaknesses

In general, of course, the benchmark tasks you choose as measuring instruments should closely represent tasks real users will perform in a real work context. Avoiding such tasks where you know there might be design problems violates the spirit of UX targets and user experience evaluation, which is about finding user experience problems so that you can fix them, not about proving you are the best designer.

22.6.2.10 Don't forget to evaluate with your power users

Often, user experience for power users is addressed inadequately in product testing (Karn, Perry, & Krolczyk, 1997). Do your product business and UX goals include power use by a trained user population? Do they require support for rapid repetition of tasks or complex and possibly very long tasks? Does their need for productivity demand shortcuts and direct commands over interactive hand-holding?

If any of these are true, you must include benchmark tasks that match this kind of skilled and demanding power use. And, of course, these benchmark tasks must be used as the measuring instrument in UX targets that match up with the corresponding user classes and UX goals.

22.6.2.11 To evaluate error recovery, a benchmark task can begin in an error state

Effective error recovery is a kind of "feature" that designers and evaluators can easily forget to include. Yet no UX design can guarantee error-free usage, and trying to recover from errors is something most users are familiar with and can relate to. A "forgiving" design will allow users to recover from errors relatively effortlessly. This ability is definitely an aspect of your design that should be evaluated by one or more benchmark tasks.

22.6.2.12 Consider tasks to evaluate performance in "degraded modes" due to partial equipment failure

In large interconnected networked systems such as military systems or large commercial banking systems, especially involving multiple kinds of hardware, subsystems can sometimes fail. When this happens, will your part of the system give up and die or can it at least continue some of its intended functionality and give partial service in a "degraded mode?" If your application fits this description,

you should include benchmark tasks to evaluate the user's perspective of this ability accordingly.

22.6.2.13 Don't try to make a benchmark task for everything

Evaluation driven by UX targets is only an engineering sampling process. It will not be possible to establish UX targets for all possible classes of users doing all possible tasks. It is often stated that about 20% of the tasks in an interactive system account for 80% of the usage and vice versa. While these figures are obviously folkloric guesses, they carry a grain of truth to guide in targeting users and tasks in establishing UX targets.

Example: Benchmark Tasks as Measuring Instruments for the Ticket Kiosk System

For the Ticket Kiosk System, the first UX target in Table 22-4 contains an objective UX measure for "Initial user performance." An obvious choice for the corresponding measuring instrument is a benchmark task. Here we need a simple and frequently used task that can be done in a short time by a casual new user in a walk-up ease-of-use situation. An appropriate benchmark task would involve buying tickets to an event. Here is a possible description to give the user participant:

> BT1: Go to the Ticket Kiosk System and buy three tickets for the Monster Truck Pull on February 28 at 7 p.m. Get three seats together as close to the front as possible. Pay with a major credit card.

In Table 22-5, we add this to the table as the measuring instrument for the first UX target.

Let us say we want to add another UX target for the "initial performance" UX measure, but this time we want to add some variety and use a different

Table 22-5

Choosing "buy special event ticket" benchmark task as a measuring instrument for "initial performance" UX measure in first UX target

Work Role: User Class	UX Goal	UX Measure	Measuring Instrument	UX Metric	Baseline Level	Target Level	Observed Results
Ticket buyer: Casual new user, for occasional personal use	Walk-up ease of use for new user	Initial user performance	BT1: Buy special event ticket				
Ticket buyer: Casual new user, for occasional personal use	Initial customer satisfaction	First impression					

Table 22-6

Choosing "buy movie ticket" benchmark task as a measuring instrument for second initial performance UX measure

Work Role: User Class	UX Goal	UX Measure	Measuring Instrument	UX Metric	Baseline Level	Target Level	Observed Results
Ticket buyer: Casual new user, for occasional personal use	Walk-up ease of use for new user	Initial user performance	BT1: Buy special event ticket				
Ticket buyer: Casual new user, for occasional personal use	Walk-up ease of use for new user	Initial user performance	BT2: Buy movie ticket				
Ticket buyer: Casual new user, for occasional personal use	Initial customer satisfaction	First impression					

benchmark task as the measuring instrument—namely, the task of buying a movie ticket. In Table 22-6, we have entered this benchmark task in the second UX target, pushing the "first impression" UX target down by one.

22.6.3 Crafting Benchmark Task Contents

22.6.3.1 Remove any ambiguities with clear, precise, specific, and repeatable instructions

Unless resolving ambiguity is what we want users to do as part of the task, we must make the instructions in benchmark task descriptions clear and not confusing. Unambiguous benchmark tasks are necessary for consistent results; we want differences in user performance to be due to differences in users or differences in designs but usually not due to different interpretations of the same benchmark task.

As a subtle example, consider this "add appointment" benchmark task for the "initial performance" UX measure for an interdepartmental event scheduling system: "Schedule a meeting with Dr. Ehrich for a month from today at 10 a.m. in 133 McBryde Hall concerning the HCI research project."

For some users, the phrase "a month from today" can be ambiguous. Why? It can mean, for example, on the same date next month or it can mean exactly four weeks from now, putting it on the same day of the week. If that difference in meaning can make a difference in user task performance, you need to make the wording more specific to the intended meaning.

You also want to make your benchmark tasks specific so that participants don't get sidetracked on irrelevant details during testing. If, for example, a "find event" benchmark task is stated simply as "Find an entertainment event for sometime next week," some participants might make it a long, elaborate task, searching

around for some "best" combination of event type and date, whereas others would do the minimum and take the first event they see on the screen. To mitigate such differences, add specific information about event selection criteria.

22.6.3.2 Tell the user what task to do, but not how to do it

This guideline is very important; the success of your evaluation based on this task will depend on it. Sometimes we find students in early evaluation exercises presenting users with task instructions that spell out a series of steps to perform. They should not be surprised when the evaluation session leads to uninteresting results.

The users are just giving a rote performance of the steps as they read them from the benchmark task description. If you wish to test whether your UX design helps users discover how to do a given task on their own, you must avoid giving any information about *how* to do it. Just tell them *what* task to do and let them figure out how.

Examples: Instructional Wording in Benchmark Task for the Ticket Kiosk System

> Example (to do): "Buy two student tickets for available adjacent seats as close to the stage as possible for the upcoming Iris Dement concert and pay with a credit card."
>
> Example (*not* to do): "Click on the Special Events button on the home screen; then select More at the bottom of the screen. Select the Iris Dement concert and click on Seating Options..."
>
> Example (*not* to do): "Starting at the Main Menu, go to the Music Menu and set it as a Bookmark. Then go back to the Main Menu and use the Bookmark feature to jump back to the Music Menu."

22.6.3.3 Don't use words in benchmark tasks that appear specifically in the UX design

In your benchmark task descriptions, you must avoid using any words that appear in menu headings, menu choices, button labels, icon pop ups, or any place in the UX design itself. For example, don't say "Find the first event (that has such and such a characteristic)" when there is a button in the UX design labeled "Find." Instead, you should use words such as "Look for…" or "Locate…"

Otherwise, it is very convenient for your users to use a button labeled "Find" when they are told to "Find" something. It does not require them to think and, therefore, does not evaluate whether the design would have helped them find the right button on their own in the course of real usage.

22.6.3.4 Use work context and usage-centered wording, not system-oriented wording

Because benchmark task descriptions are, in fact, descriptions of user tasks and not system functionality, you should use usage-centered words from the user's work context and not system-centered wording. For example, "Find information about xyz" is better than "Submit query about xyz." The former is task oriented; the latter is more about a system view of the task.

22.6.3.5 Have clear start and end points for timing

In your own mind, be sure that you have clearly observable and distinguishable start and end points for each benchmark task and make sure you word the benchmark task description to use these end points effectively. These will ensure your ability to measure the time on task accurately, for example.

At evaluation time, not only must the evaluators know for sure when the task is completed, but *the participant must know when the task is completed.* For purposes of evaluation, the task cannot be considered completed until the user experiences closure.

The evaluator must also know when the user knows that the task has been completed. Don't depend on the user to say when the task is done, even if you explicitly ask for that in the benchmark task description or user instructions. Therefore, rather than ending task performance with a mental or sensory state (i.e., the user knowing or seeing something), it is better to incorporate a user action confirming the end of the task, as in the (to do) examples that follow.

Examples: Clear Starting and End Points in Benchmark Tasks

Example (*not* to do): "Find out how to set the orientation of the printer paper to 'landscape'." Completion of this task depends on the user knowing something and that is not a directly observable state. Instead, you could have the user actually set the paper orientation; this is something you can observe directly.

Example (*not* to do): "View next week's events." Completion of this task depends on the user seeing something, an action that you may not be able to confirm. Perhaps you could have the user view and read aloud the contents of the first music event next week. Then you know whether and when the user has seen the correct event.

Example (to do): "Find next week's music event featuring Rachel Snow and add it to the shopping cart."

Example (to do): Or, to include knowing or learning how to select seats, "Find the closest available seat to the stage and add to shopping cart."

Example (to do): "Find the local weather forecast for tomorrow and read it aloud."

22.6.3.6 Keep some mystery in it for the user

Don't always be too specific about what the users will see or the parameters they will encounter. Remember that real first-time users will approach your application without necessarily knowing how it works. Sometimes try to use benchmark tasks that give approximate values for some parameters to look for, letting the rest be up to the user. You can still create a prototype in such a way that there is only one possible "solution" to this task if you want to avoid different users in the evaluation ending in a different state in the system.

Example (to do): "Purchase two movie tickets to *Bee Movie* within 1.5 hours of the current time and showing at a theater within five miles of this kiosk location."

22.6.3.7 Annotate situations where evaluators must ensure preconditions for running benchmark tasks

Suppose you write this benchmark task: "Your dog, Mutt, seems perfectly healthy and energetic. Delete your appointment with the vet for Mutt's annual checkup from your calendar."

Every time a user performs this task during evaluation, the prototype calendar must start on the same "current" date and it must contain an existing appointment at some future date in the calendar so that each user can find it and delete it. You must attach a note in the form of rubrics (next point later) to this benchmark task to that effect—a note that will be read and followed in the evaluation activity.

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 and 22.6.4.4).

22.6.3.8 Use "rubrics" for special instructions to evaluators

When necessary or useful, add a "rubrics" section to your benchmark task descriptions as special instructions to evaluators, not to be given to participants in evaluation sessions. Use these rubrics to communicate a heads up about anything that needs to be done or set up in advance to establish task preconditions, such as an existing event in the kiosk system, work context for ecological validity, or a particular starting state for a task.

Benchmark tasks for addressing designer questions are especially good candidates for rubrics. In a note accompanying your benchmark task, you can alert evaluators to watch for user performance or behavior that might shed light on these specific designer questions.

22.6.4 Other Benchmark Task Mechanics

22.6.4.1 Put each benchmark task description on a separate sheet of paper

Yes, we want to save trees but, in this case, it is necessary to present the benchmark tasks to the participant only one at a time. Otherwise, the participant will surely read ahead, if only out of curiosity, and can become distracted from the task at hand.

As another reason for separate task descriptions, it is possible that not all participants will complete all tasks. There is no need for anyone to see that they have not accomplished them all. If they see only one at a time, they will never know and never feel bad.

Finally, if a task has a surprise step, such as a mid-task change of intention, that step should also be on a separate piece of paper, not shown to the participant initially. To save trees you can cut (with scissors) a list of benchmark tasks so that only one task appears on one smaller piece of paper.

22.6.4.2 Write a "task script" for each benchmark task

Sometimes it's useful to write a "task script" describing the steps of a representative or typical way to do the task and include it in the benchmark task document "package." This is just for use by the evaluator and is definitely not given to the participant. The evaluator may not have been a member of the design team and initially may not be too familiar with how to perform the benchmark tasks, and it helps the evaluator to be able to anticipate a possible task performance path. This is especially useful in cases where the participant cannot determine a way to do the task; then, the evaluation facilitator knows at least one way.

22.6.4.3 How many benchmark tasks and UX targets do you need?

As in most things UX, it depends. The size and complexity of the system should be reflected in the quantity and complexity of the benchmark tasks and UX targets. We cannot even give you an estimate of a typical number of benchmark tasks.

You have to use your engineering judgment and make enough benchmark tasks for reasonable, representative coverage without overburdening the evaluation process. If you are new to this, we can say that we have often seen a dozen UX targets, but 50 would probably be too much—not worth the cost to pursue in evaluation.

How long should your benchmark tasks be (in terms of time to perform)? The typical benchmark task takes a range of a couple of minutes to 10–15 minutes. Some short and some long are good. Longer sequences of related tasks are needed to evaluate transitions among tasks. Try to avoid really long benchmark tasks because they may be tiring to participants and evaluators during testing.

22.6.4.4 Ensure ecological validity

The extent to which your UX evaluation setup matches the user's real work context is called *ecological validity* (Thomas & Kellogg, 1989). One of the valid criticisms of lab-based user experience testing is that a UX lab can be kind of a sterile environment, not a realistic setting for the user and the tasks. But you can take steps to add ecological validity by asking yourself, as you write your benchmark task descriptions, how can the setting be made more realistic?

- What are the constraints in the user or work context?
- Does the task involve more than one person or role?
- Does the task require a telephone or other physical props?
- Does the task involve background noise?
- Does the task involve interference or interruption?
- Does the user have to deal with multiple simultaneous inputs, for example, multiple audio feeds through headsets?

As an example for a task that might be triggered by a telephone call, instead of writing your benchmark task description on a piece of paper, try calling the participant on a telephone with a request that will trigger the desired task. Rarely do task triggers arrive written on a piece of paper someone hands you. Of course, you will have to translate the usual boring imperative statements of the benchmark task description to a more lively and realistic dialogue: "Hi, I am Fred Ferbergen and I have an appointment with Dr. Strangeglove for a physical exam tomorrow, but I have to be out of town. Can you change my appointment to next week?"

Telephones can be used in other ways, too, to add realism to work context. A second telephone ringing incessantly at the desk next door or someone talking loudly on the phone next door can add realistic task distraction that you would not get from a "pure" lab-based evaluation.

For an anecdote about the need for ecological validity (the extent to which your UX evaluation setup matches the user's real work context) in the early testing of the early A330 Airbus, see Section 28.5.

Example: Ecological Validity in Benchmark Tasks for the Ticket Kiosk System

To evaluate use of the Ticket Kiosk System to manage the work activity of ticket buying, you can make good use of physical prototypes and representative locations to enhance ecological validity. By this we mean building a touchscreen display into a cardboard or wooden kiosk structure and placing it in the hallway of a relatively busy work area. Users will be subject to the gawking and questions of curiosity seekers. Having coworkers join the kiosk queue will add extra realism.

Exercise 22.2: Create Benchmark Tasks and UX Targets for Your System

Goal: To gain experience in writing effective benchmark tasks and measurable UX targets.

Activities: We have shown you a rather complete set of examples of benchmark tasks and UX targets for the Ticket Kiosk System. Your job is to do something similar for the system of your choice.

Begin by identifying which work roles and user classes you are targeting in evaluation (brief description is enough).

Write three or more UX table entries (rows), including your choices for each column. Have at least two UX targets based on a benchmark task and at least one based on a questionnaire.

Create and write a set of about three benchmark tasks to go with the UX targets in the table. Do NOT make the tasks too easy.

Make tasks increasingly complex.

Include some navigation.

Create tasks that you can later "implement" in your low-fidelity rapid prototype.

The expected average performance time for each task should be no more than about 3 minutes, just to keep it short and simple for you during evaluation.

Include the questionnaire question numbers in the measuring instrument column of the appropriate UX target.

Cautions and hints:

Do not spend any time on design in this exercise; there will be time for detailed design in an upcoming exercise.

Do not plan to give users any training.

Deliverables:

Two user benchmark tasks, each on a separate sheet of paper.

Three or more UX targets entered into a blank UX target table on your laptop or on paper.

If you are doing this exercise in a classroom environment, finish up by reading your benchmark tasks to the class for critique and discussion.

Schedule: Work efficiently and complete in about an hour and a half.

Subjective UX evaluation data

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

22.7 MEASURING INSTRUMENT: USER SATISFACTION QUESTIONNAIRES

As a measuring instrument for a subjective UX measure, a questionnaire related to various user UX design features can be used to determine a user's satisfaction with the UX design. Measuring a user's satisfaction provides a subjective, but still quantitative, UX metric for the related UX measure. As an aside, we should point out that objective and subjective measures are not always orthogonal. For example, very low user satisfaction can degrade user performance over a long period of time. In the following examples, we use the QUIS questionnaire (Section 24.3.2.4), but there are other excellent choices, including the System Usability Scale or SUS (Section 24.3.2.5).

Example: Questionnaire as Measuring Instrument for the Ticket Kiosk System

If you think the first two benchmark tasks (buying tickets) make a good foundation for assessing the "first-impression" UX measure, then you can specify that a particular user satisfaction questionnaire or a specific subset thereof be administered following those two initial tasks, stipulating it as the measuring instrument in the third UX target of the growing UX target table, as we have done in Table 22-7.

Example: Goals, Measures, and Measuring Instruments

Before moving on to UX metrics, in Table 22-8 we show some examples of the close connections among UX goals, UX measures, and measuring instruments.

Table 22-7

Choosing a questionnaire as the measuring instrument for first-impression UX measure

Work Role: User Class	UX Goal	UX Measure	Measuring Instrument	UX Metric	Baseline Level	Target Level	Observed Results
Ticket buyer: Casual new user, for occasional personal use	Walk-up ease of use for new user	Initial user performance	BT1: Buy special event ticket				
Ticket buyer: Casual new user, for occasional personal use	Walk-up ease of use for new user	Initial user performance	BT2: Buy movie ticket				
Ticket buyer: Casual new user, for occasional personal use	Initial customer satisfaction	First impression	Questions Q1–Q10 in the QUIS questionnaire				

Table 22-8

Connections among UX goals, UX measures, and measuring instruments

UX Goal	UX Measure	Potential Metrics
Ease of first-time use	Initial performance	Time on task
Ease of learning	Learnability	Time on task or error rate, after given amount of use and compared with initial performance
High performance for experienced users	Long-term performance	Time and error rates
Low error rates	Error-related performance	Error rates
Error avoidance in safety critical tasks	Task-specific error performance	Error count, with strict target levels (much more important than time on task)
Error recovery performance	Task-specific time performance	Time on recovery portion of the task
Overall user satisfaction	User satisfaction	Average score on questionnaire
User attraction to product	User opinion of attractiveness	Average score on questionnaire, with questions focused on the effectiveness of the "draw" factor
Quality of user experience	User opinion of overall experience	Average score on questionnaire, with questions focused on quality of the overall user experience, including specific points about your product that might be associated most closely with emotional impact factors
Overall user satisfaction	User satisfaction	Average score on questionnaire, with questions focusing on willingness to be a repeat customer and to recommend product to others
Continuing ability of users to perform without relearning	Retainability	Time on task and error rates reevaluated after a period of time off (e.g., a week)
Avoid having user walk away in dissatisfaction	User satisfaction, especially initial satisfaction	Average score on questionnaire, with questions focusing on initial impressions and satisfaction

22.8 UX METRICS

A *UX metric* describes the kind of value to be obtained for a UX measure. It states what is being measured. There can be more than one metric for a given measure. As an example from the software engineering world, software complexity is a measure; one metric for the software complexity measure (one way to obtain values for the measure) is "counting lines of code."

Probably the most common UX metrics are objective, performance-oriented, and taken while the participant is doing a benchmark task. Other UX metrics can be subjective, based on a rating or score computed from questionnaire results. Typical objective UX metrics include time to complete task[1] and number of errors made by the user. Others include frequency of help or documentation use; time spent in errors and recovery; number of repetitions of failed commands (what are users trying to tell us by repeating an action that did not work before?); and the number of commands, mouse clicks, or other user actions to perform task(s).

If you are feeling adventurous, you can use a count of the number of times the user expresses frustration or satisfaction (the "aha and cuss count") during his or her first session as an indicator of his or her initial impression of the UX design. Of course, because the number of remarks is directly related to the length of the session, plan your levels accordingly or you can set your levels as a count per unit time, such as comments per minute, to factor out the time differences. Admittedly, this measuring instrument is rather participant-dependent, depending on how demonstrative a participant feels during a session, whether a participant is generally a complainer, and so on, but this metric can produce some interesting results.

Typically, subjective UX metrics will represent the kind of numeric outcome you want from a questionnaire, usually based on simple arithmetic statistical measures such as the numeric average. Remember that you are going only for an engineering indicator of user experience, not for statistical significance.

And don't overlook a combination of measures for situations where you have performance tradeoffs. If you specify your UX metric as some function, such as a sum or an average, of two other performance-related metrics, for example, time on task and error rate, you are saying that you are willing to give up some performance in one area if you get more in the other.

[1]Although the time on task often makes a useful UX metric, it clearly is not appropriate in some cases. For example, if the task performance time is affected by factors beyond the user's control, then time on task is not a good measure of user performance. This exception includes cases of long and/or unpredictable communication and response-time delays, such as might be experienced in some website usage.

We hope you will explore many other possibilities for UX metrics, extending beyond what we have mentioned here, including:

- Percentage of task completed in a given time.
- Ratio of successes to failures.
- Time spent moving cursor (would have to be measured using software instrumentation, but would give information about the efficiency of such physical actions, necessary for some specialized applications).
- For visibility and other issues, fixations on the screen, cognitive load as indicated by correlation to pupil diameter, and so on using eye tracking.

Finally, be sure you match up your UX measures, measuring instruments, and metrics to make sense in a UX target. For example, if you plan to use a questionnaire in a UX target, don't call the UX measure "initial performance." A questionnaire does not measure performance; it measures user satisfaction or opinion.

Example: UX Metrics for the Ticket Kiosk System

For the initial performance UX measure in the first UX target of Table 22-8, as already discussed in the previous section, the length of time to buy a special event ticket is an appropriate value to measure. We specify this by adding average "time on task" as the metric in the first UX target of Table 22-9.

As a different objective performance measure, you might measure the average number of errors a user makes while buying a movie ticket. This was chosen as the

Table 22-9

Choosing UX metrics for UX measures

Work Role: User Class	UX Goal	UX Measure	Measuring Instrument	UX Metric	Baseline Level	Target Level	Observed Results
Ticket buyer: Casual new user, for occasional personal use	Walk-up ease of use for new user	Initial user performance	BT1: Buy special event ticket	Average time on task			
Ticket buyer: Casual new user, for occasional personal use	Walk-up ease of use for new user	Initial user performance	BT2: Buy movie ticket	Average number of errors			
Ticket buyer: Casual new user, for occasional personal use	Initial customer satisfaction	First impression	Questions Q1–Q10 in the QUIS questionnaire	Average rating across users and across questions			

value to measure in the second UX target of Table 22-9. You will often want to measure both these metrics during a participant's single performance of the same single task. A participant does not, for example, need to perform one "buy ticket" task while you time performance and then do a different (or repeat the same) "buy ticket" task while you count errors.

Finally, for the UX metric in the third UX target of Table 22-9, the subjective UX target for the first impression UX measure, let us use the simple average of the numeric ratings given across all users and across all the questions for which ratings were given (i.e., Q1–Q10).

22.9 BASELINE LEVEL

The baseline level is the benchmark level of the UX metric; it is the "talking point" level against which other levels are compared. It is often the level that has been measured for the current version of the system (automated or manual).

22.10 TARGET LEVEL

The target-level specification is a quantitative statement of an aimed-at or hoped-for value for a UX metric. The target level is an operationally defined criterion for success of the expected user experience. The target level for a UX metric is the minimum value indicating attainment of user experience success. Target levels not met in evaluation serve as focal points for improvement by designers.

22.11 SETTING LEVELS

The baseline level and target level in the UX target table are key to *quantifying user experience metrics*. But sometimes setting baseline and target levels can be a challenge. The answer requires determining what level of user performance and user experience the system is to support.

Obviously, level values are often "best guesses" but with practice, UX people become quite skilled at establishing reasonable and credible target levels and setting reasonable values.

Among the yardsticks you can use to set both baseline and target levels are:

- An existing system or previous version of the new system being designed.
- Competing systems, such as those with a large market share or with a widely acclaimed user experience.

Table 22-10

Setting baseline levels for UX measures

Work Role: User Class	UX Goal	UX Measure	Measuring Instrument	UX Metric	Baseline Level	Target Level	Observed Results
Ticket buyer: Casual new user, for occasional personal use	Walk-up ease of use for new user	Initial user performance	BT1: Buy special event ticket	Average time on task	Three minutes		
Ticket buyer: Casual new user, for occasional personal use	Walk-up ease of use for new user	Initial user performance	BT2: Buy movie ticket	Average number of errors	<1		
Ticket buyer: Casual new user, for occasional personal use	Initial customer satisfaction	First impression	Questions Q1–Q10 in questionnaire XYZ	Average rating across users and across questions	7.5/10		

Although it may not always be explicitly indicated in a UX target table, the baseline and target *levels shown refer to the mean over all participants* of the corresponding measure. That is, the levels shown don't have to be achieved by every participant in every session. So, for example, if we specify a target level of four errors for benchmark task BT 2 in the second UX target of Table 22-10 as a worst acceptable level of performance, there must be no more than *an average of* four errors, as averaged across all participants who perform the "buy movie ticket" task.

22.11.1 Setting the Baseline Level

Example: Baseline Level Values for the Ticket Kiosk System

To determine the values for the first two UX target baseline levels for the Ticket Kiosk System, we can have someone perform the benchmark tasks for buying a ticket for a special event and a movie using the MUTTS ticket counter. That might be quite different from what you expect users will be able to achieve using our new system, but it is a stake in the sand, something for comparison. Measuring a baseline level helps ensure that the UX metric is, in fact, measurable.

Suppose that buying a ticket for a special event takes about 3 minutes. If so, this value, 3 minutes, makes a plausible baseline level for the first UX target in Table 22-10. Because most people are already experienced with ticket offices, this value is not really for initial performance, but it gives some idea for that value.

MUTTS

MUTTS is the acronym for Middleburg University Ticket Transaction Service, our running example for most of the process chapters (Section 5.5).

To set a baseline value for the second UX target, for buying a movie ticket, it can be assumed that almost no one should make any errors doing this at a ticket counter, so let us set the baseline level as less than 1, as in Table 22-10.

To establish a baseline value for the first impression UX measure in the third UX target, we could administer the questionnaire to some users of MUTTS. Let us say we have done that and got an average score of 7.5 out of 10 for the first impression UX measure (a value we put in Table 22-10).

22.11.2 Setting the Target Level

Because "passing" the user experience test means meeting all your target levels simultaneously, you have to ensure that the target levels for all UX measures in the entire table must be, in fact, simultaneously attainable. That is, don't build in tradeoffs of the kind where meeting one target level goal might make it much more difficult to meet another related target level.

So how do you come up with reasonable values for your target levels? As a general rule of thumb, a target level is usually set to be an improvement over the corresponding baseline level. Why build a new system if it is not going to be better? Of course, improved user performance is not the only motivation for building a new system; increased functionality or just meeting user needs at a higher level in the design can also be motivating factors. However, the focus here is on improving user experience, which often means improving user performance and satisfaction.

For initial performance measures, you should set target levels that allow enough time, for example, for unfamiliar users to read menus and labels, think a bit, and look around each screen to get their bearings. So don't use levels for initial performance measures that assume users are familiar with the design.

Example: Target Level Values for the Ticket Kiosk System

In Table 22-11, for the first initial performance UX measure, let us set the target level to 2.5 minutes. In the absence of anything else to go on, this is a reasonable choice with respect to our baseline level of 3 minutes. We enter this value into the "Target level" column for the first UX target of the UX target table in Table 22-11.

With a baseline level of less than one error for the "Buy movie ticket" task, it would again be tempting to set the target level at zero, but that does not allow for *anyone ever* to commit an error. So let us retain the existing level, <1, as the target level for error rates, as entered into the second UX target of Table 22-11.

For the first impression UX measure, let us be somewhat conservative and set a target level of a mean score of 8 out of 10 on the questionnaire. Surely 80% is passing in most anyone's book or course. This goes in the third UX target of Table 22-11.

Table 22-11

Setting target levels for UX metrics

Work Role: User Class	UX Goal	UX Measure	Measuring Instrument	UX Metric	Baseline Level	Target Level	Observed Results
Ticket buyer: Casual new user, for occasional personal use	Walk-up ease of use	Initial user performance	BT1: Buy special event ticket	Average time on task	Three minutes, as measured at the MUTTS ticket counter	2.5 minutes	
Ticket buyer: Casual new user, for occasional personal use	Walk-up ease of use for new user	Initial user performance	BT2: Buy movie ticket	Average number of errors	<1	<1	
Ticket buyer: Casual new user, for occasional personal use	Initial customer satisfaction	First impression	Questions Q1–Q10 in questionnaire XYZ	Average rating across users and across questions	7.5/10	8/10	
Ticket buyer: Frequent music patron	Accuracy	Experienced usage error rate	BT3: Buy concert ticket	Average number of errors	<1	<1	
Casual public ticket buyer	Walk-up ease of use for new user	Initial user performance	BT4: Buy Monster Truck Pull tickets	Average time on task	Five minutes (online system)	2.5 minutes	
Casual public ticket buyer	Walk-up ease of use for new user	Initial user performance	BT4: Buy Monster Truck Pull tickets	Average number of errors	<1	<1	
Casual public ticket buyer	Initial customer satisfaction	First impression	QUIS questions 4–7, 10, 13	Average rating across users and across questions	6/10	8/10	
Casual public ticket buyer	Walk-up ease of use for user with a little experience	Just postinitial performance	BT5: Buy Dunkirk movie tickets	Average time on task	Five minutes (including review)	Two minutes	
Casual public ticket buyer	Walk-up ease of use for user with a little experience	Just postinitial performance	BT6: Buy Ben Harper concert tickets	Average number of errors	<1	<1	

22.11.3 A Few Additional Targets

Just for illustration purposes, we have added a few additional UX targets to Table 22-11. The UX target in the fourth row is for a regular music patron's task of buying a concert ticket using a frequent-customer discount coupon. The UX measure for this one is to measure experienced usage error rates using the "Buy concert ticket" benchmark task, with a target level of 0.5 (average).

Additional benchmark tasks used in the last two UX targets of the table are:

BT5: You want to buy a ticket for the movie *Dunkirk* for between 7–8 p.m. tonight at a theater within a 10-minute walk from the Metro station. First check to be sure this movie is rated PG-13 because you will be with your 15-year-old son. Then go to the reviews for this movie (to show us you can find the reviews, but you don't have to spend time reading them now) and then buy two general admission tickets.

BT6: Buy three tickets to the Ben Harper concert on any of the nights on the weekend of Sep. 29–Oct. 1. Get the best seats you can for up to $50 per ticket. Print out the directions for taking the Metro to the concert.

22.12 OBSERVED RESULTS

The final column in Table 22-11 is for o*bserved results*, a space reserved for recording values measured while observing users performing the prescribed tasks during formative evaluation sessions. As part of the UX target table, this column affords direct comparisons between specified levels and results of testing.

Because you typically will have more than one user from which observed results are obtained, you can either record multiple values in a single observed results column or, if desired, add more columns for observed results and use this column for the average of the observed values. If you maintain your UX target tables in spreadsheets, as we recommend, it is easier to manage observed data and results later in UX evaluation analysis (Chapter 26).

Exercise 22-3: Creating Benchmark Tasks and UX Targets for Your System

Write out descriptions of a few (3–4) key/interesting user tasks for your product or system. Say what to do, but not how. Using these as measuring instruments, and anything else appropriate, fill out a UX target table.

22.13 PRACTICAL TIPS AND CAUTIONS FOR CREATING UX TARGETS

Here we present some hints about filling out your UX target table:

- Be prepared to adjust your target level values based on initial observed results.

Sometimes in evaluation, you observe that users perform dramatically differently than you had expected when you set the levels. These results can mean serious problems with the design, but they can help you refine the target levels in UX targets, too. While it is possible to set the levels too leniently, it is also possible that you make your initial UX targets too demanding, especially in early cycles of iteration.

- Don't set nearly impossible average goals such as zero errors.

Because the target-level value is an average, even one error occurring anywhere in the session will prevent the result value from being zero.

- What about UX goals, metrics, and targets for usefulness and emotional impact?

Questionnaires and interviews can also be used to assess usefulness, emotional impact (such as branding issues), and meaningfulness.

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).

22.14 RAPID APPROACH TO UX GOALS, METRICS, AND TARGETS

As in most of the other process chapters, the process here can be abridged, trading rigor (e.g., completeness) for speed and lower cost. Possible steps of increasing abridgement include:

- Eliminate objective UX measures and metrics, but retain UX goals and quantitative subjective measures. Metrics obtained with questionnaires are easier and far less costly than metrics requiring empirical testing, lab-based or in the field.
- Eliminate all UX measures and metrics and UX target tables. Retain benchmark tasks as a basis for user task performance and behavior to observe in limited empirical testing for gathering qualitative data (UX problem data).
- Ignore UX goals, metrics, and targets altogether and use only rapid evaluation methods that later produce only qualitative data.