Building User Interfaces

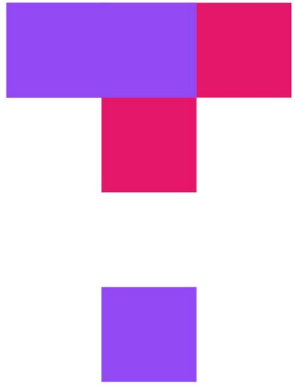
Usability Testing

Professor Bilge Mutlu

What we will learn today?

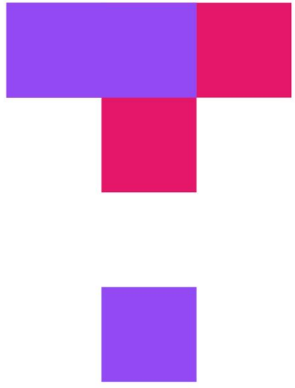
- Why Evaluate?
- Redefining Usability
- Usability Testing Basics
- Designing a User Test
- Measurement
- Assignment Preview

TopHat Attendance



TOP HAT

TopHat Questions



TOP HAT

Why evaluate?

Recap: What is UX design?[1]

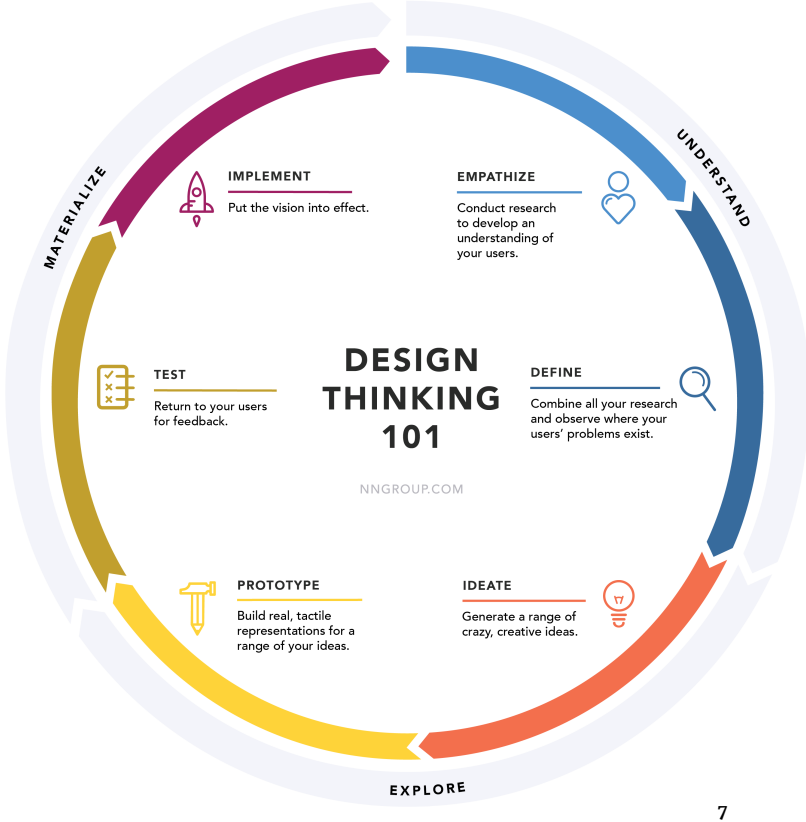
Definitions: User experience (UX) design is the **process** design teams use to create products that provide meaningful and relevant experiences to users.

[1] Source: Interaction Design Foundation

Recap: What is the design *process*?[2]

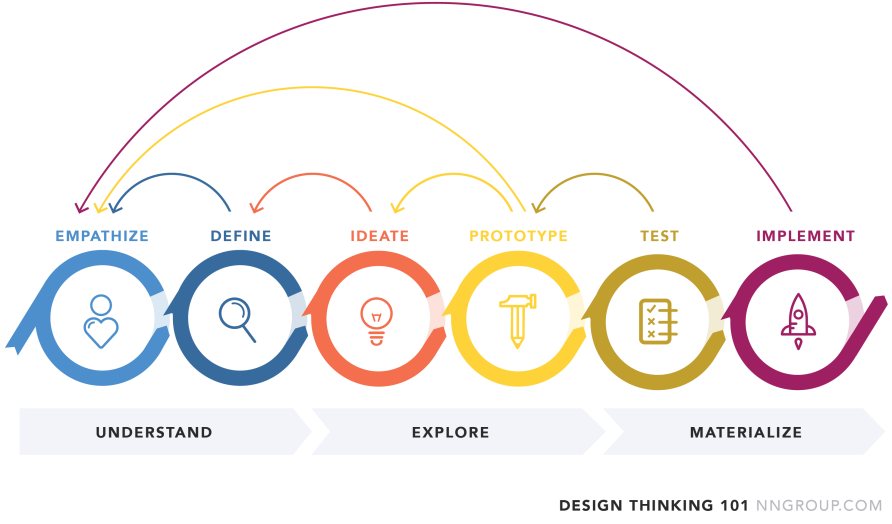
UX design usually involves the steps:

1. Empathize
2. Define
3. Ideate
4. Prototype
5. Test
6. Implement



[2] Image source: NN/g Design Thinking

3



[3] Image source: NN/g Design Thinking

Recap: Usability Evaluation

Usability: The *effectiveness*, *efficiency*, and *satisfaction* with which a specified set of users can achieve a specified set of tasks in a particular environment. — ISO 9241-11

Usability Evaluation: The assessment of the usability of design solutions.

Recap: Types of Usability Evaluation

1. **Testing-based** methods
2. **Expert-review-based** methods

Recap: Testing-based methods

Definition: Empirical, i.e., based on data, testing with users who represent the target population of design solutions.

Today, we will cover testing-based methods.

Redefining Usability

Redefining Usability

Definition: The *effectiveness*, *efficiency*, and *satisfaction* with which a specified set of users can achieve a specified set of tasks in a particular environment. — ISO 9241-11

We can detail this definition a bit more…

The Five-E Model of Usability[4]

Dimension	Definition
Effective	How completely and accurately the work or experience is completed or goals reached
Efficient	How quickly this work can be completed
Engaging	How well the interface draws the user into the interaction and how pleasant and satisfying it is to use
Error tolerant	How well the product prevents errors and can help the user recover from mistakes that do occur
Easy to learn	How well the product supports both the initial orientation and continued learning throughout the complete lifetime of use

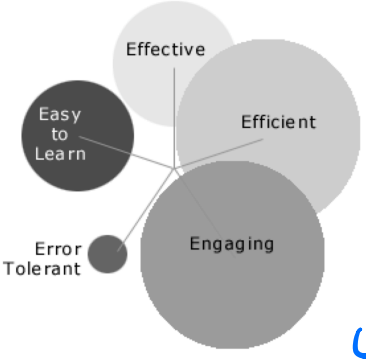
how does it prevent?
how does it help recovery

[4] Quesenbery, 2003, Dimensions of Usability

Different design and evaluation problems will require considering the five Es with different weights.[5]

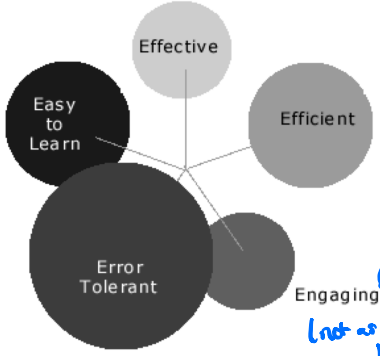
different types of interfaces may have different priorities

Five Es for a *museum site*:



(much more important)

Five Es for a *registration form*:



(not as important here)

[5] Image source: Quesenbery, 2003, Dimensions of Usability

How can the five Es guide evaluation?

Dimension 1: *Effective*

» Create scenarios with difficult or ambiguous tasks.

» Evaluate tasks for how accurately they are completed and how often they produce undetected errors.

Dimension 2: *Efficient*

» Construct the test with enough repetitions of typical tasks to create a realistic work rhythm. e.g. repeated copy / paste

» Use working software or a high fidelity prototype.

» Collect timing data, but also interview participants for their subjective impression of the program

↑ down-time may be appreciated

Dimension 3: *Engaging*

» Use satisfaction interview questions or surveys as part of the evaluation.

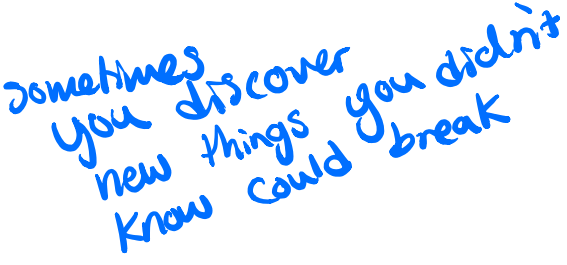
» Do comparative preference testing of presentation design.

Construct the test so that participants are able to abandon a product if they want.

Dimension 4: *Error Tolerant*

» Construct scenarios to create situations in which errors or other problems are likely.

» Observe how easily or accurately users are able to recover from problems when they occur.



Dimension 5: *Easy to Learn*

» Control how much instruction is given to test participants, or recruit participants with different levels of experience or knowledge.

» Mix frequently used task with scenarios for functions used less often or tasks with unusual variations.

balance availability of frequently used functionality with critical, but less frequently used functionality

Usability Evaluation Basics

Usability Testing

Definition: Observing users performing tasks with a design solution and asking them questions about their experience with the solution.

Observations include user actions, behavior, and verbal descriptions.

When do we use usability testing?

Depending on where usability testing is used in the design process, the testing can take two forms:

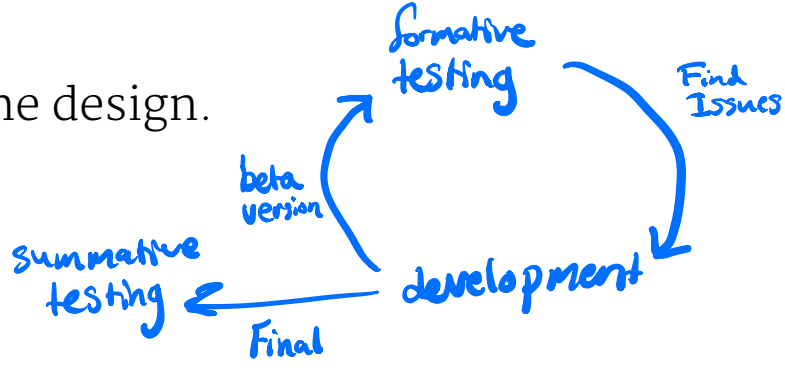
1. *Formative* testing → diagnose & respond to issues (iterative)
2. *Summative* testing → assess a final result

Formative Testing

Definition: Testing done throughout the design process to diagnose and address design problems.

Involves small number of users; used repeatedly; informs design improvements.

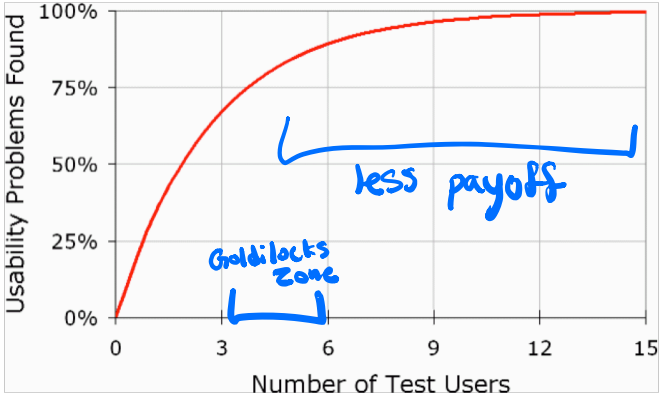
It "forms" the next iteration of the design.



How small can the testing be?[6]

Formative testing is considered to be a "discount" usability method, as 85% of usability solutions can be identified through testing with 5 users.

$$N(1-(1-L)^n)$$



[6] Image source: NN/g: Why you only need to test with 5 users

Summative Testing

Definition: Testing done at the end of the design process to establish the baseline usability of the design solution.

Involves a larger number of users; comparative testing; utilizes large number of metrics and statistics methods.

Usability Testing Contexts

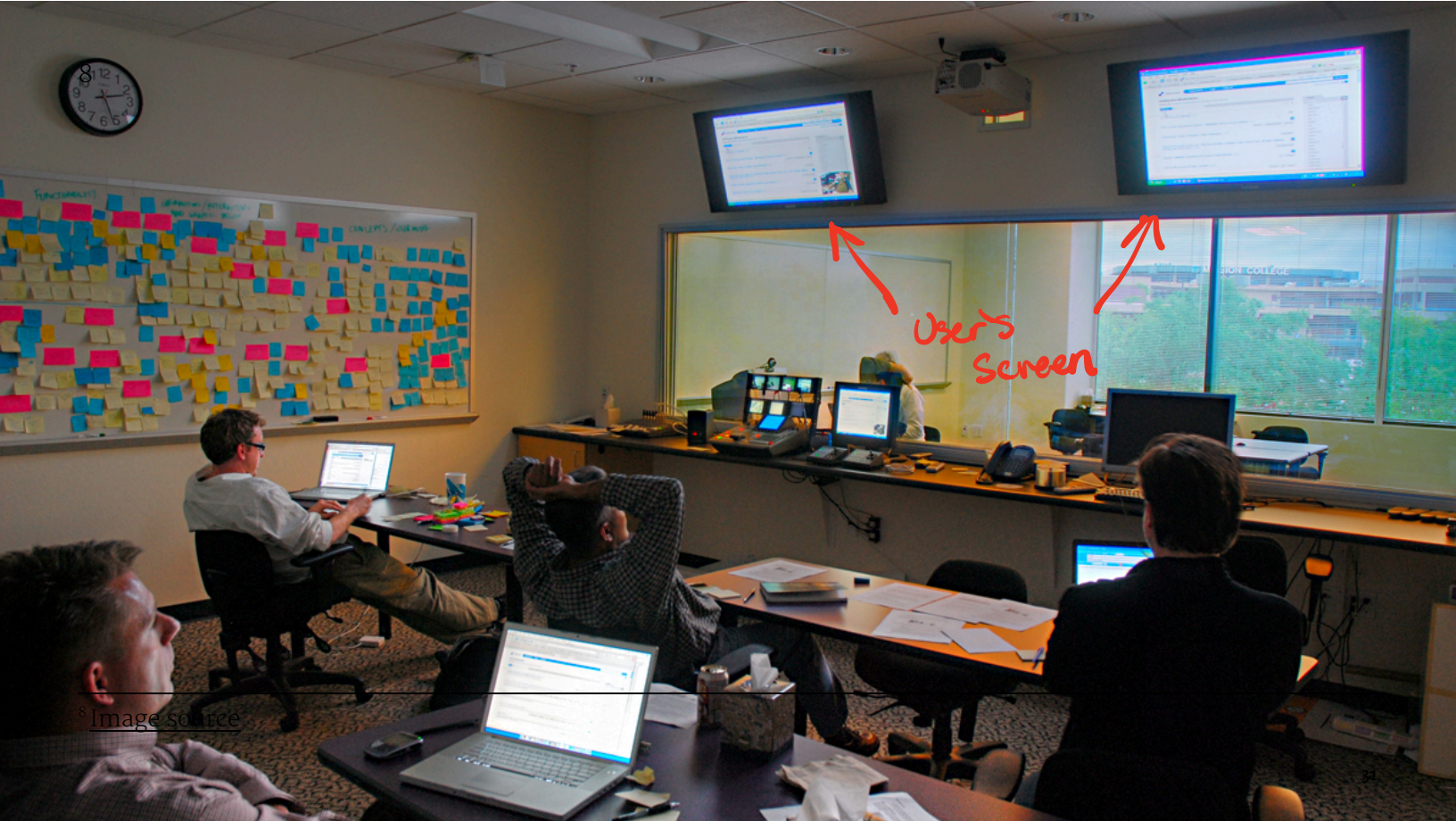
- Laboratory testing
- Field testing
  - Guerilla testing
  - Remote testing

Laboratory Testing[7]

Definition: Testing in the lab set up to capture user behavior through screen recording, software logging, over-the-shoulder video recording, eye tracking, etc. and to allow the design team to observe and analyze the test session.

[7] Image source


WebUsability
Usability
USABILITY



[8] Image source

Field Testing

Definition: Testing in the target setting of use for the design solution with the target profile of users.

Field testing is often also used to more easily recruit study participants.

Field Methods: *Guerilla Testing*[9]

Definition: Low-cost usability testing set up in a public space where passersby are recruited as test participants as volunteers or small compensation.

has critics, less controlled than laboratory testing

[9] Image source

Field Methods: *Remote Testing*

Definition: Testing a hi-fi prototype or early version of a deployed product over the internet.

Different forms of remote testing:

» **Moderated**: expert guides and observes, asks questions

» **Unmoderated**: application does testing, captures behavior

» **Automated**: data is collected over time, e.g., A/B testing

Designing a Usability Test

Key Dimensions of Usability Testing

When designing a usability test, we need to define and characterize the following four dimensions:

Why	What	How	Who
Goals	Scope, task/ scenarios	Approach, metrics	User subgroups, study team

The *Why*

Defining the why will involve determining test goals and provide a 10,000-feet view of the design goals.

» E.g., "improving accessibility of a website for older adults"

Different goals will result in entirely different test designs.

Formulating Test Goals

Formulate goals as questions that the test is designed to answer and specify two components.

Does our solution significantly improve accessibility for older adults over the previous design?

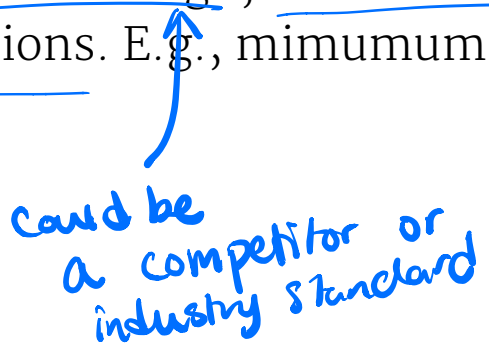
Points toward a comparative test with older adults.

To what extent do users consider our solution to be usable?

Points toward a study with standard metrics.

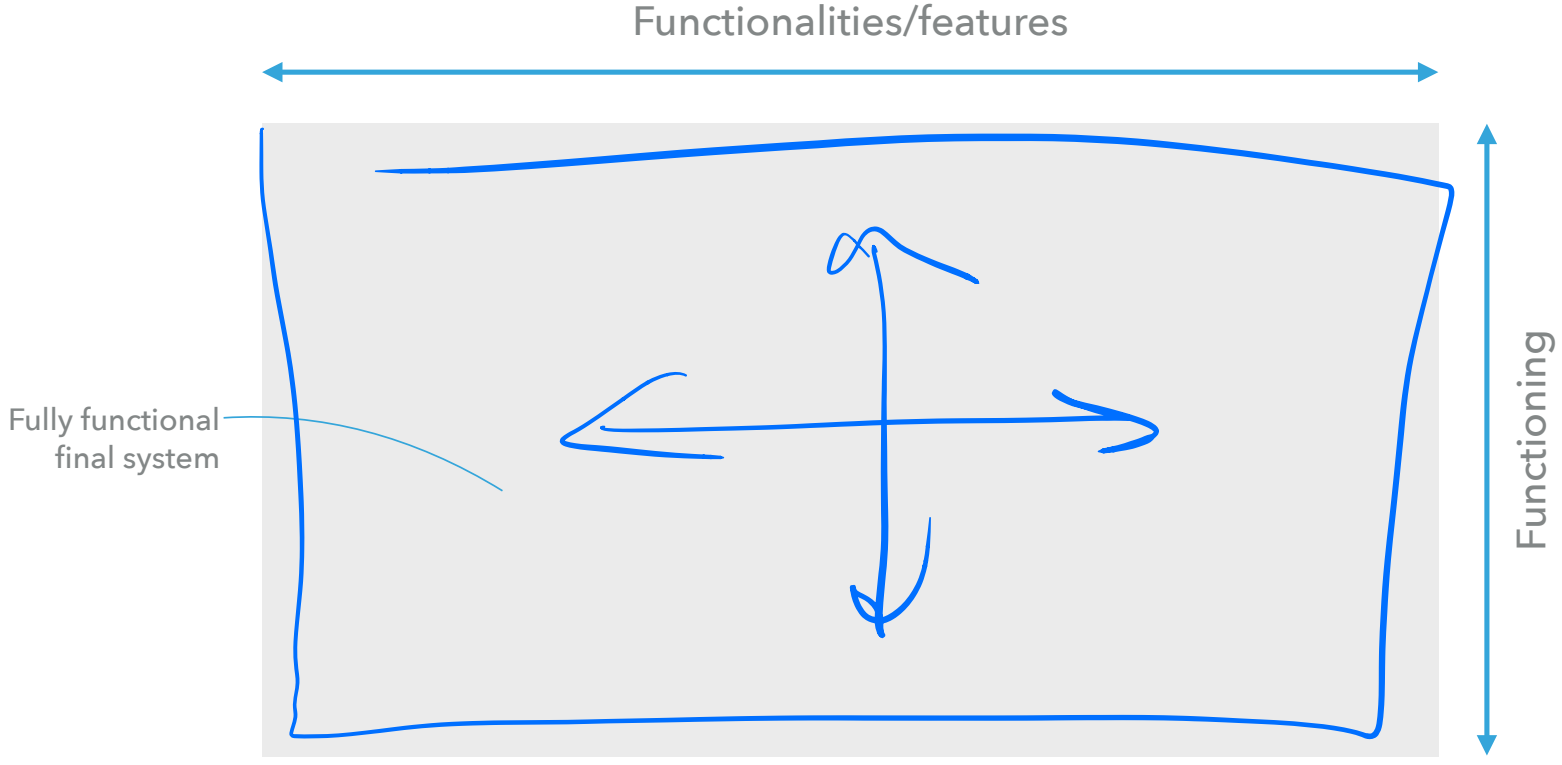
Test goals should specify:

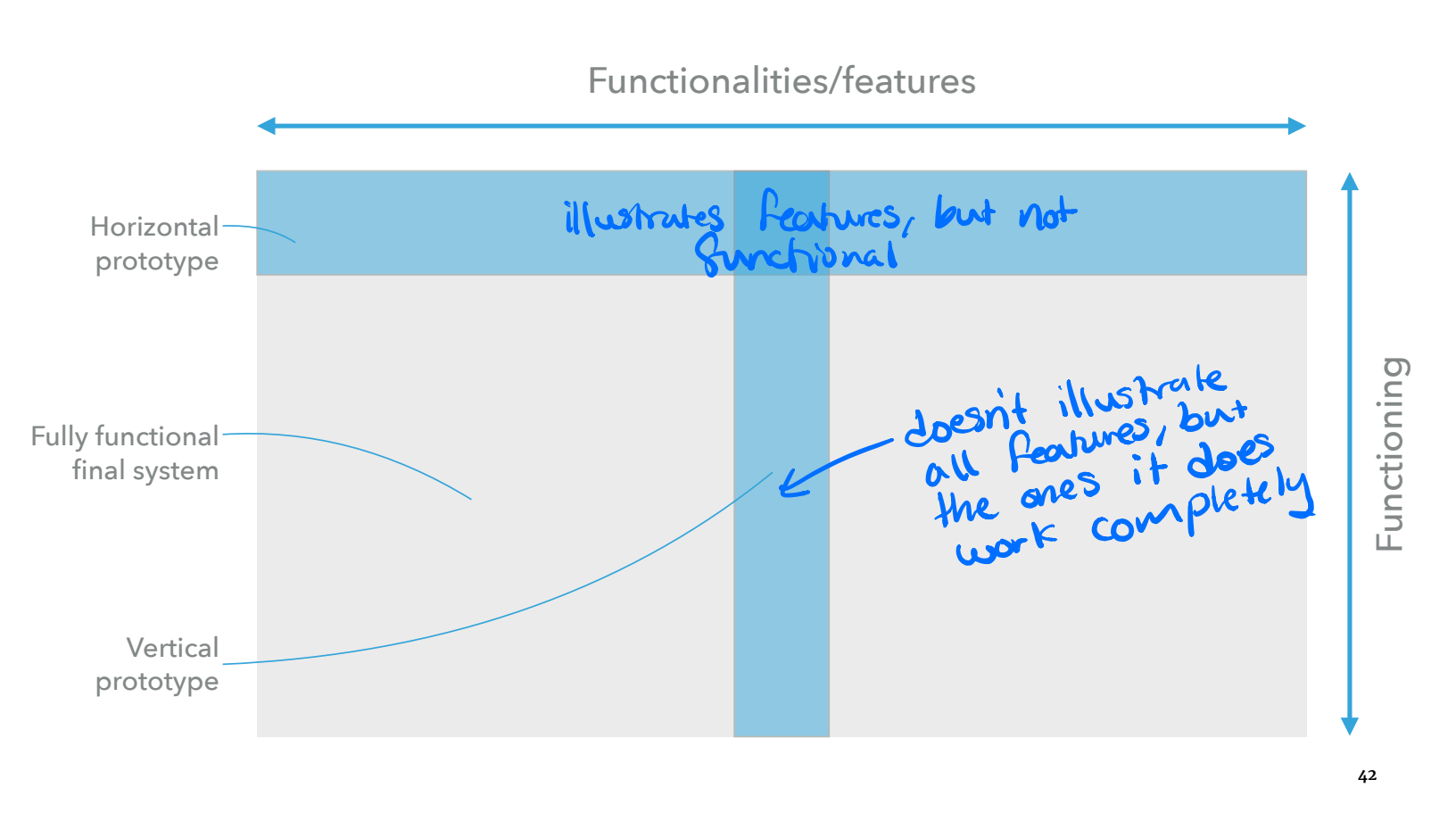
1. **Desired outcomes** capture how the design is expected to achieve. E.g., improved accessibility, reduced error rate, etc.
2. **Basis for comparison** specifies whether the outcome is with respect to a baseline, such as a previous design, established guidelines, or performance expectations. E.g., mimumum score on a standardized test.

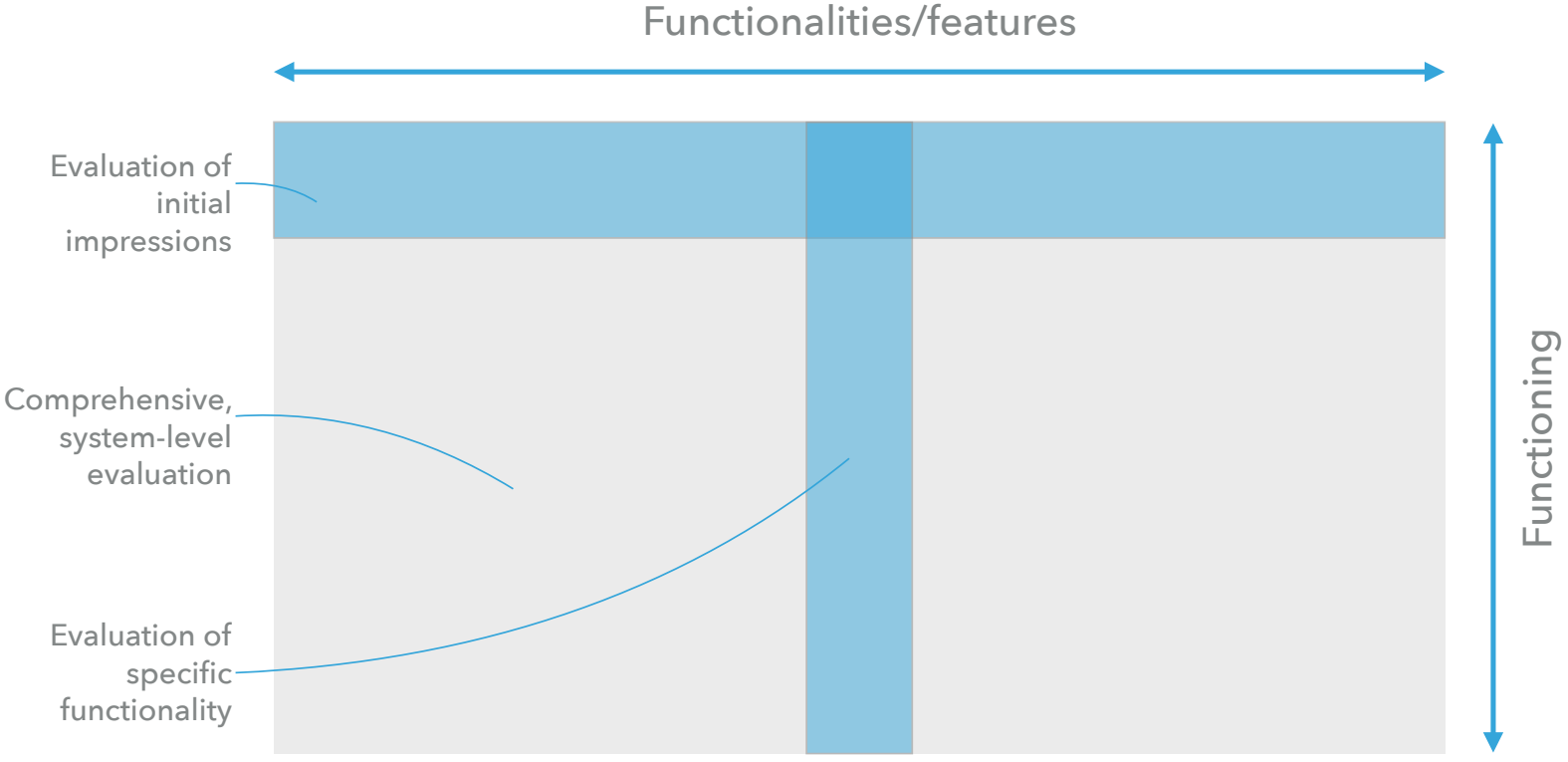


The *What*

We need to determine the scope of the testing, including what aspects of the system design, what tasks, and what scenarios will be included in the testing.


Functionalities/features
Fully functional final system
Functioning


Functionalities/features
Horizontal prototype
illustrates features, but not functional
Fully functional final system
doesn't illustrate all features, but the ones it does work completely
Vertical prototype
Functioning


Functionalities/features
Evaluation of initial impressions
Comprehensive, system-level evaluation
Evaluation of specific functionality
Functioning

Developing the *What*

Defining the *what* involves defining:

» **Questions:** Expectations of specific outcomes, e.g., whether or not the user will successfully achieve a particular goal.

» **Tasks:** The sequence of actions that users are expected to perform to achieve goals.

» **Scenarios:** Brief stories that provide users with context and goals in using the system.

An Example[10]

Question: Will users look at the top navigation bar to start their search for information?

Task: Seeking information about online programs for military personnel. Correct choice is Featured Degrees in top navigation bar. Users can also find a link to programs for military personnel in the description of featured programs in the center of the homepage, but it may be below the fold on their computer screen.

[10] Barnum, 2011, Usability Testing Essentials

Scenario: You have a friend in the military who wants to enroll in college courses while serving. You want to see if there are any online programs your friend could apply for. How would you go about doing this on this website?

How do we present multiple scenarios?

Scenarios should be ordered:

» From initial impressions to specific tasks

» From general to specific

» From short to long

» From simple to complex

Scenarios can also be presented *all at once* or *one at a time* depending on the testing goals.

The *How*

The how of the usability test will depend on:

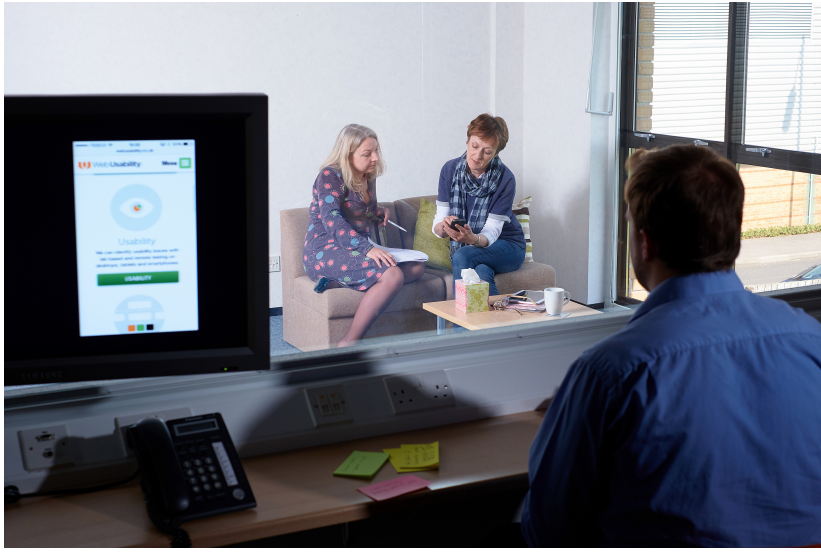
1. Whether the test is *formative* or *summative*
2. Whether the test is for a *single design* or *comparative*

What are you evaluating?
Comparison conditions?

Types of Measurement[11]

We can collect two types of data:

1. **Qualitative data:** observations of user actions and behavior, comments, and answers to questions
2. **Quantitative data:** measurements of user performance, error, and perceptions of the design

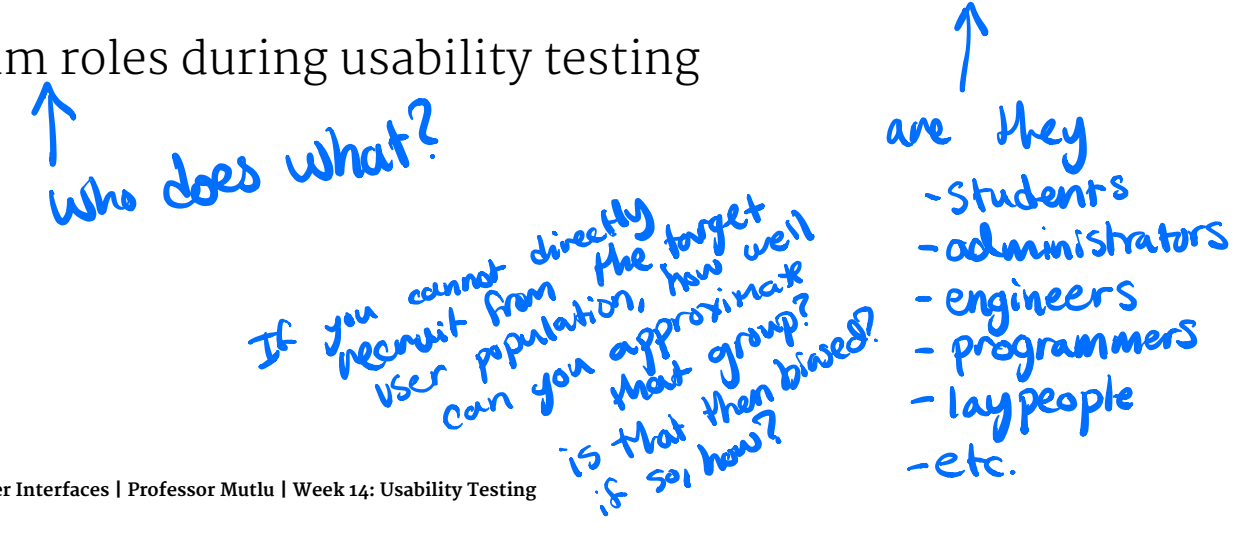


[11] Image source

The *Who*

The *who* of the usability test includes:

1. Participants who represent the target user population
2. Team roles during usability testing



Participants

Participants should represent the user subgroups that is targeted by the design. Subgroups characteristics can be defined by *experience*, *familiarity*, *skill*, *occupation*, *domain knowledge*, and *demographics*.

Once user subgroups are identified, several sessions of the study can be planned for each or a subset of the subgroups. The participants should be representative of the targeted subgroups.

The problem domain should also dictate participant representation. E.g., in a test for a budgeting app, users from different income levels can provide different insights.

Team Roles

The testing team usually involves:

1. **Moderator** who guides the participant and probes them with questions.
2. **Note-taker** who captures data.
3. **Observer(s)** from the UX team.
4. **Technician**, who operates the tested system or the testing equipment.

↑ sometimes beta versions break, sometimes recording equipment fails

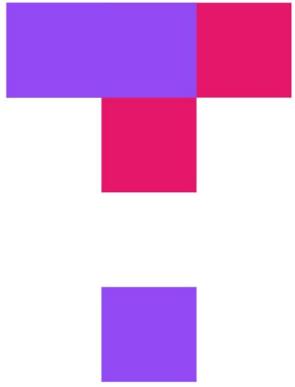
The Test Plan

The outcome of the design of the usability test is a *test plan* document that captures the *why*, *what*, *how*, and *who*.

Sometimes called a *test protocol*.

Supplements can include checklists for each role, moderator script, consent form, non-disclosure agreement (NDA) form.

TopHat Quiz



TOP HAT

Measurement

Types of Measures

1. Performance metrics → how fast / well / errors / etc
2. Self-report metrics → subjective measures / ratings
3. Issue-based metrics → events that occur

Performance metrics

Five basic types of performance metrics:

1. Task success
2. Time on task
3. Errors
4. Efficiency
5. Learnability

Task success: measures how effectively users are able to complete a given set of tasks. Can be used as binary or levels of success.

Time-on-task measures how much time is required to complete a task.

Errors measures the mistakes made during a task.

Efficiency measures the level of effort required to complete the task.

Learnability measures how performance changes over time.

Self-report metrics

Definition: Asking participants about their perceptions of and experience with a design solution using a set of questions.

Participants provide quantitative (e.g., ratings, rankings) or qualitative (e.g., open-ended, narrative) responses.

Could be a questionnaire or an interview

Commonly Used Self-Report Metrics

» System usability scale (SUS)

» USE scale

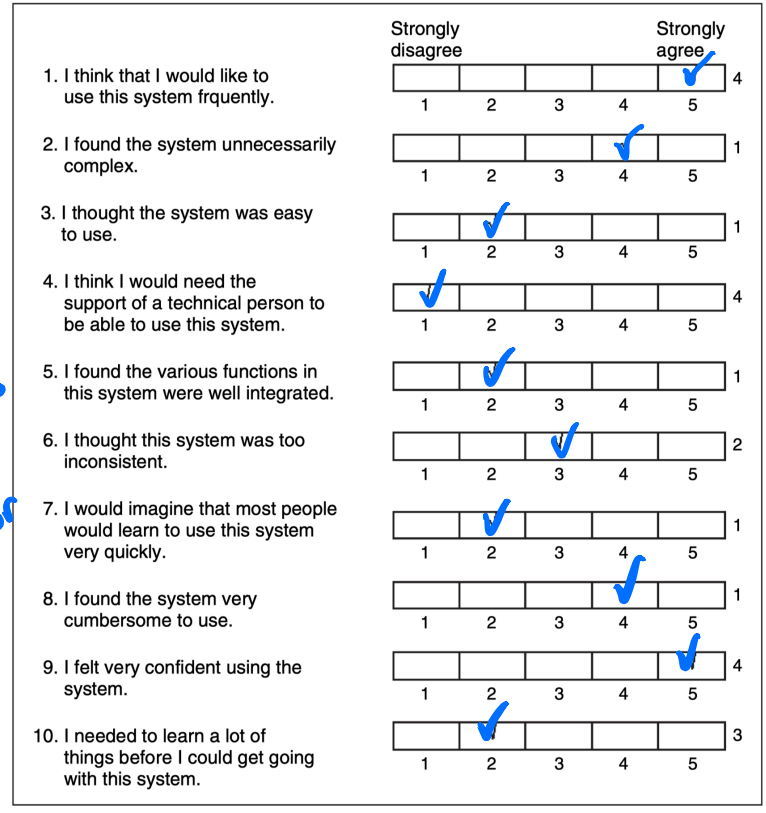
SUS[12] [13]

System Usability Scale

Ten-item questionnaire that focuses on usability.

Can be used for relative comparison or absolute benchmarking.

↑ Compared to previous score or industry competitor



[12] How to use the SUS

[13] Image source: Albert & Tullis, 2013, Measuring the User Experience

USE[14]

Includes four sub-scales for usefulness, ease of use, ease of learning, and satisfaction.

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 don't 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.

Users rate agreement with these statements on a 7-point Likert scale, ranging from strongly disagree to strongly agree. Statements in *italics* were found to weight less heavily than the others.

[14] Image source: Albert & Tullis, 2013, Measuring the User Experience

Likert Scales

A numerical 3-5-7-9-11 point scale with descriptive labels. E.g.:

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

Issue-based Metrics[15]

Definition: Problems that users encounter in using a system.

Examples:

» Behaviors that prevent task completion

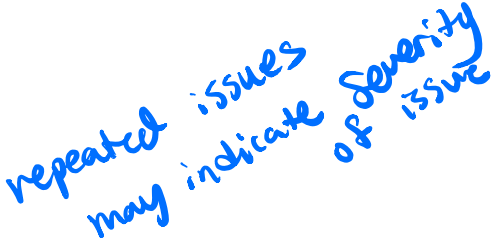
» Behaviors that takes someone "off course"

» An expression of frustration by the participant → "Ughh!"

» Not seeing something that should be noticed

[15] Albert & Tullis, 2013, Measuring the User Experience

- Participant says a task is complete when it is not
- Performing an action that leads away from task success
- Misinterpreting some piece of content
- Choosing the wrong link to navigate through web pages



How do we identify issues?

- User task actions
- User behavior
  - Verbal expressions of confusion, frustration, dissatisfaction, pleasure, surprise, confidence or indecision about a particular action that may be right or wrong
  - Not saying/doing what they should have done/said
  - Nonverbal behaviors, e.g., facial expressions, gaze

Severity ratings

Definition: Assessments of issues that help the design team prioritize design efforts. Based on:

1. Impact on user experience
2. Predicted frequency of occurrence
3. Impact on the business goals → if user couldn't complete purchase, etc.
4. Technical/implementation costs

Severity levels

Low: Issues that annoy or frustrate participants but do not play a role in task failure.

Medium: Issues that contribute to significant task difficulty but do not cause task failure.

High: Issues that lead directly to task failure.

Assignment Preview

Final Design Assignment

You will conduct a usability test of the shopping assistant.

Will be released tomorrow night, due on December 13.

Assignment Steps

Step 1. Design usability test, developing a *test plan*

Step 2. Execute usability test with ~3 participants

Step 3. Analyze findings, generate design insight

What did we learn today?

- Why Evaluate?
- Redefining Usability
- Usability Testing Basics
- Designing a User Test
- Measurement
- Assignment Preview