Aalto University
School of Arts, Design
and Architecture

Design B + Evaluation planning

MUO-E3036 Interaction Design (IxD)

31 January 2022

Antti Salovaara

MyCourses > Interaction design > Split S > Lecture slides > Week4-Day1-Evaluation-planning.pdf

Contents of today's teaching

Choosing what features to compare in evaluation

Planning the evaluation

Lots of group work during the day:

- Discussion on which designs you want to compare
- Discussions on useful evaluation arrangements

Friday's presentation instructions

Reading materials for Friday

Tutor meetings

Contents for all the remaining weeks

Week 4

Choose what the Design B will be
Plan the evaluation planning
Start recruitment of participants for the evaluation
Finish the creation of both designs for the evaluations
Present the designs and evaluation plan
Pilot test

Week 5

Do concrete preparations for the evaluations (e.g., staging, mockups)
Conduct the evaluations
Start the analysis of data from the evaluations
Lecture only on Monday morning (afternoon + Friday are free)

Week 6

Finish the analysis of data + find which design was better
Fine-tune the final design
Prepare final report

Fuzzy boundaries

Friday’s presentation instructions

Friday's presentation contents

10 minutes / group

Part 1: A vs B presentation

- Your UX goal(s)
- Final design A vs Final design B
- How these designs address your UX goal in different ways

Part 2: Evaluation plan

- How you will evaluate the designs with users
- How you will measure the UX goal(s)

Also: Submission of the evaluation plan to MyCourses

- https://mycourses.aalto.fi/mod/assign/view.php?id=861451
- DL: Friday 13:00

Evaluation plan template

1. Your UX goal(s): ½ page

   Name each goals + tell using your own words what it means in the case of your app or service

2. Present your designs (A and B): 1–2 pages

   Screenshots from your final designs + main interactions within and between the screens

   Clearly indicate + explain what makes A vs B different

   Explain the reasons for the two designs (e.g., how the differences related to your UX goals)

3. Evaluation plan: 2–4 pages

   Details of your methods: interview questions, usability test script (and division of work), questionnaires

   How these methods will answer which design (A or B) meets your UX goal better

Finishing the Designs A and B

Choosing what to compare

Designing A and B so that this comparison is possible

Ways by which designs may differ

Different interaction sequences

- “First step A, then B” vs “First step B, then A”

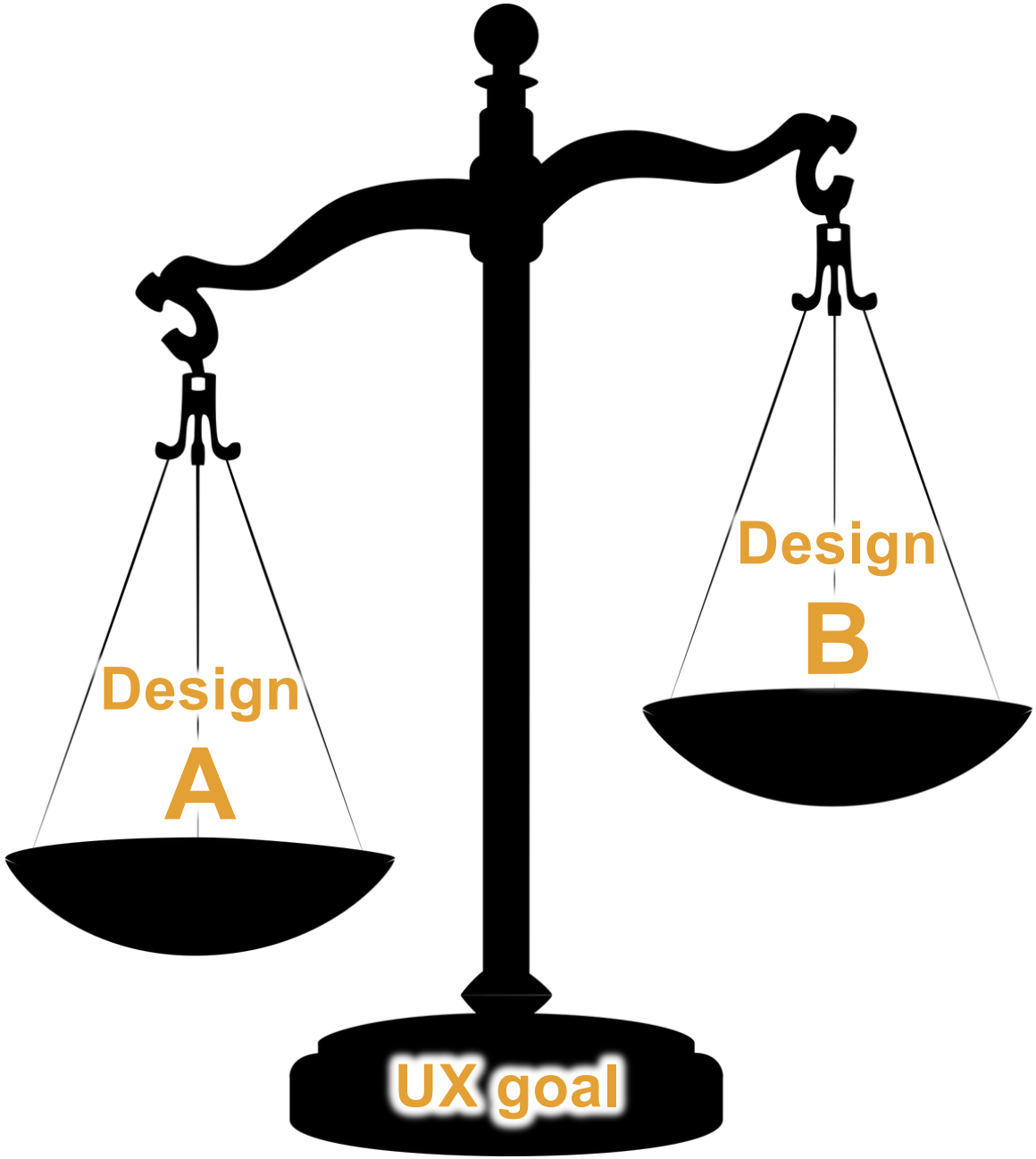
Different IxD patterns

- Wizard vs. accordion

Different solutions to the same problem

- Different information visualizations

Etc.


Design
A
Design
B
UX goal

Discussion in groups (15 mins)

Overall question:

- "What kind of design B should we have?"

Guiding questions:

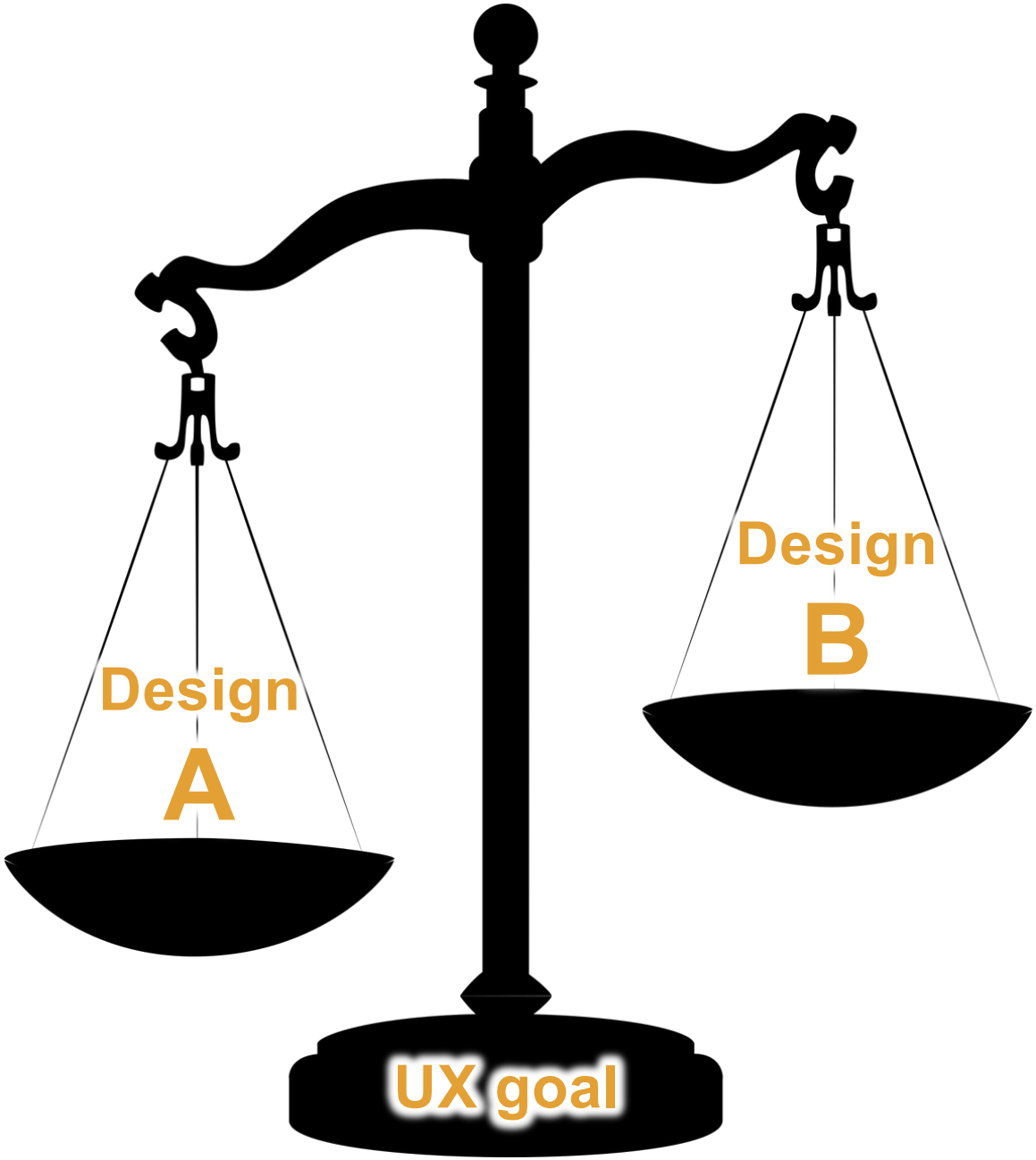
- What could be a better way to reach the UX goal than Design A?
- What kind of difference would be worth of a user evaluation?

Join the main session if/when you wish to discuss with me

After the discussion:

- Groups present their ideas to each other (**11**⇔12, **13**⇔14, **15**⇔16) + get feedback (15 mins)
- Joint discussion about questions that emerged

Planning the evaluation



What determines which design has more weight?

=> It is the data that you collect in your evaluation

Quick group task

Consider your most important UX goal

E.g., ease of use

Discuss: What kinds of data do you need so that you can evaluate this UX goal?

E.g., user's stress level, number of errors at the first try on the task, ...

Steps:

1. Start by brainstorming individually (5 mins)
2. Then share ideas within your group (5 mins)

Break

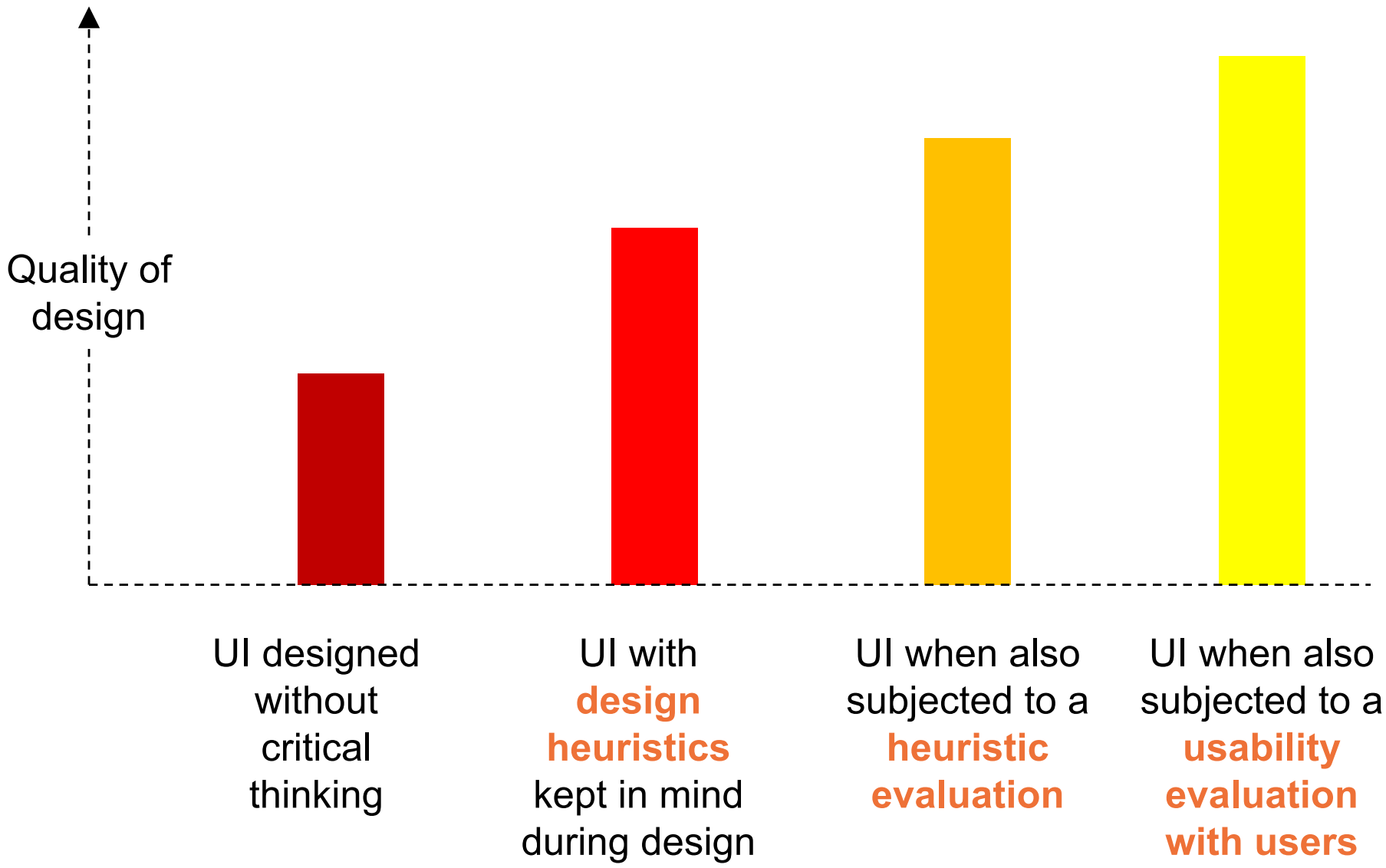
Types of evaluations

Heuristic evaluation: evaluation without users

Traditional scenario-driven usability evaluation

In-the-wild evaluation

How to reach a good usability and UX



Heuristic evaluation

UI's analysis using the design heuristics

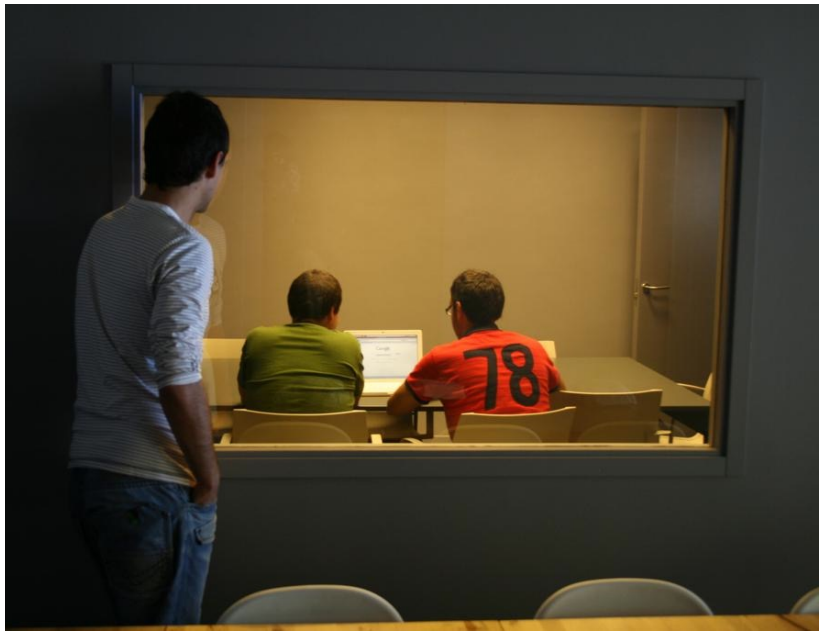
- Use both knowledge in the world and in the head
- Simplify the structure of tasks
- Make things visible
- Get the mappings right
- Exploit the power of constraints
- Design for error
- When all else fails: Standardize!

1. Strive for consistency
2. Enable frequent users to use shortcuts
3. Offer informative feedback
4. Design dialog to yield closure
5. Offer simple error handling
6. Permit easy reversal of actions
7. Support internal locus of control
8. Reduce short-term memory load

- Visibility of system status
- Match between system and the real world
- User control and freedom
- Consistency and standards
- Help users recognize, diagnose, and recover from errors
- Error prevention
- Recognition rather than recall
- Flexibility and efficiency of use
- Aesthetic and minimalist design
- Help and documen-tation

- Easy to learn
- Efficient
- Memorable
- Designed for error
- Satisfaction

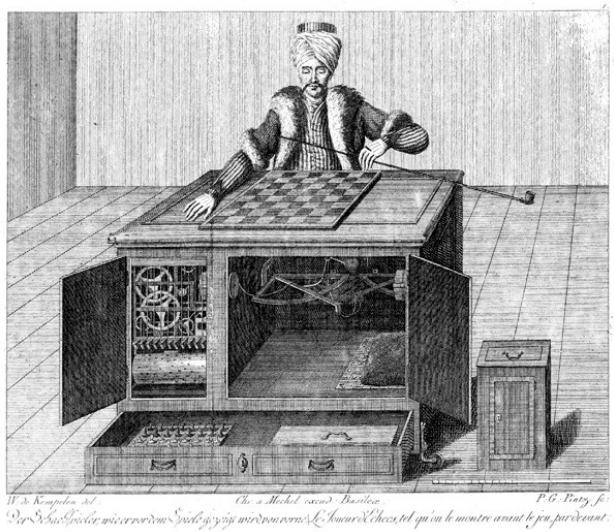
Traditional usability evaluation



Scenario-driven test:

1. Write realistic task scenarios for the features that need evaluation
2. Create mockup materials that make the unfinished system feel real
3. Present the scenario for the participant and ask him/her carry out the tasks.
4. Record with video
5. Repeat with more participants until findings “saturate”

Wizard-of-Oz evaluations



Copper engraving in Karl Gottlieb von Windisch, Briefe über den Schachspieler des Hrn. von Kempelen, nebst drei Kupferstichen die diese berühmte Maschine vorstellen. 1783. Public domain: Copyright expired.

Chess-playing automaton constructed by Wolfgang von Kempelen in 1770

https://hcde498processlog.wordpress.com/2015/05/11/wizard-of-oz-a-pen-that-corrects-you-when-you-write-off-line/

Definition:

"a research experiment in which subjects interact with a computer system that subjects believe to be autonomous, but which is actually being operated or partially operated by an unseen human being." (Wikipedia)

Use when:

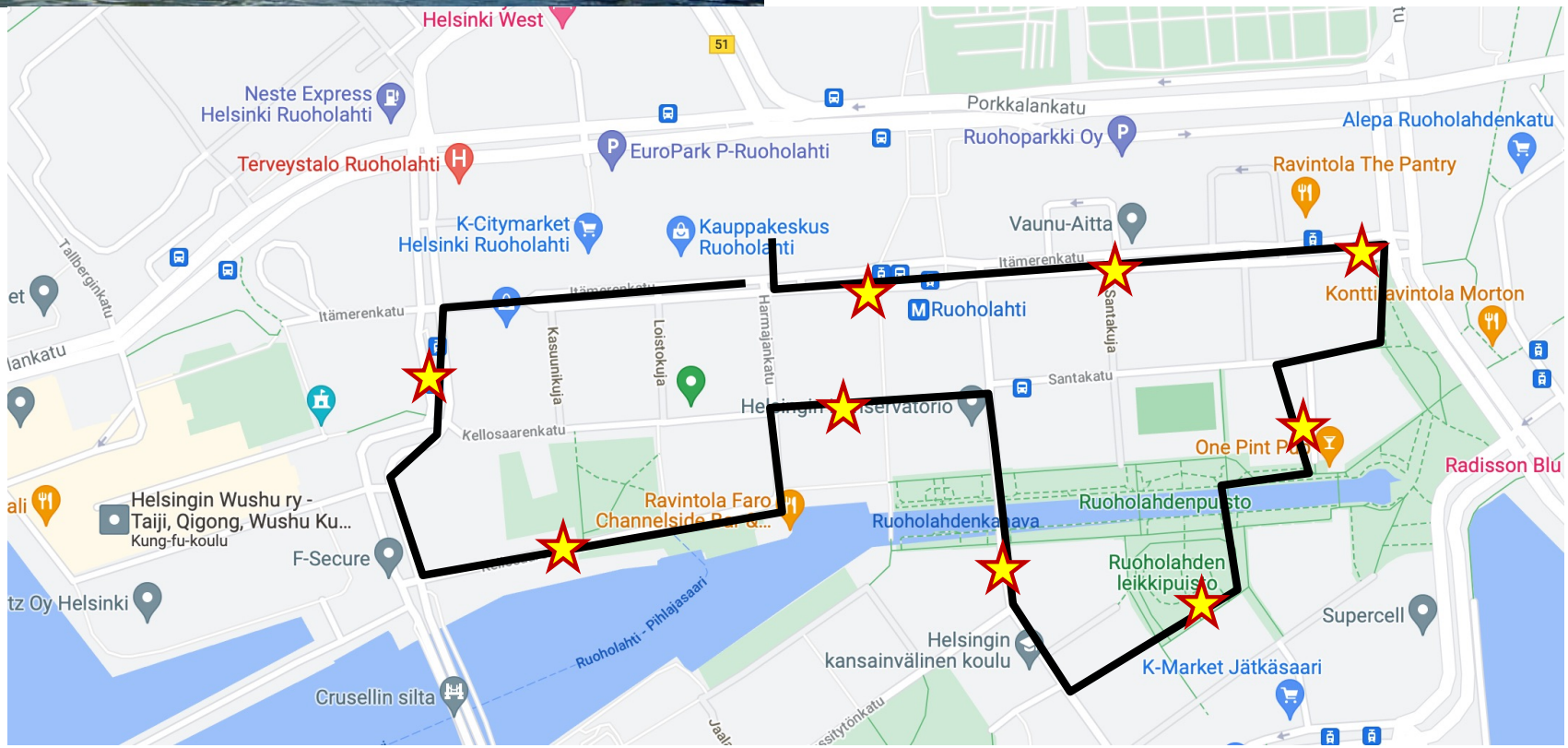
you can't prototype a computer to perform interactions

Ethics issue:

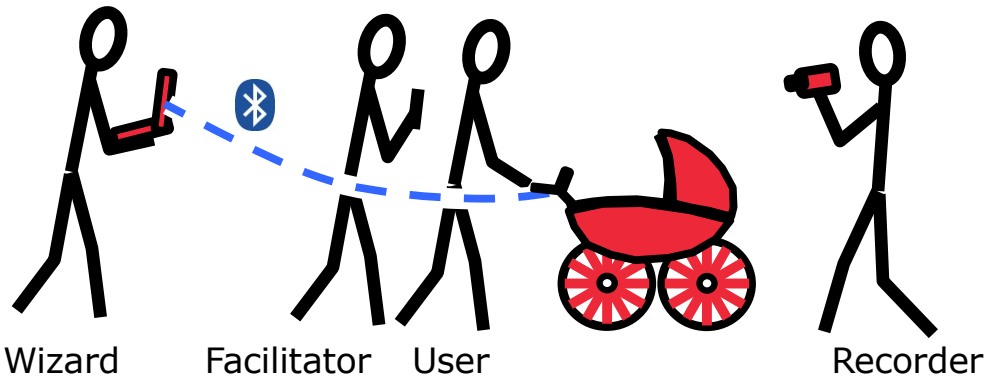
Setup is revealed after the study

Example: In-the-wild Wizard-of-Oz study

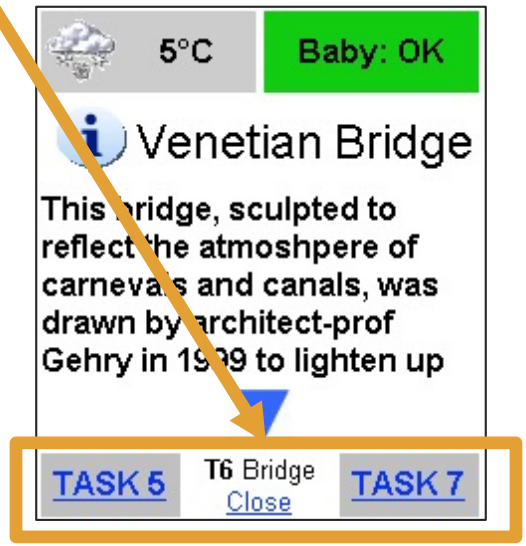
Would parents with babies be interested in location-based advertisements?



Wizard-of-Oz setup



Wizard's controls



How to control a prototype remotely

“In-the-wild” wizard-of-Oz studies on mobile phones:

- It is a great benefit if you can make user’s screen contents change at desired times
- Prepare a phone for the participant and allow its control from another phone

Investigate these options:

- https://joyofandroid.com/how-to-remotely-control-android-phone/
- https://www.androidauthority.com/how-to-remote-control-android-device-41969/

Think aloud method

Origins in psychological research on problem-solving and creativity*

Encourage the users to talk aloud:

- What they are trying to do
- What they are thinking

!!! Thinking aloud is not natural to many people

- A demonstration by the moderator and a practice task are needed to give the user an idea on what is expected
- Remember to remind the user politely ("Can you tell what you are now thinking?")

* E.g., Ericsson, K. A. (2006). Protocol analysis and expert thought: concurrent verbalizations of thinking during experts' performance on representatve task. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), Cambridge Handbook of Expertise and Expert Performance, ch. 13 (pp. 223--242). Cambridge University Press.

Break?

Arrangement of a user evaluation

Who should be recruited as users

How much data is needed

What users should be recruited?

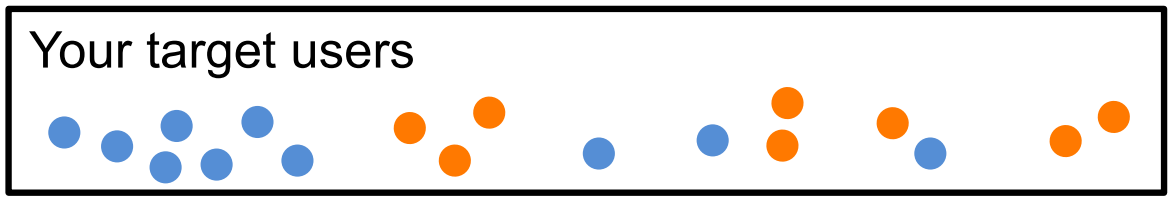
Random sampling

Each participant that you recruit has a **known probability** of being chosen for the study

Practically impossible in studies on humans

Convenience sampling

Studying people who you have a good access to (the typical method)



Choosing between heterogeneous vs homogeneous samples

Homogeneous (users very similar): If you need “deep” findings

Heterogenous (users differ a lot): Generalizable but shallower findings

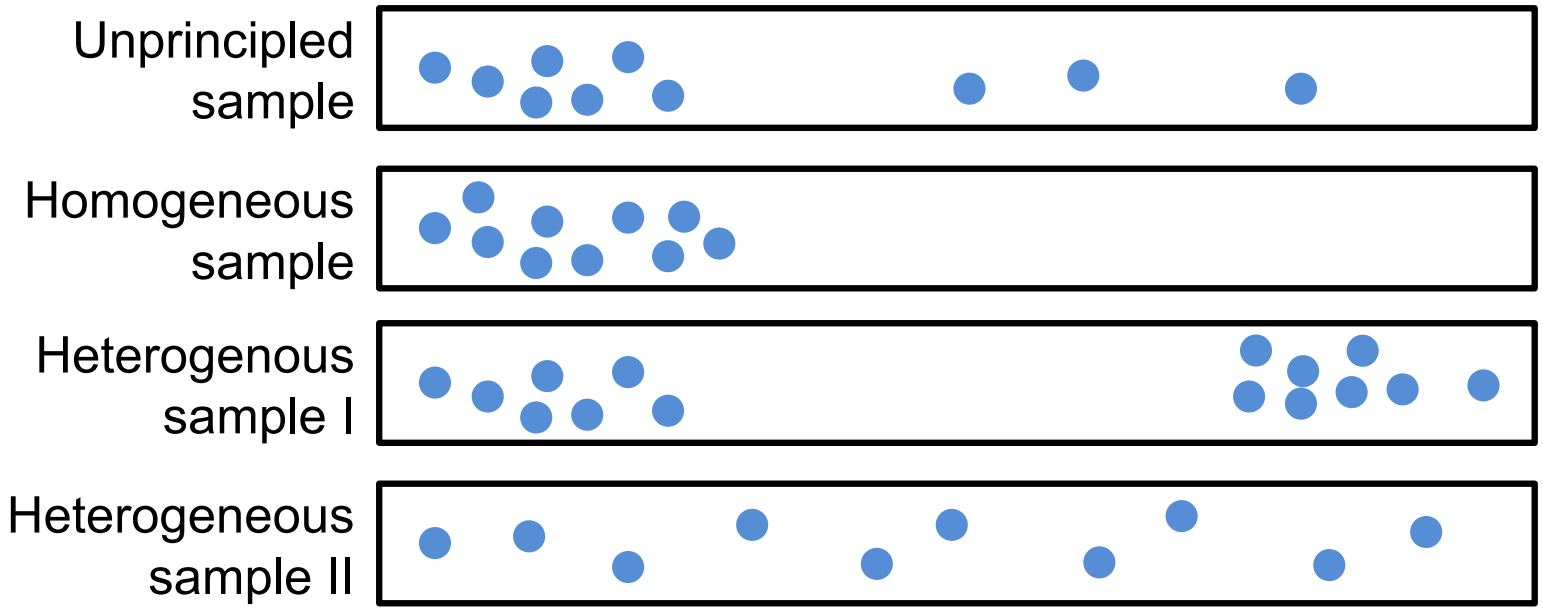
Choose between heterogeneous vs homogeneous samples

Homogeneous sample:

- Users are very similar
- Little noise in your data => You can get “deeper” findings

Heterogenous sample:

- Users differ a lot (e.g., in terms of age, gender, expertise, life values)
- A lot of noise and variability => Generalizable but shallower findings



Using same participants again?

Pros and cons of using your Sprint's test participants again:

Pros:

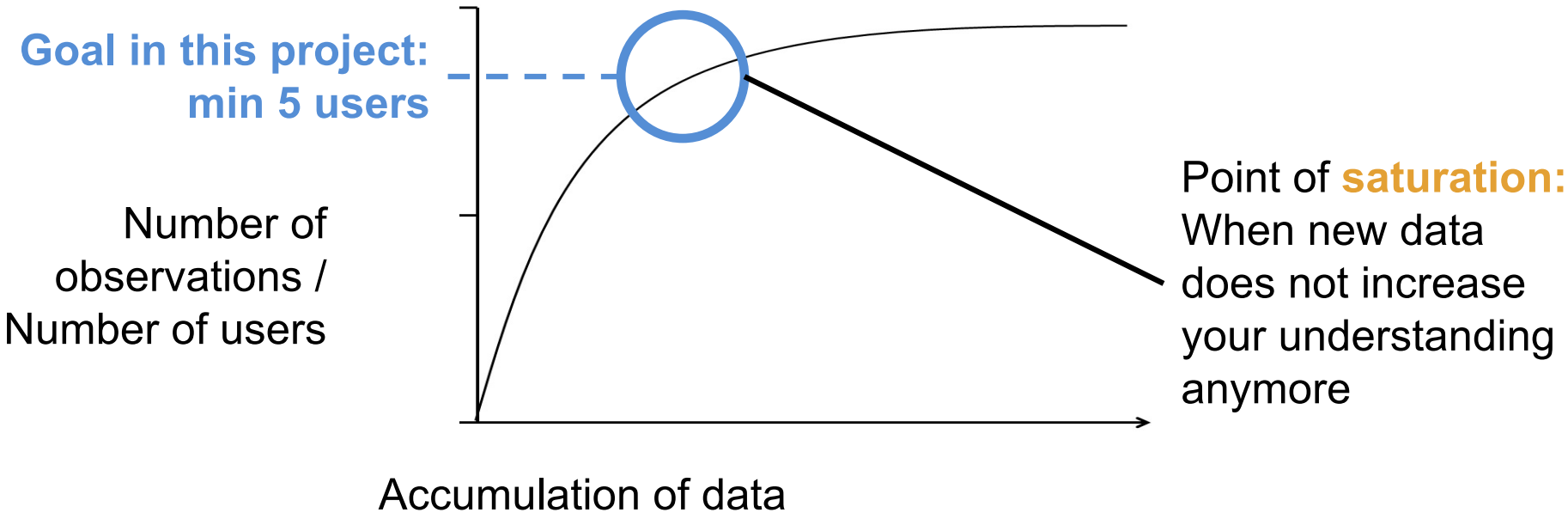
- More detailed feedback
- Easier recruitment
- No need to explain prototype in detail

Cons:

- Overfitting your design to individual users' needs
- Learning effect
- May go against your UX goal evaluation (e.g., is it possible to evaluate ease of use with a user that already knows the product?)

Recruitment from this course: Same issues

How much data is enough?



In usability evaluations, data can both quantitative and qualitative, but the analysis is almost always qualitative

Making the most out of every participant

To gather more data, include repetition in the scenario

- E.g., plan the first task to lead to suboptimal outcome, in order to make the user do something also another time
- “Ok, now I have done almost what I wanted, but this is not perfect. I’ll try to find a better solution, just a minute…”

Gather data in many ways simultaneously:

- Measure speed, errors etc.
- Use think-aloud to also find out what the user thinks
- Take video to observe behaviour and interactions
- Use a questionnaire (SUS, AttrakDiff, your own questions…)
- Interview about the experience after the evaluation

Mockups and staging

Although evaluations are unnatural…

(since user are recruited to carry out artificially constructed tasks)

…they should feel natural and believable

(to help the participants engage in the tasks and behave naturally)

Mockups: Preparation of authentic-feeling task materials

=> To evaluate a CAD software, prepare an unfinished 3D design that the user can work on

Staging: Making believable physical and social surroundings

=> To evaluate a wayfinding app for busy shopping malls, you have to create a context of a busy shopping mall

Comparative evaluations

A/B tests

Between-subjects vs within-subjects research designs

A/B tests

Gold standard for A/B tests:

- A randomized statistical test between two systems that differ only by one factor
- Example: in an online service, 10,000 visitors are directed to Design A, another set of 10,000 visitors are directed to Design B. The length of the visit is measured to find out which Design keeps them longer at the service.
- Read more: Wikipedia: “A/B testing”, Google: “ab test ux design”

In our course, the A/B test will be qualitative

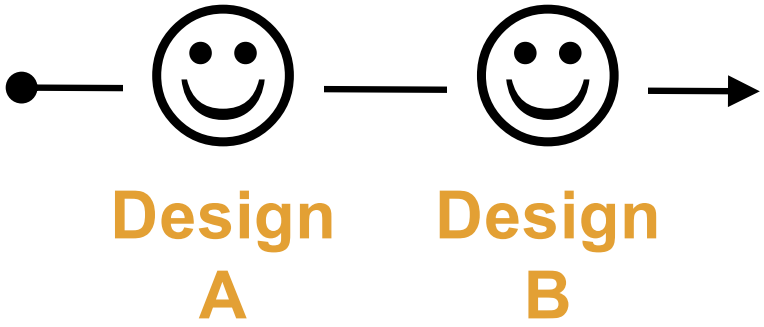
- What people say about Design A vs Design B
- How they use the designs differently
- Etc.

What designs will each user interact with?

Participant uses both Design A and Design B

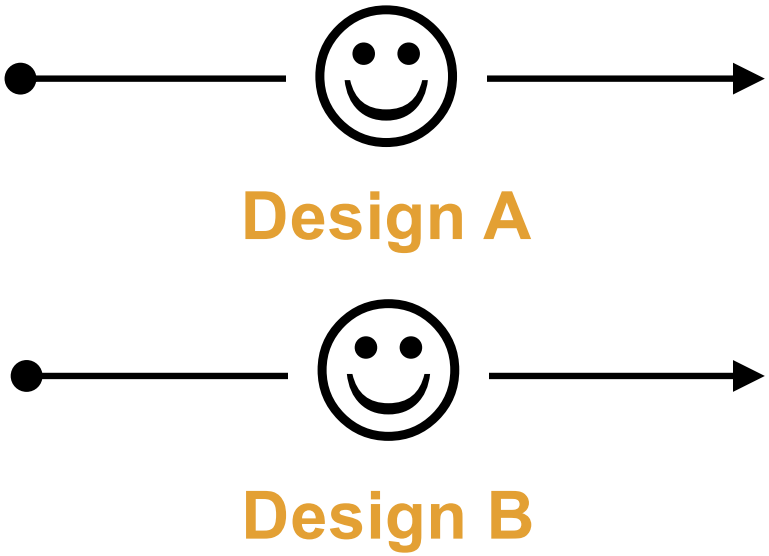
Within-subjects

(aka repeated measures)

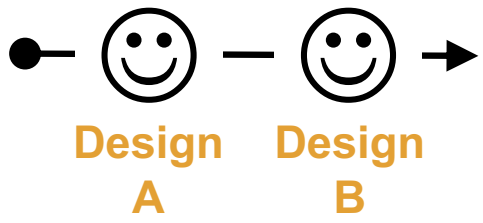


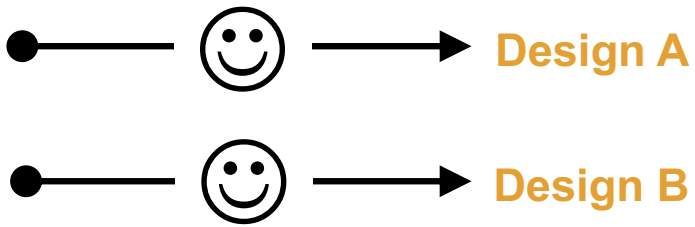
Participant user either Design A or Design B, but not both

Between-subjects



Pros and cons of between- and within-subjects tests





Within-subjects

- \+ A and B can be compared easily on user-by-user level
- \+ You get more data with a small number of people
- – Learning effect: participants learn to carry out Task B by carrying out Task A

Between-subjects

- \+ No learning effects
- – Need more participants

Counter-balancing helps:
50% of users start with Design A, the other 50% with Design B

Finally:

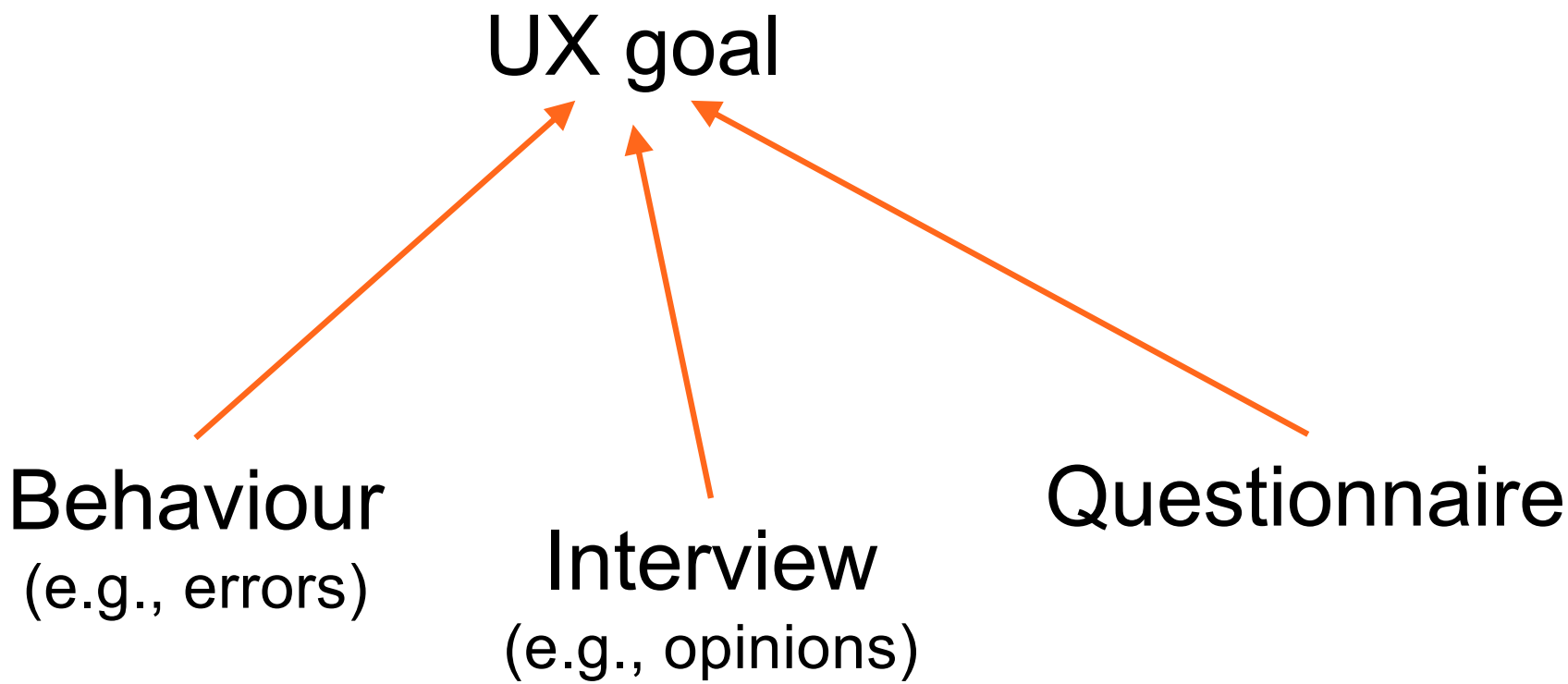
What data should you gather?

Triangulation

Examples of basic usability metrics

Triangulation

Try to measure the same question in several, complementary ways



Group your ideas from the quick group task

You did this task:

Quick group task

Consider one of your UX goals

E.g., ease of use

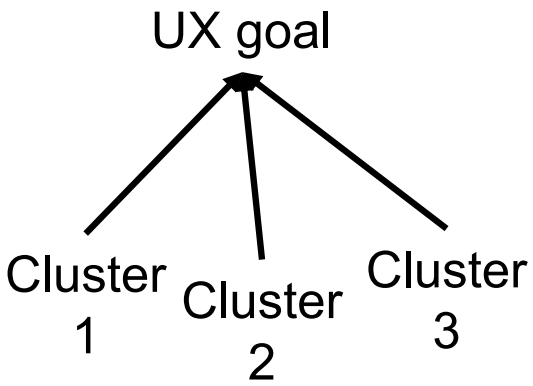
What kinds of data do you need so that you can evaluate this UX goal?

E.g., user's stress level, number of errors at the first try on the task, ...

Steps:

1. Start by brainstorming individually (5 mins)
2. Then share ideas within your group (5 mins)

Group you ideas into clusters that support each other



If you notice missed opportunities, add them to your lists

Common usability metrics

Table 10.3 Common usability metrics

Usability objective	Effectiveness measures	Efficiency measures	Satisfaction measures
Overall usability	• Percentage of tasks successfully completed • Percentage of users successfully completing tasks	• Time to complete a task • Time spent on non-productive actions	• Rating scale for satisfaction • Frequency of use if this is voluntary (after system is implemented)
Meets needs of trained or experienced users	• Percentage of advanced tasks completed • Percentage of relevant functions used	• Time taken to complete tasks relative to minimum realistic time	• Rating scale for satisfaction with advanced features
Meets needs for walk up and use	• Percentage of tasks completed successfully at first attempt	• Time taken on first attempt to complete task • Time spent on help functions	• Rate of voluntary use (after system is implemented)
Meets needs for infrequent or intermittent use	• Percentage of tasks completed successfully after a specified period of non-use	• Time spent re-learning functionsNumber of persistent errors	• Frequency of reuse (after system is implemented)
Learnability	• Number of functions learned • Percentage of users who manage to learn to a pre-specified criterion	• Time spent on help functions • Time to learn to criterion	• Rating scale for ease of learning

Benyon p. 226

Ask comparative questions

Use the comparative setup to gather deep answers:

"If you would need to analyse **how easy** these Designs were to use, how would you describe them?"

"Can you tell 2 good aspects from both Designs? How about 2 negative aspects?"

"How many stars, from 1 to 5, would you give to these Designs along the following dimensions: ease of use, efficiency, simplicity, beauty. Explain why."

Questionnaire-based measures

SUS

AttrakDiff

NASA TLX

System Usability Scale (SUS)

Usability.gov's description:

> "Quick and dirty", reliable tool for measuring the usability. It consists of a 10 item questionnaire with five response options for respondents; from Strongly agree to Strongly disagree.
>
> https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html

Example statements:

1. I think that I would like to use this system frequently.
3. I thought the system was easy to use.
6. I thought there was too much inconsistency in this system.

These can be great discussion topics after the user has given their responses

AttrakDiff

http://attrakdiff.de/index-en.html

- Contains a web tool to carry out all the analyses

AttrakDiff measures users' perceptions with 28 "semantic differentials":

- Ugly — Beautiful
- Confusing — Clear
- …

Result: three measures:

- Pragmatic (utilitarian) quality
- Hedonic (enjoyment-oriented) quality
- Attractiveness

Check out the use for A/B tests:

- http://attrakdiff.de/index-en.html#tab-vergleich-ab

NASA TLX (task load index)

Measures subjective perception of task load

Traditional version:

- 6 statements
- Ranking of the statements task
- Score calculation

"Raw NASA":

- Plain average of the 6 statements

More info + where to get it:

- https://humansystems.arc.nasa.gov/groups/TLX/
- https://en.wikipedia.org/wiki/NASA-TLX

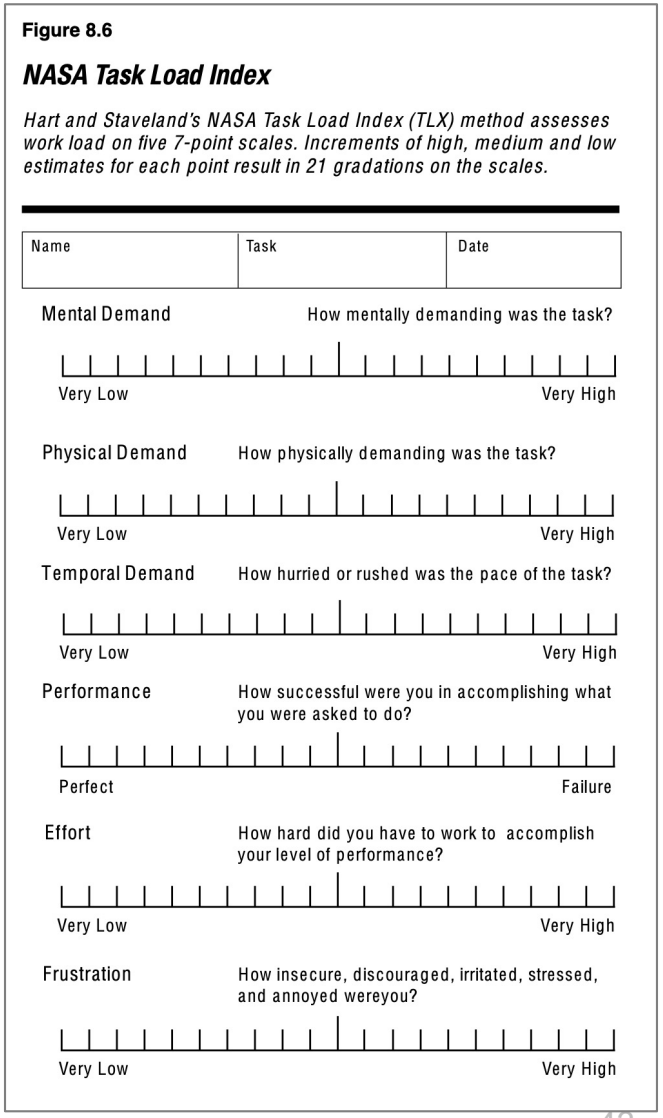


Figure 8.6

NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand — How mentally demanding was the task?

Very Low — Very High

Physical Demand — How physically demanding was the task?

Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?

Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?

Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low — Very High

Your own questionnaire

If you want, you can make your own questionnaire

Tip:

- Use Likert statements (“Totally disagree – Totally agree”)
- Ask about the same topic using multiple prompts
- Do a pilot study

How to visualize questionnaire data

Use bar charts to visualize answers

1. Calculate user-level averages:

   Example: Average of all NASA-TLX answers from a user to Design A, and another average from answers to Design B

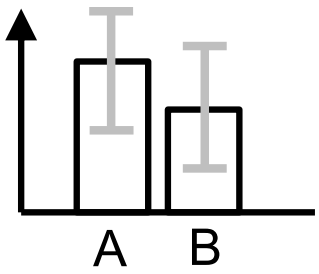
2. Calculate averages across all the users

   Example: Average of all the user-level averages from step 1 for Design A, and the similar average for Design B

3. Present the two bars in a diagram

4. Draw confidence intervals

   If you have time – See YouTube tutorials



Group task: “Methods shopping”

Review the slides in the group

Pick measures that you think you would like to have

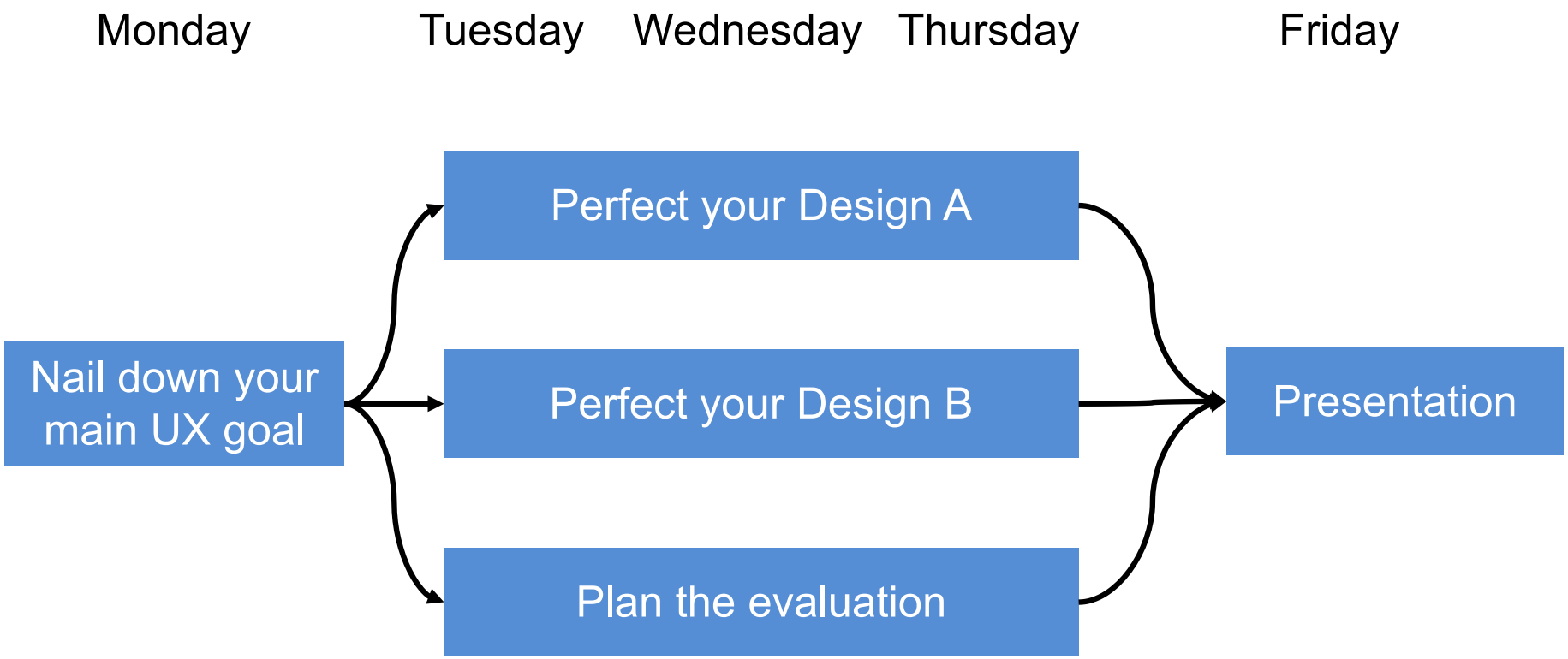
Discuss if this is possible

Discuss why the measure would be useful

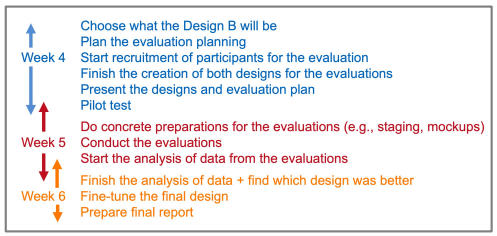
- Does it relate to your UX goal(s)?
- Is it important because of another reason?

How to “survive” this week

Divide yourselves into sub-teams

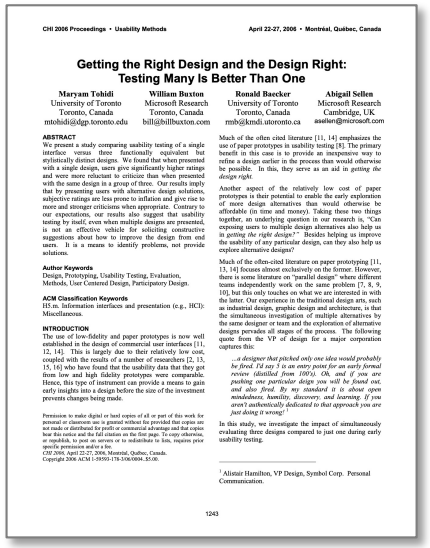


Remember that weeks 4, 5 and 6 also have fuzzy boundaries



Reading materials

Reading materials for week 4



CHI 2006 Proceedings • Usability Methods April 22-27, 2006 • Montréal, Québec, Canada

Getting the Right Design and the Design Right: Testing Many Is Better Than One

Maryam Tohidi, University of Toronto, Toronto, Canada, mtohidi@dgp.toronto.edu
William Buxton, Microsoft Research, Toronto, Canada, bill@billbuxton.com
Ronald Baecker, University of Toronto, Toronto, Canada, rmb@kmdi.utoronto.ca
Abigail Sellen, Microsoft Research, Cambridge, UK, asellen@microsoft.com

ABSTRACT
We present a study comparing usability testing of a single interface versus three functionally equivalent but stylistically-distinct designs. We found that when presented with a single design, users give significantly higher ratings and were more reluctant to criticize than when presented with the same design in a group of three. Our results imply that by presenting users with alternative design solutions, subjective ratings are less prone to inflation and give rise to more and stronger criticisms when appropriate. Contrary to our expectations, our results also suggest that usability testing by itself, even when multiple designs are presented, is not an effective vehicle for soliciting constructive suggestions about how to improve the design from end users. It is a means to identify problems, not provide solutions.

Author Keywords
Design, Prototyping, Usability Testing, Evaluation, Methods, User Centered Design, Participatory Design.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
The use of low-fidelity and paper prototypes is now well established in the design of commercial user interfaces [11, 12, 14]. This is largely due to their relatively low cost, coupled with the results of a number of researchers [2, 13, 15, 16] who have found that the usability data that they get from low and high fidelity prototypes were comparable. Hence, this type of instrument can provide a means to gain early insights into a design before the size of the investment prevents changes being made.

CHI 2006, April 22-27, 2006, Montréal, Québec, Canada.

Much of the often cited literature [11, 14] emphasizes the use of paper prototypes in usability testing [8]. The primary benefit in this case is to provide an inexpensive way to refine a design earlier in the process than would otherwise be possible. In this, they serve as an aid in *getting the design right*.

Another aspect of the relatively low cost of paper prototypes is their potential to enable the early exploration of more design alternatives than would otherwise be affordable (in time and money). Taking these two things together, an underlying question in our research is, "Can exposing users to multiple design alternatives also help us in *getting the right design*?" Besides helping us improve the usability of any particular design, can they also help us explore alternative designs?

Much of the often-cited literature on paper prototyping [11, 13, 14] focuses almost exclusively on the former. However, there is some literature on "parallel design" where different teams independently work on the same problem [7, 8, 9, 10], but this only touches on what we are interested in with the latter. Our experience in the traditional design arts, such as industrial design, graphic design and architecture, is that the simultaneous investigation of multiple alternatives by the same designer or team and the exploration of alternative designs pervades all stages of the process. The following quote from the VP of design for a major corporation captures this:

> *...a designer that pitched only one idea would probably be fired. I'd say 5 is an entry point for an early formal review (distilled from 100's). Oh, and if you are pushing one particular deign you will be found out, and also fired. By my standard it is about open mindedness, humility, discovery, and learning. If you aren't authentically dedicated to that approach you are just doing it wrong!* [1]

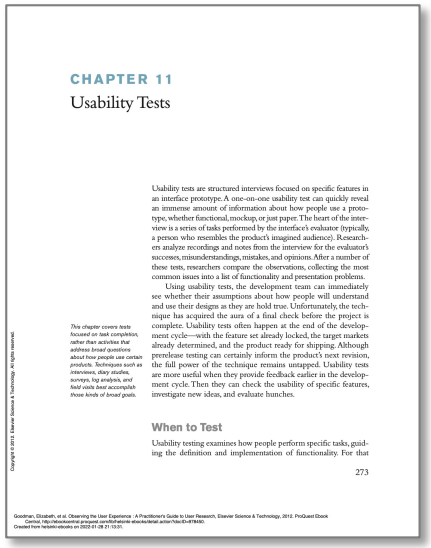
In this study, we investigate the impact of simultaneously evaluating three designs compared to just one during early usability testing.

[1] Alistair Hamilton, VP Design, Symbol Corp. Personal Communication.

Tohidi et al (CHI2006):

Getting the right design and the design right: Testing many is better than one

https://dl-acm-org.libproxy.aalto.fi/doi/10.1145/1124772.1124960



CHAPTER 11

Usability Tests

This chapter covers tests focused on task completion, rather than activities that address broad questions about how people use certain products. Techniques such as interviews, diary studies, surveys, log analysis, and field visits best accomplish those kinds of broad goals.

Usability tests are structured interviews focused on specific features in an interface prototype. A one-on-one usability test can quickly reveal an immense amount of information about how people use a prototype, whether functional, mockup, or just paper. The heart of the interview is a series of tasks performed by the interface's evaluator (typically, a person who resembles the product's imagined audience). Researchers analyze recordings and notes from the interview for the evaluator's successes, misunderstandings, mistakes, and opinions. After a number of these tests, researchers compare the observations, collecting the most common issues into a list of functionality and presentation problems.

Using usability tests, the development team can immediately see whether their assumptions about how people will understand and use their designs as they are hold true. Unfortunately, the technique has acquired the aura of a final check before the project is complete. Usability tests often happen at the end of the development cycle—with the feature set already locked, the target markets already determined, and the product ready for shipping. Although prerelease testing can certainly inform the product's next revision, the full power of the technique remains untapped. Usability tests are more useful when they provide feedback earlier in the development cycle. Then they can check the usability of specific features, investigate new ideas, and evaluate hunches.

When to Test

Usability testing examines how people perform specific tasks, guiding the definition and implementation of functionality. For that

Goodman & Kuniavsky (2012):

Chapter 11: Usability tests

https://pdfroom.com/books/observing-the-user-experience-second-edition-a-practitioners-guide-to-user-research/wW5mwke4gYo

or

https://primo.aalto.fi/permalink/358AALTO_INST/ha1cg5/alma9985689444064526

Tutor meetings

https://doodle.com/poll/xvr4fgimhs6w8m9f?utm_source=poll&utm_medium=link