Aalto University
School of Arts, Design
and Architecture

Evaluation planning

MUO-E3055 Interaction Design (IxD)

5 February 2024

Antti Salovaara

MyCourses > Slides > 05a Evaluation planning.pdf

Contents of today's teaching

Overview on the next tasks

Tutor meeting time on Tuesday + how to prepare for it

User evaluation planning

- Different types of evaluations
- Triangulation + Selection of data collection methods
- User recruitment
- Questionnaire-based methods
- Practical arrangements + tips

Research ethics

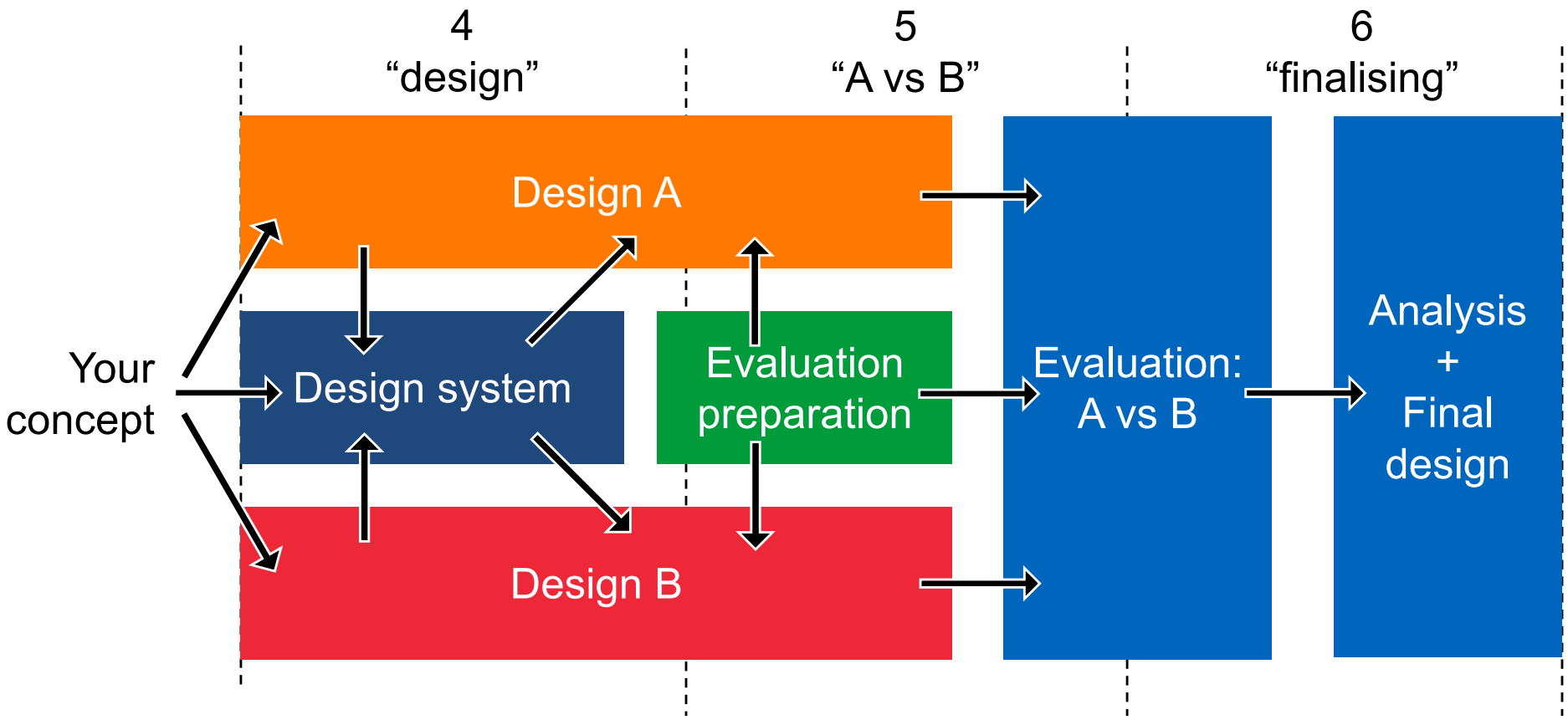
Friday's program (heuristic evaluations)

Reading materials + quiz for Friday

Overview of next tasks

Week 5’s main tasks

Requirements for the final presentation in week 6



Main tasks this week:

- Integrating your A and B wireflows and your design system
- Selecting the research methods and recruiting 5 participants
- Creating interactable A and B prototypes for evaluations
- Making other preparations for the evaluations (propping, actors, …)
- Starting the evaluations (and continuing them on week 6)

Possible schedule for this week

Monday:

- Prepare a research design diagram (see later slide)
- Recruit users (5)

Tuesday:

- Tutor meetings
- Stitch the prototypes for Designs A and B
- Prepare mockup materials + do other practical preparations

Wednesday:

- Carry out pilot test within your group (or by getting a user from another group)

Thursday:

- Evaluate

Friday:

- Evaluate
- Lecture:
  - discussion on reading material
  - no presentations
  - heuristic evaluations
  - getting prepared for data analysis

Week 6: finalise the evaluations; analyse the data; prepare final design illustration

Requirements for week 6's final design

The final design is just an illustration!

- It is non-functional but realistic-looking
- Screenshots or pictures are enough
- "What it would look like if we would really build it"

Its design is based on your findings

- From your learnings during the creation A+B's interactive versions
- From week 5's Friday's heuristic evaluations
- From user evaluations

Final presentation

15 mins presentation + 10 mins discussion

1 slide from each point, please!

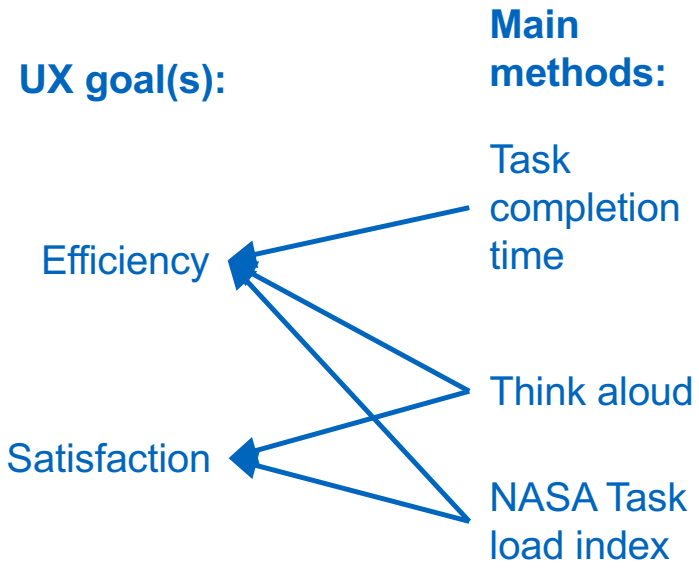
1. Title slide
2. Pictures of Designs A and B
   So that the final design can be compared to them
3. Evaluation's research design diagram
   UX goal(s) ⟷ Methods that you used to measure the goals
4. Technologies that you used to build the prototypes
   How you operated the prototype (Wizard of Oz? Programming? Lots of Figma screens? …)
5. Evaluation setup
   Tasks that users performed, how data was gathered
6. Picture from the entire affinity wall
7. 2–3 main findings from the analysis
8. Final design, based on the findings
   Presented in a manner that is comparable to Designs A and B in the 2nd slide
9. Lessons learned for other teams (2 items)
   Dos and donts, methodological recommendations, …

Tutor meeting times for Tuesday

Times

How to prepare for the meeting

In the tutor meeting:
Present your evaluation plan template



Meeting times:

9:30, 10:15, 11:00,
12:30, 13:15, 14:00

Location: P210

1. Research design diagram
   - Main methods
   - ”Sub-methods” such as interview questions, sub-tasks within A and B tests, repetitions, …
2. Recruitment plan
   - Who will be your participants?
   - When will you run the evaluations?
3. Arrangements for designs A and B
   - Prototyping technology
   - Where you will run the study, where you will interview?
   - Division of work: facilitator, wizard, video-taker, …
   - Data collection (video, note-taking, …)

Planning the evaluations

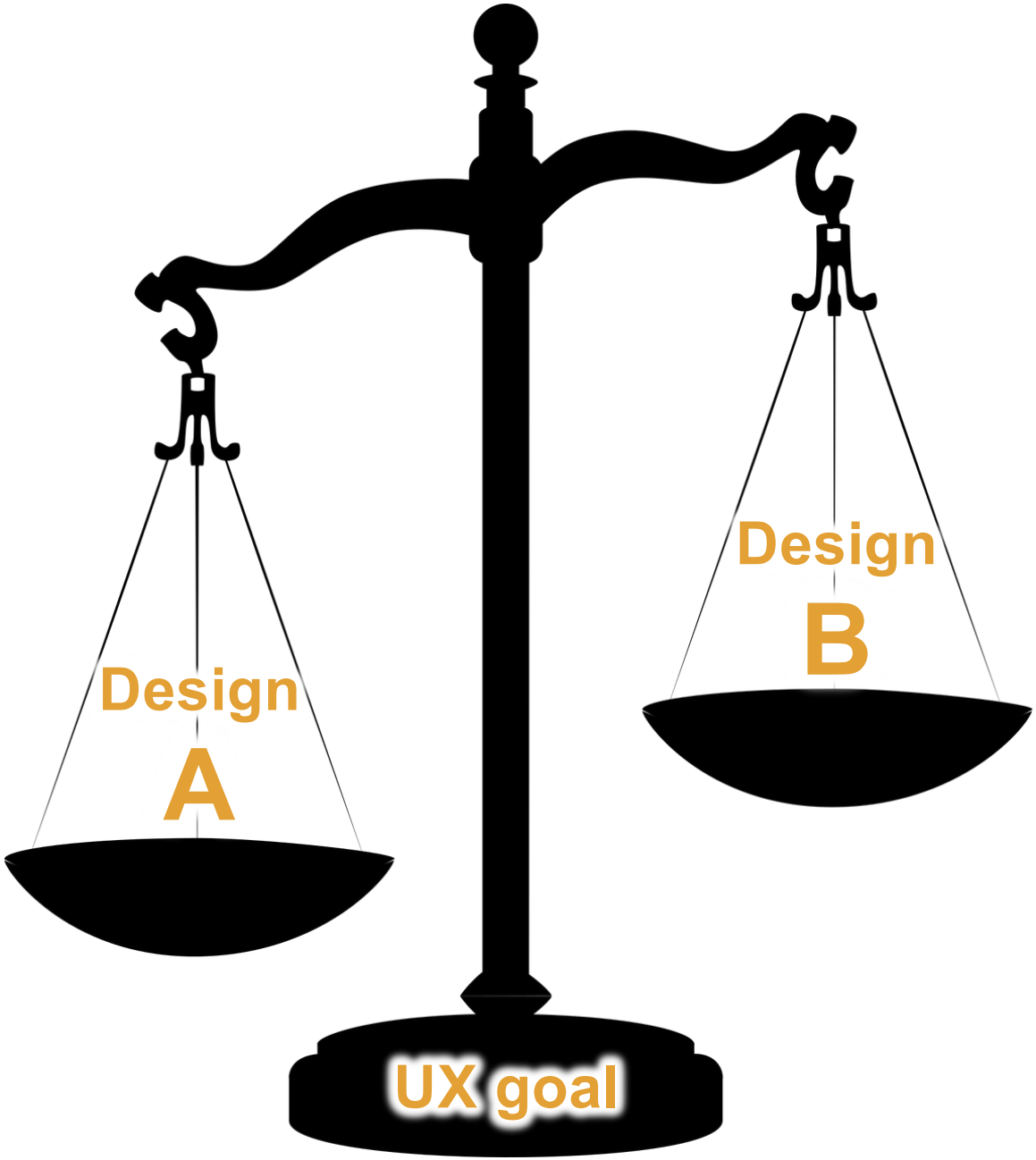
Different types of evaluation

Triangulation + Selection of data collection methods

User recruitment

Questionnaire-based methods

Practical arrangements + tips



What determines which design has more weight?

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

Group exercise (15 mins)

Consider your most important UX goal

- E.g., ease of use

Discuss: What kinds of data could let you measure and learn about this UX goal in your prototypes?

- User's stress level?
- Number of errors at the first try on the task?
- Asking the participant how easy the interaction was?

Steps:

1. Brainstorm individually (5 mins)
2. Then share ideas within your group (10 mins)

Break (~5 mins, until 10:10)

Different types of evaluation

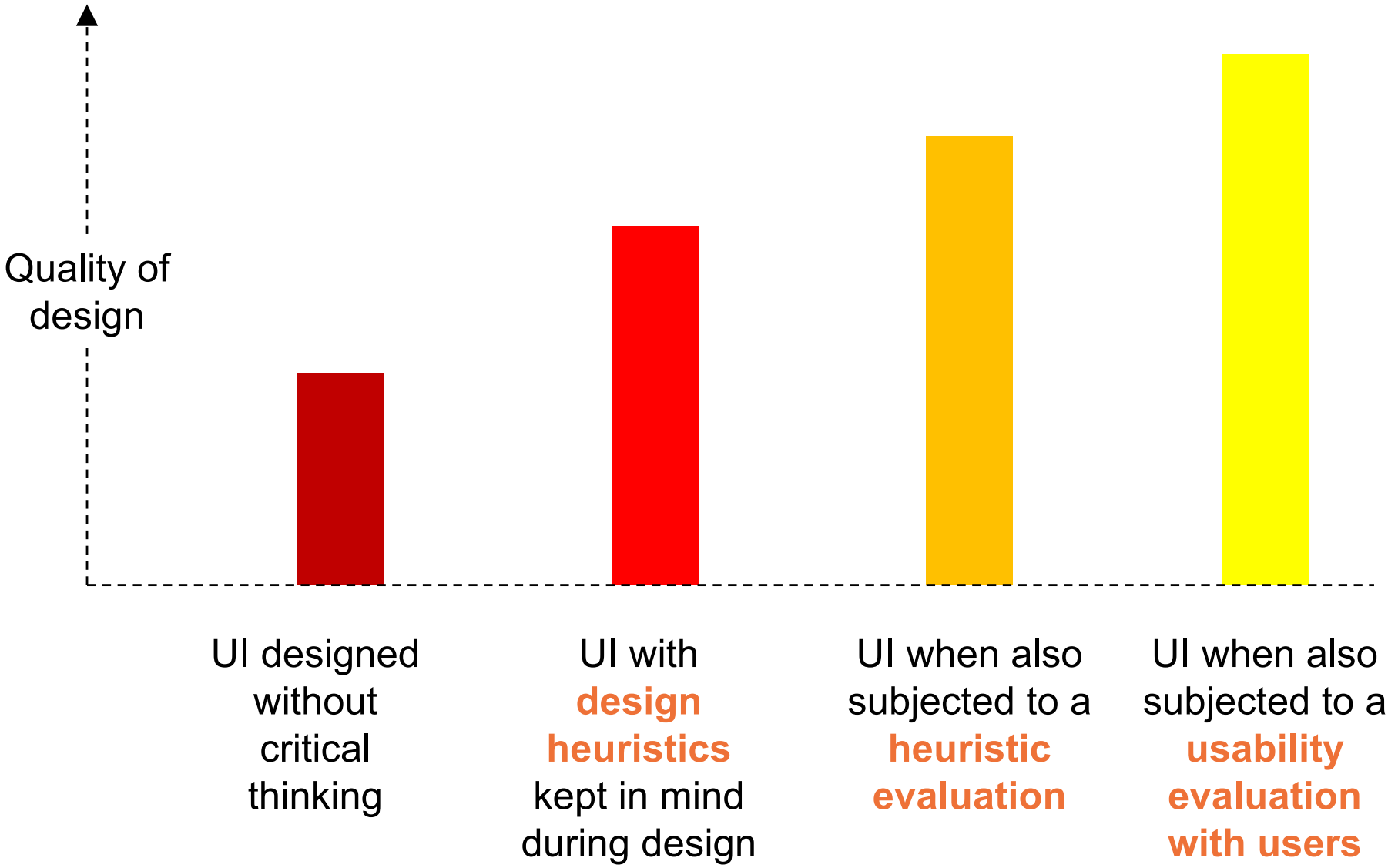
Heuristic evaluations: evaluation without users

Traditional scenario-driven usability evaluations

In-the-wild studies (field trials)

Wizard of Oz evaluations

How to reach a good usability and UX

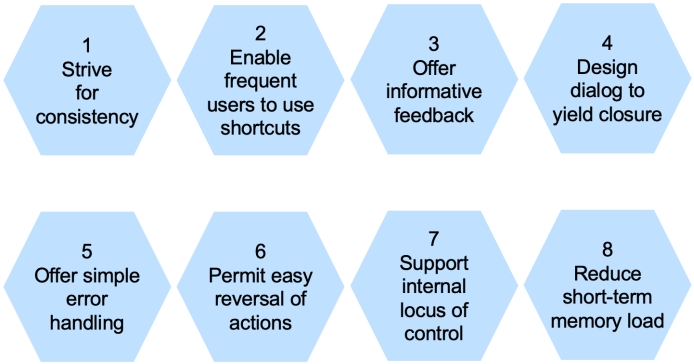


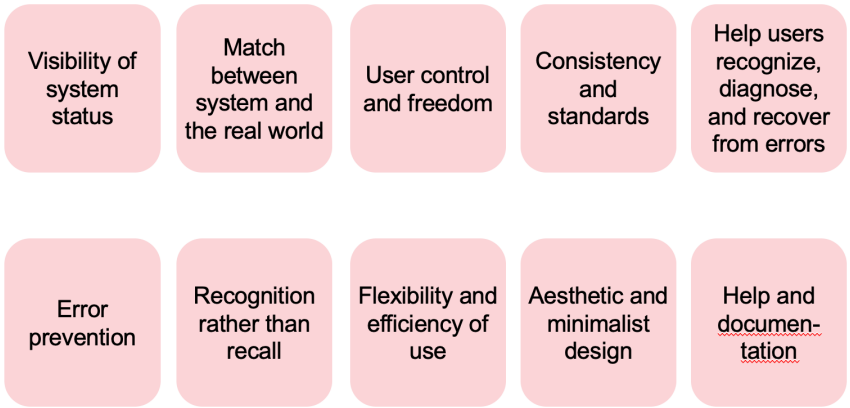
Heuristic evaluation

UI's analysis using design heuristics

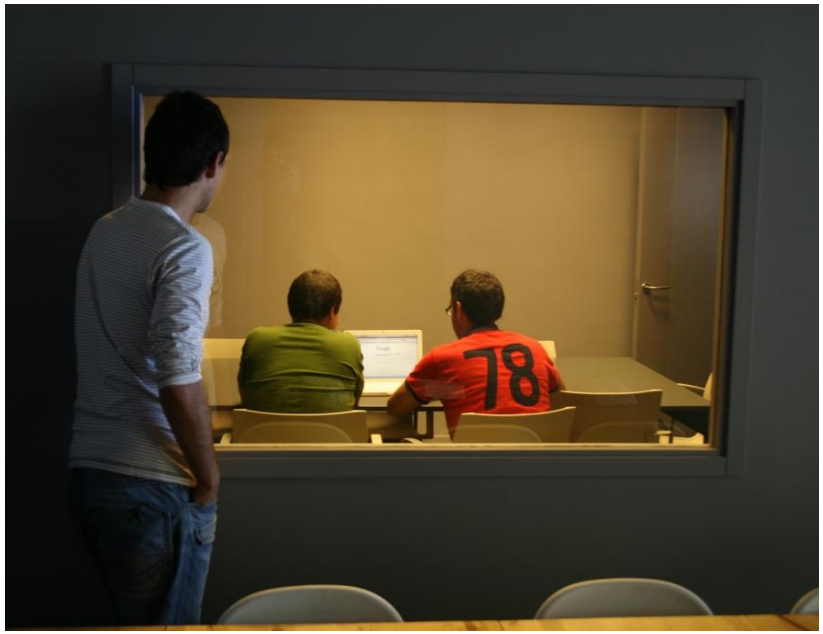
Performed by an expert, without users

- 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!





Traditional usability evaluation



Preparations:

1. Write realistic task scenarios for the features that need evaluation
2. Create mockup materials + propping to create a believable UX with the prototype

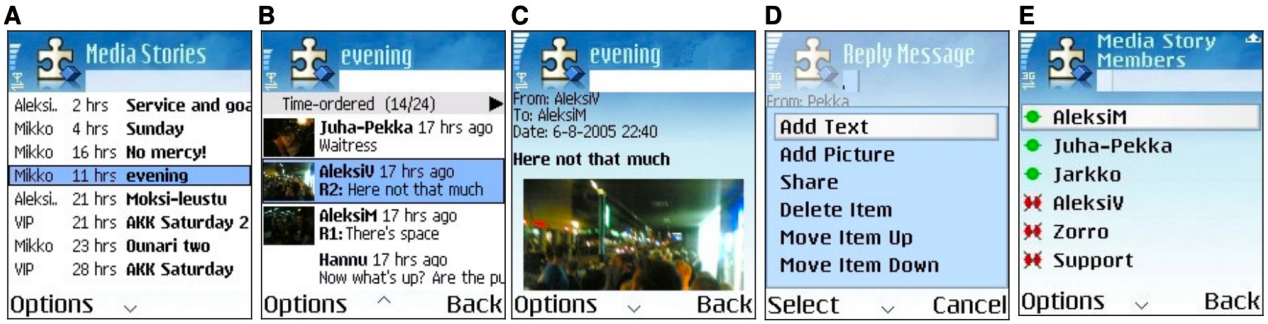
The actual test with a participant:

1. Present the scenario and ask them to carry out the tasks.
2. Record with video

Repeat with more participants until findings “saturate”

In-the-wild study (field trial)

Example: Study of a Whatsapp-like mobile app in 2005 *



Msg 4
AleksiV Sat 17:03:58
Hey gents how is your day going? Any luck with the ladies last night? You guys heading out for a big night again? DaMo.

Msg 5
AleksiV Sat 17:18:45

The jump

Msg 6 (Reply to **4**)
Eljas Sat 17:26:50
Hi guys! Sat night plan -> First to Killeri, there is a rallycross competition from 19.30 to 22.00. After that "little" bit drinking and then to pub. How about you?

A
Msg 13 Hannu Fri 00:15:53
The boys keep it going, but we have...

(Shows two mattresses in a living room, a sleeping bag and a pillow)

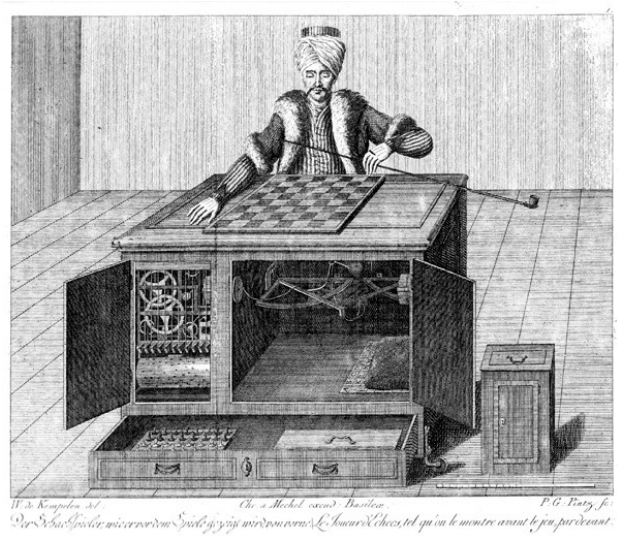
B

C

D

* Salovaara et al. (CHI2006): Collective creation and sense-making of mobile media.

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

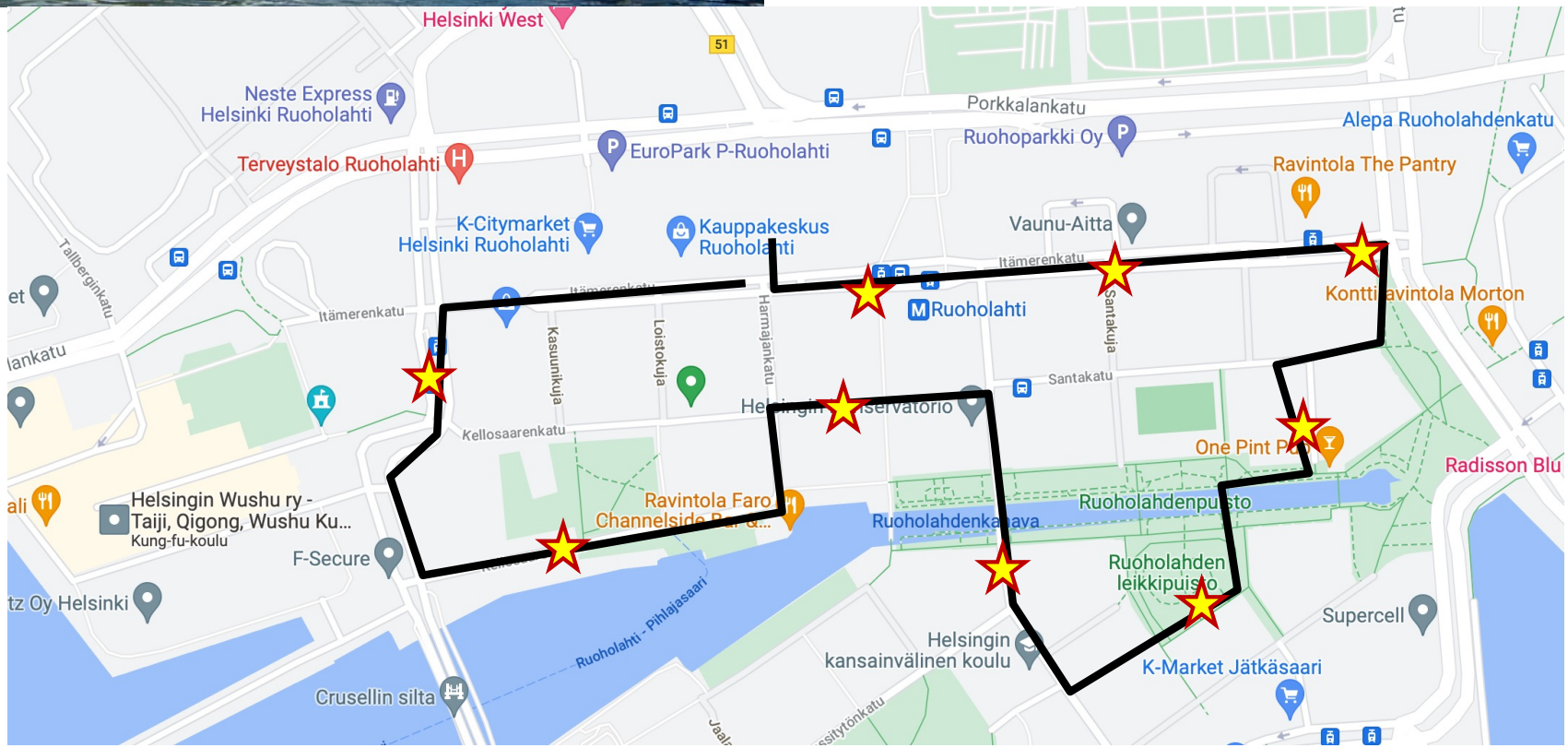
Ethics issue:

Setup is revealed after the study

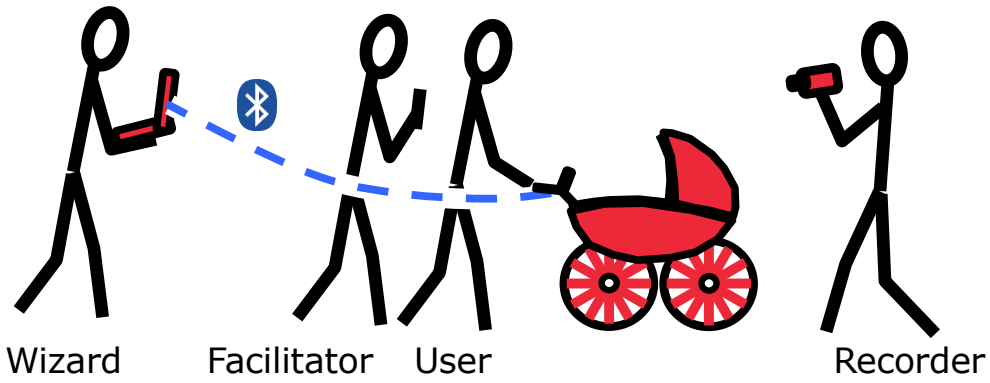
Example: In-the-wild Wizard-of-Oz study

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

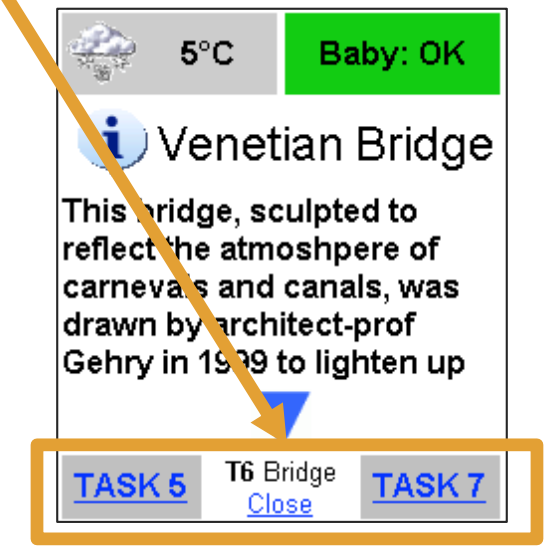
The Ruoholahti canal by Oghmoir.
http://commons.wikimedia.org/wiki/File:Ruoholahden_kanava.jpg. Licensed under Creative Commons Attribution-Share Alike 3.0 Unported



Wizard-of-Oz setup



Wizard's controls



How to control a prototype remotely in a WoZ study

One possibility is Protopie

See my very rapidly created demo video on how to do that with Protopie:

https://aalto.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=e35fc051-249f-458a-907c-af9d00745109

The video explains how to create a very simple Augmented Reality mobile wayfinding app: a phone that shows view from a camera, and a triangle-shaped turning sign on top of it. The turns of the triangle can be controlled from another phone.

Think aloud method

Encourage the users to talk aloud:

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

!!! Thinking aloud is not natural to many people

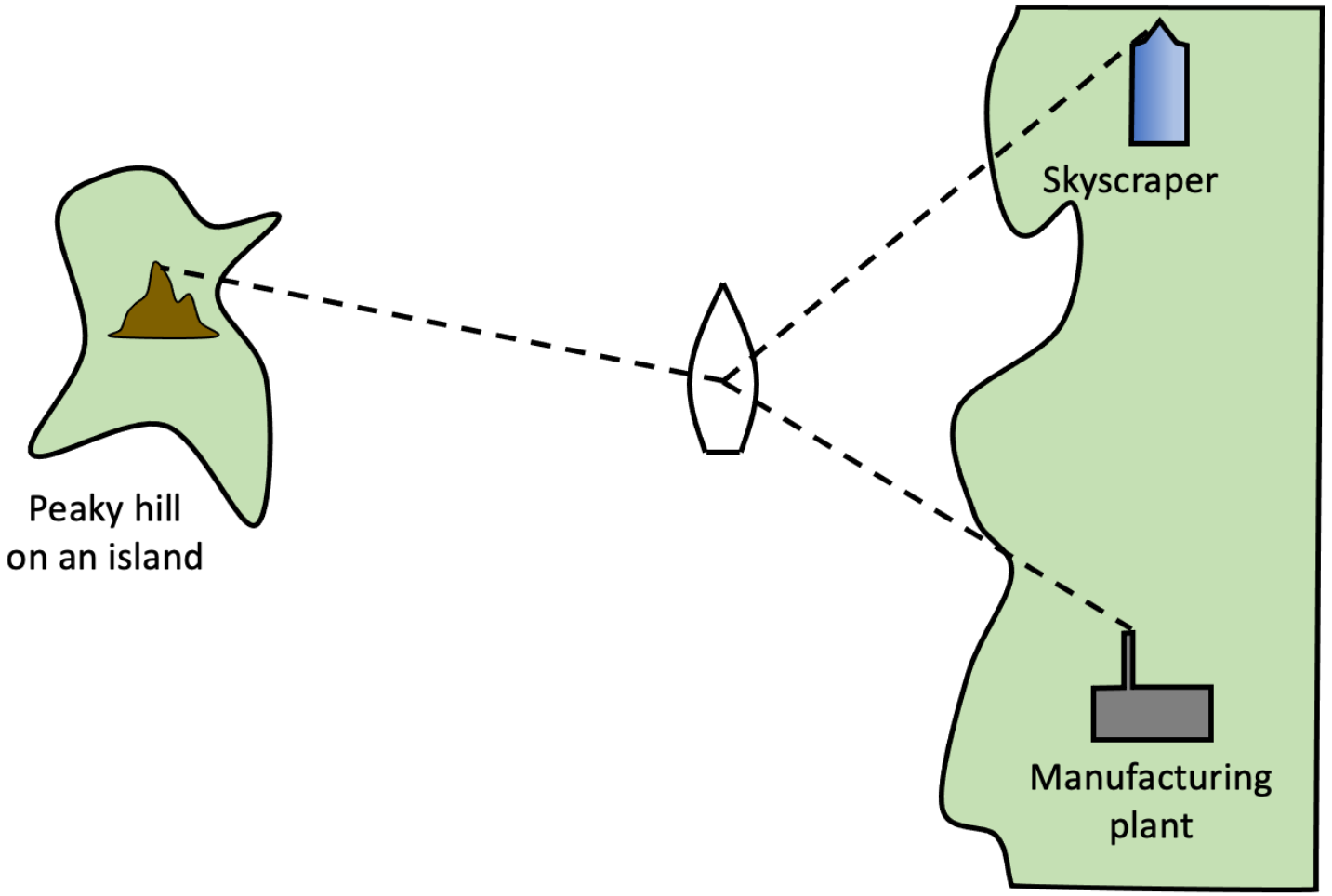
- A demonstration by the moderator and a practice task are needed to give the user an idea on what is expected
- Remember to remind the user politely (“Can you tell what you are now thinking?”)

The method’s origins are in psychological research on problem-solving and creativity*

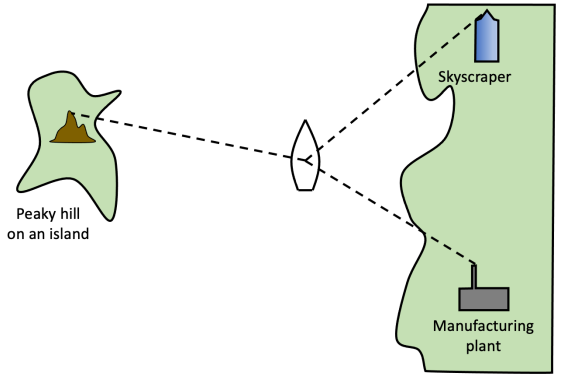
* E.g., Ericsson, K. A. (2006). Protocol analysis and expert thought: concurrent verbalizations of thinking during experts' performance on representatve task. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), Cambridge Handbook of Expertise and Expert Performance, ch. 13 (pp. 223--242).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

Triangulation +
Selection of data collection methods

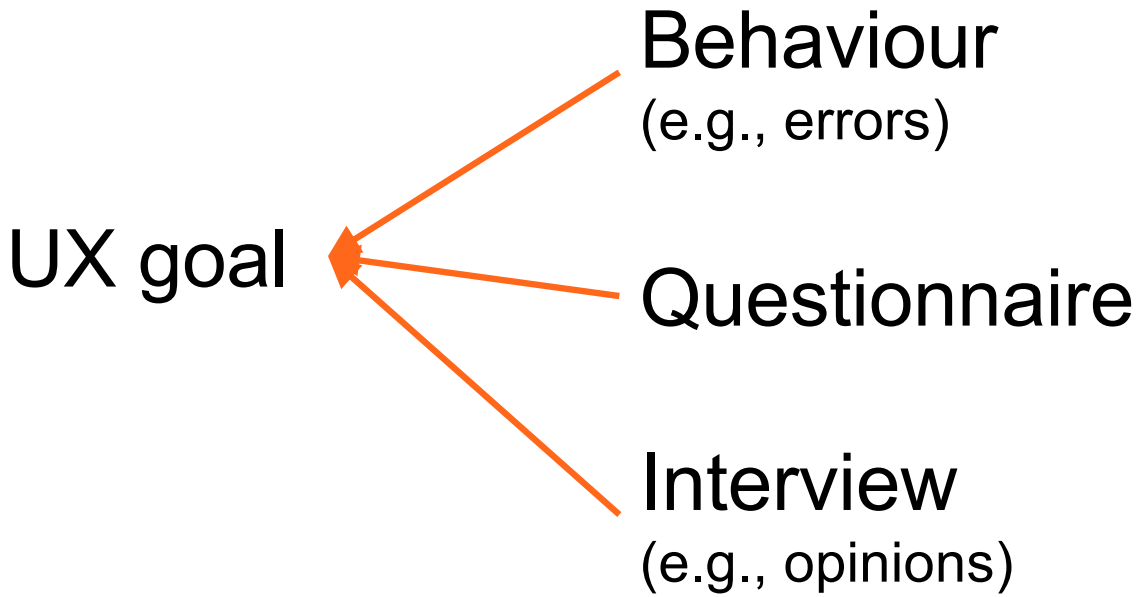
Triangulation



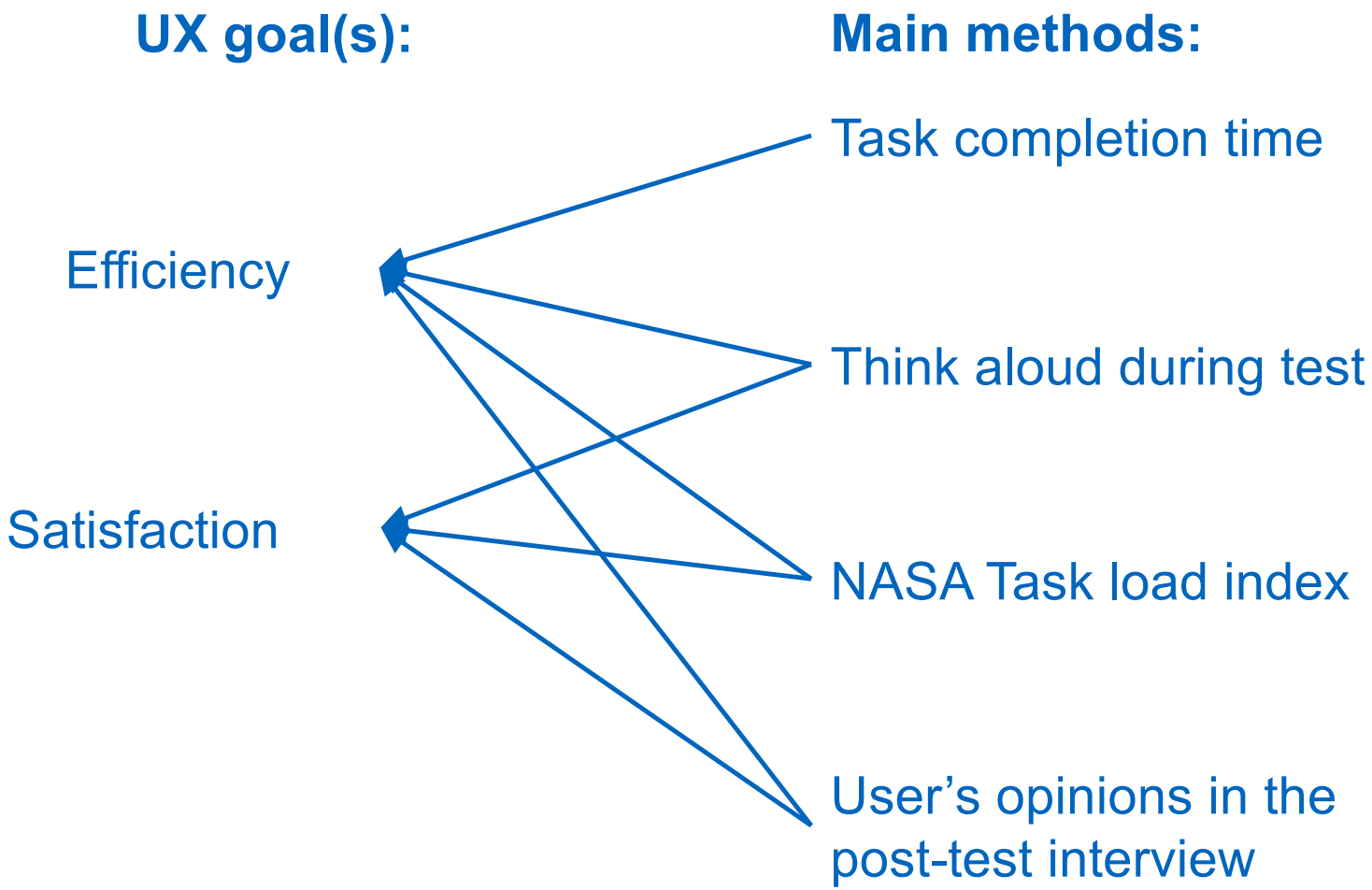
Triangulation



Try to measure the same question in several, complementary ways

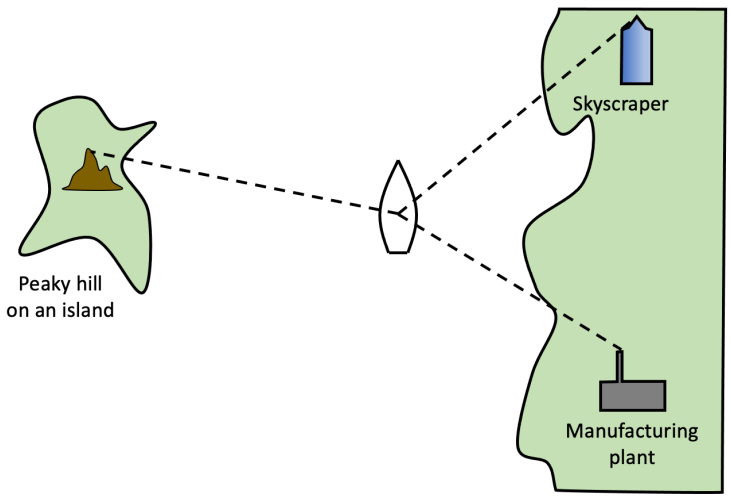


Research design diagram (example)



Group exercise (20 mins)

Break until 11:15



Examples of complementary methods:

Self-report vs action

Quantitative vs qualitative data

Before an IX event vs. after it

Preparation:

Write your UX goal(s) in the middle of a paper

10 mins:

Brainstorm techniques and methods that can “measure” the UX goal

Write them on sticky notes

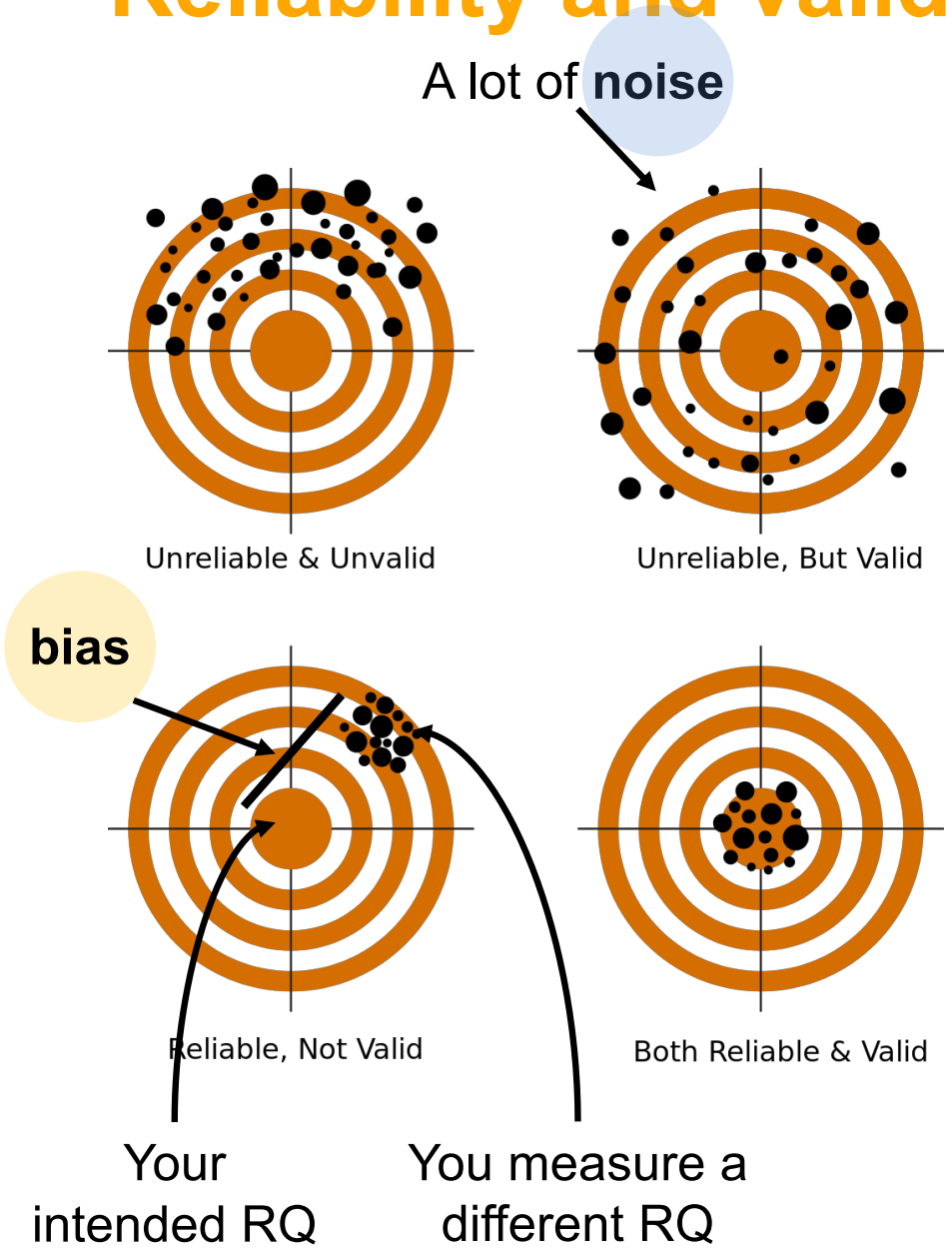
10 mins:

Identify triangulations: place methods around the UX goals, so that complementary methods are in 90 degrees angles

Identify more methods

Break?

Reliability and validity of a method



When the method is **valid…**

… it measures the intended RQ

Validity is destroyed with **bias**

When the method is reliable:

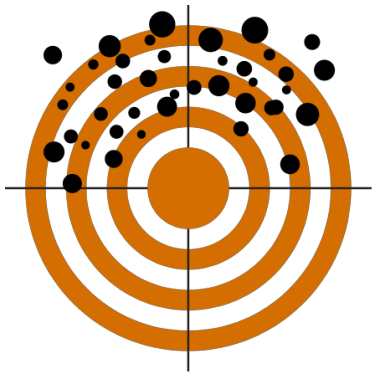
… it is precise and detailed

Reliability is destroyed with **noise** and **lack of data**

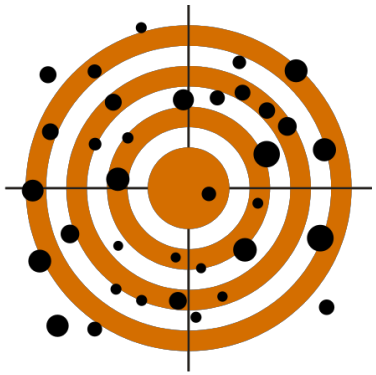
A good method is valid and reliable

https://commons.wikimedia.org/wiki/File:Reliability_and_validity.svg

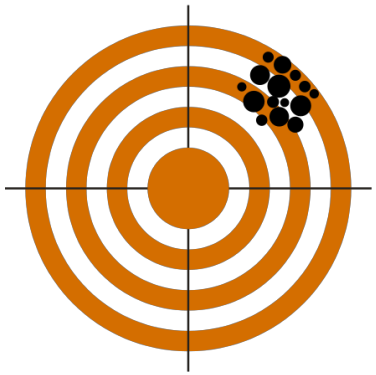
Exercise



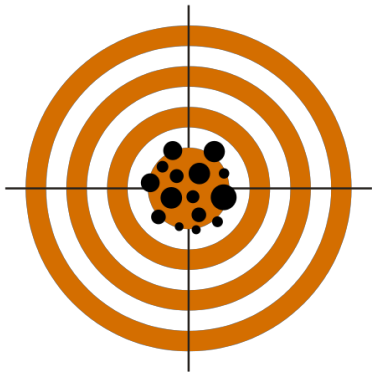
Unreliable & Unvalid



Unreliable, But Valid



Reliable, Not Valid



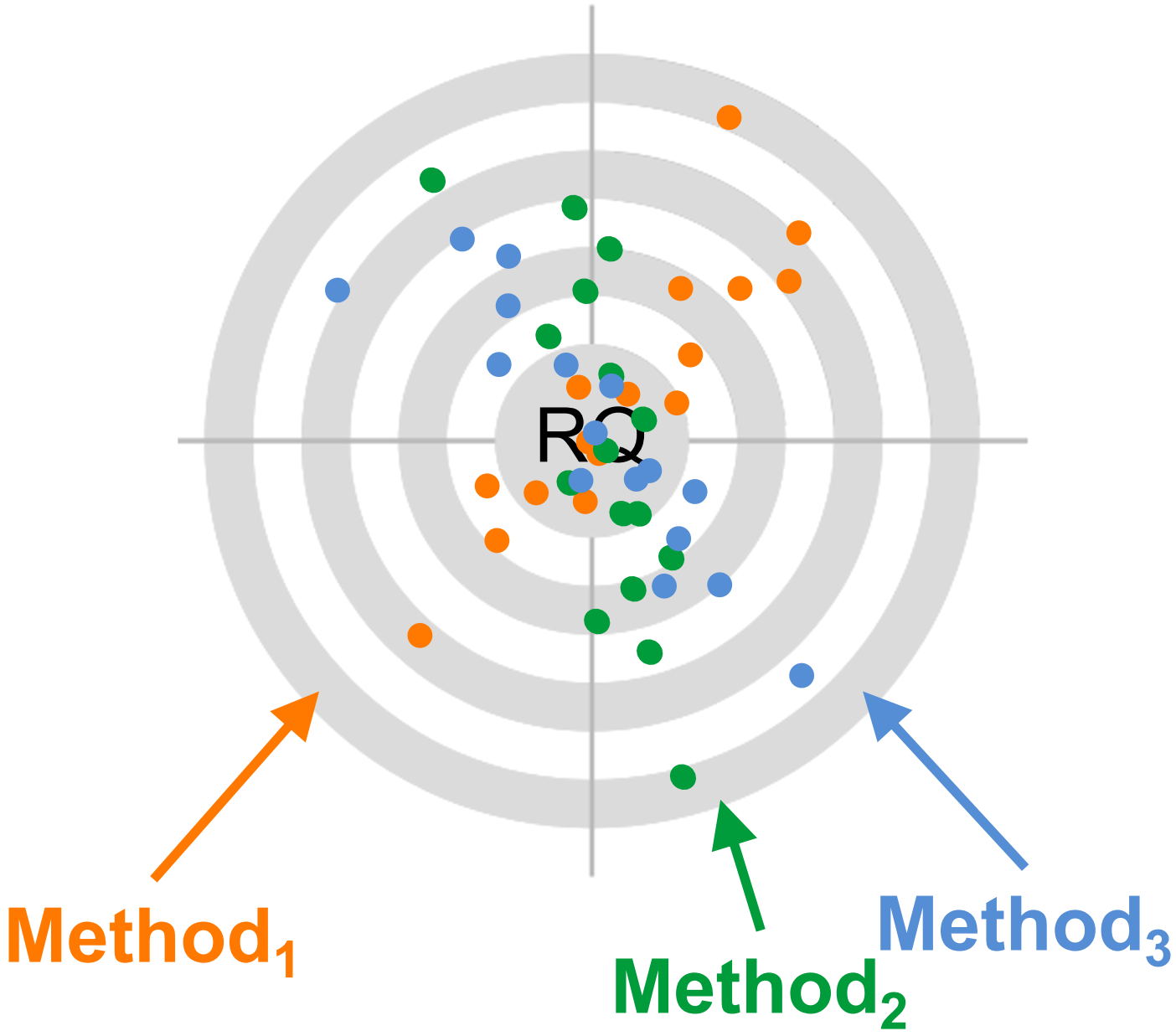
Both Reliable & Valid

How would you categorise the following methods when you are interested in **efficiency?**

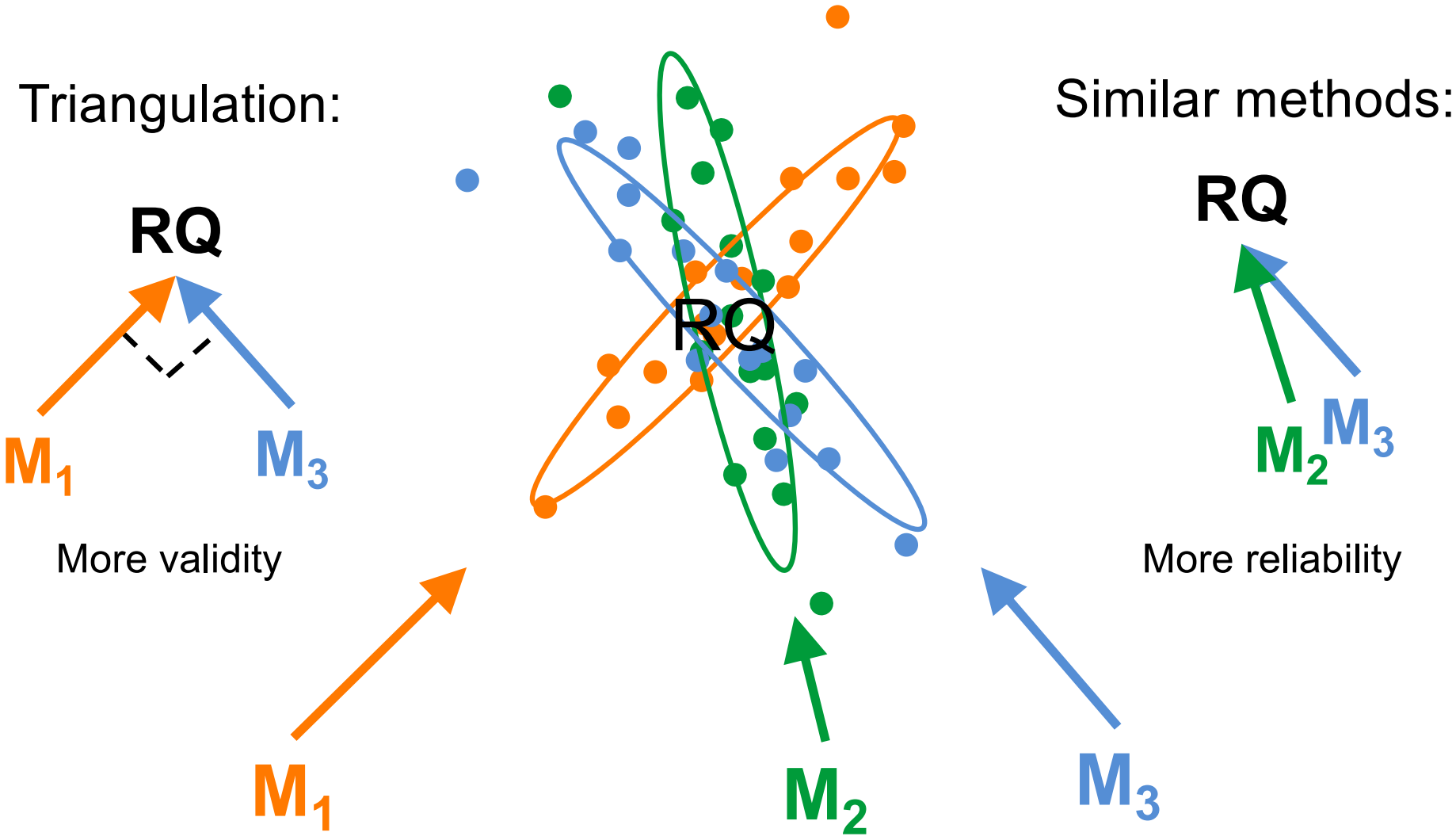
1. Interview question: “Is this app fast to use?”
2. Speed test: User completes a task with app A vs app B
3. Repeated speed test: User completes the task with apps A and B many times
4. Repeated speed test with apps A and B with expert users

https://commons.wikimedia.org/wiki/File:Reliability_and_validity.svg

How to improve validity and reliability

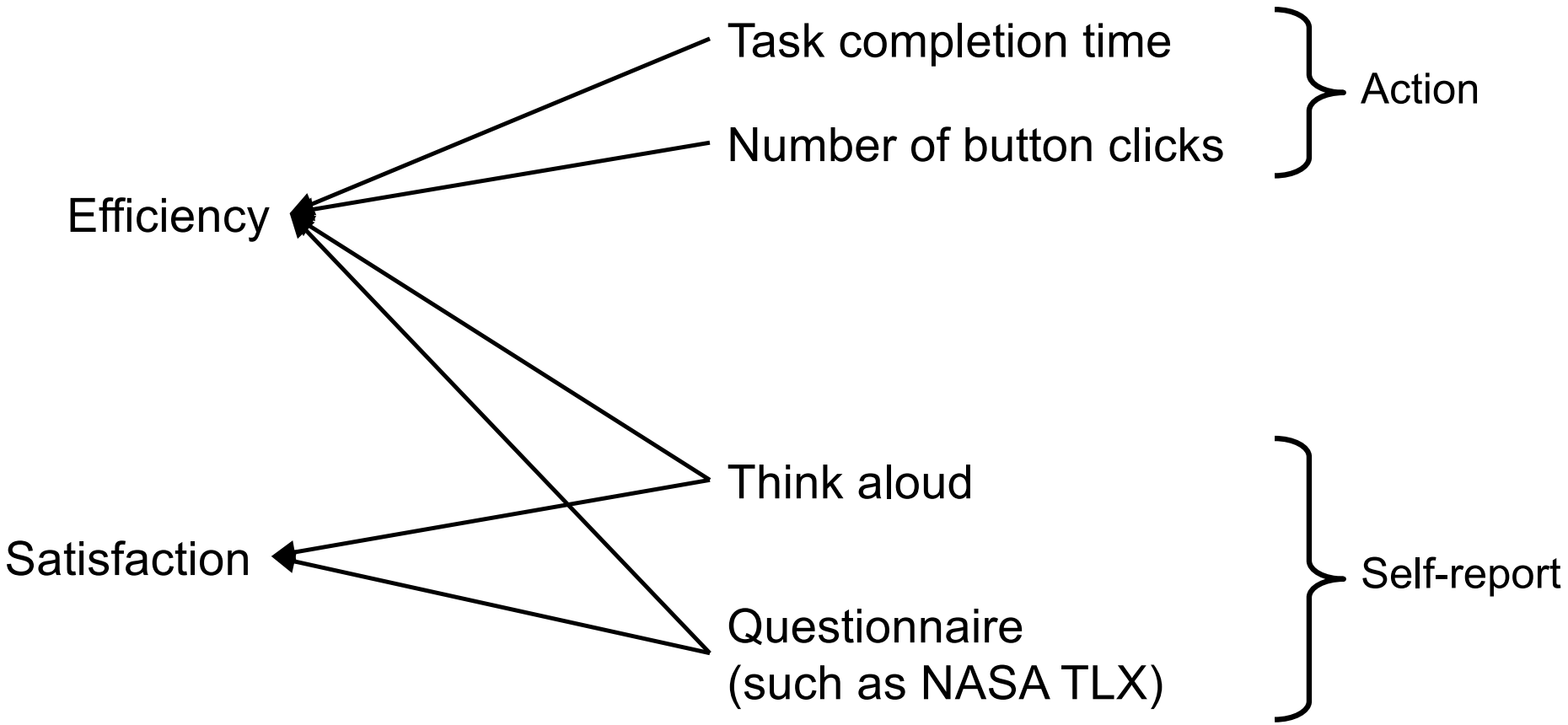


Triangulation and redundancy



Example

What factor differentiates these groups?



Group exercise (until 11:55 or more)

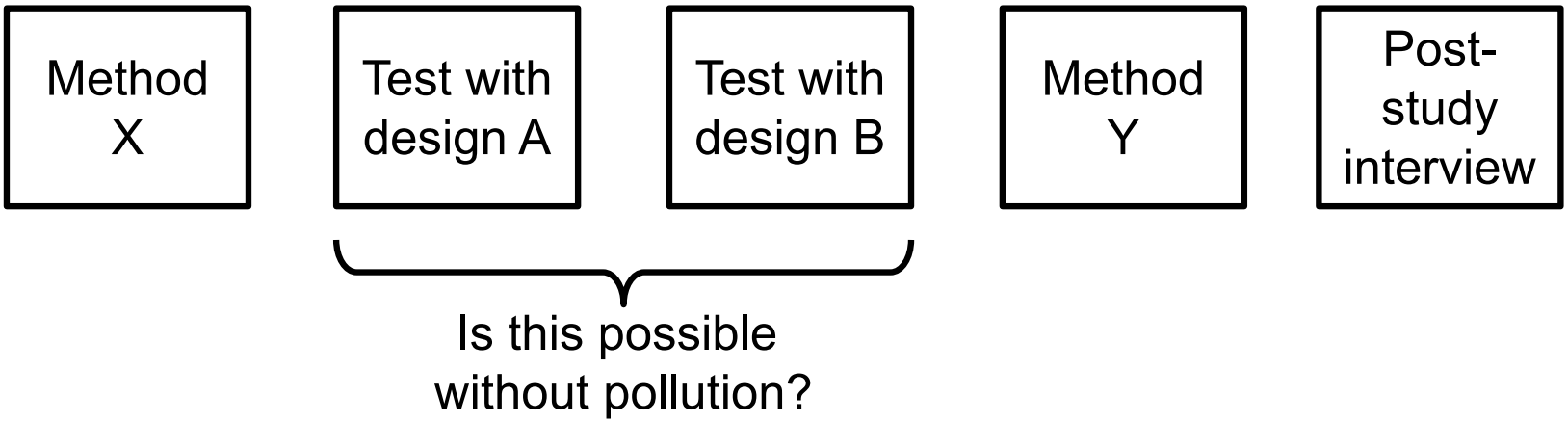
Invent a relatable + believable scena… …luation

- The scenario should enable you… …ethods

Guidelines for a good e…

- Evaluation shou…
- Methods … other: an earlier method 1 shou… …e a later method 2
- Good … …es naturally repeating tasks because that increas… …ndancy and creates more data

Be back 13:20



User recruitment

What users should be recruited?

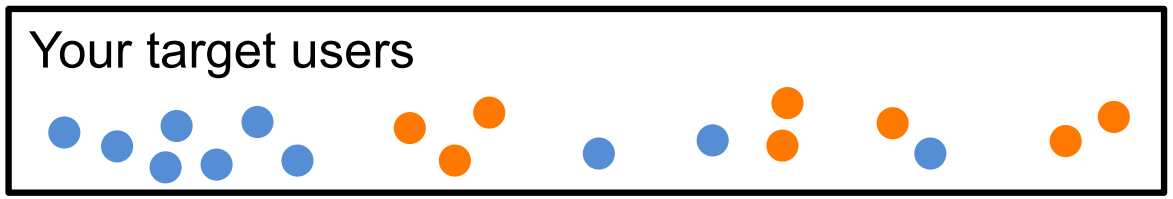
Random sampling

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)

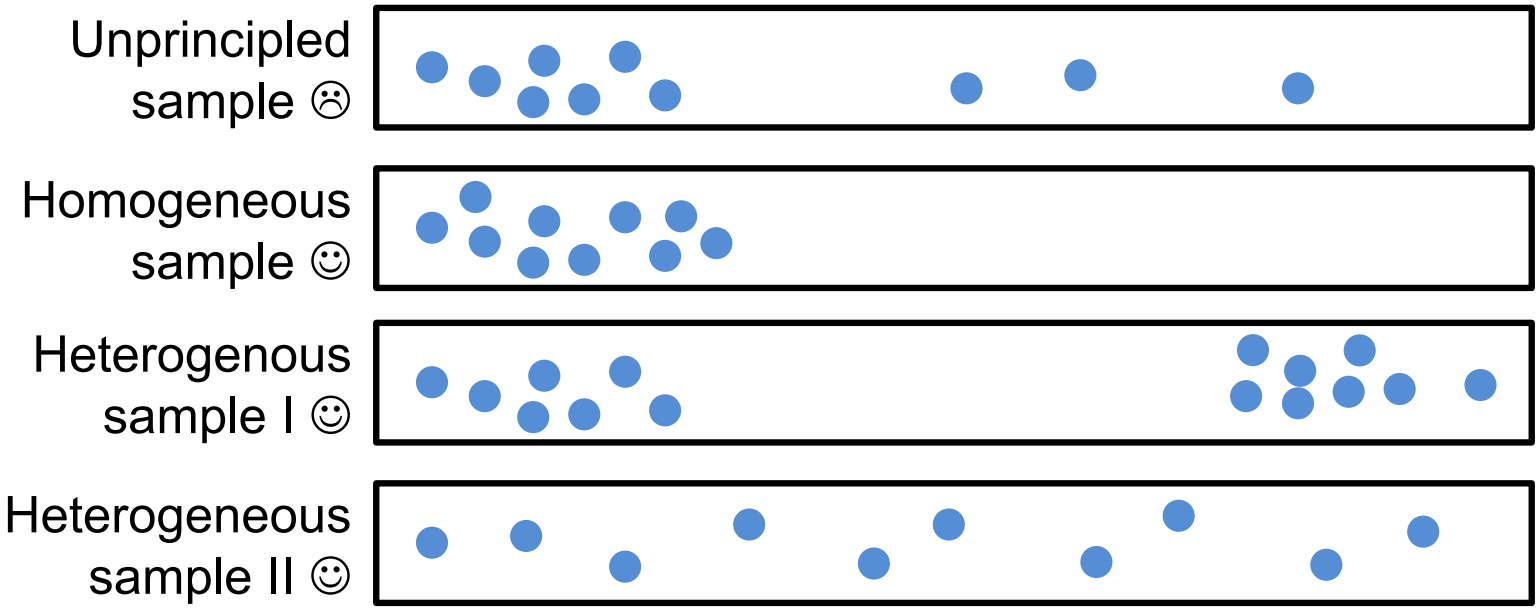


Choosing between heterogeneous vs homogeneous samples

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

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

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

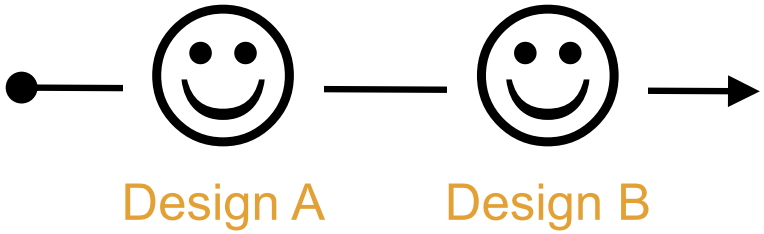
Which designs will each user interact with?

"User takes both medicines"

Participant uses both Design A and Design B

Within-subjects

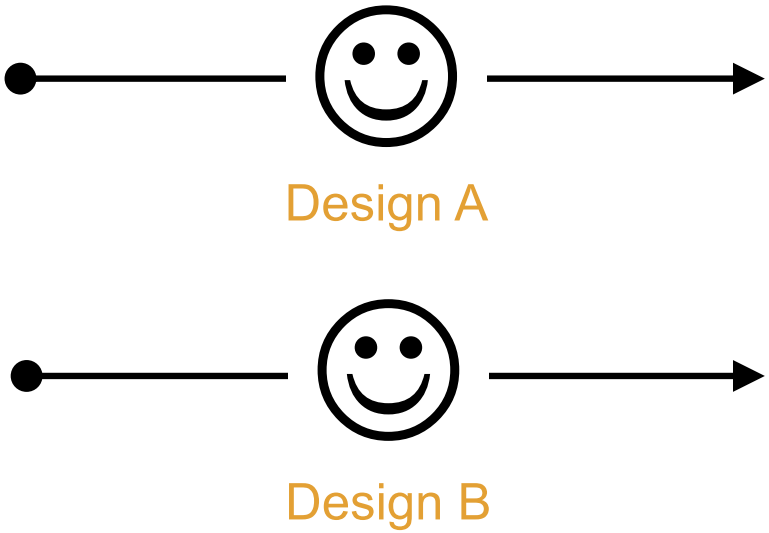
(aka repeated measures)



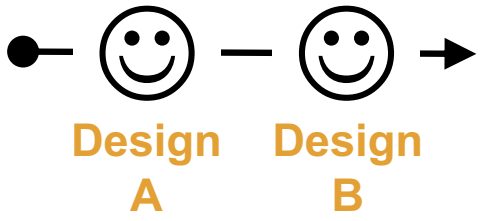
Participant user either Design A or Design B, but not both

Between-subjects

"Different users take different medicines"

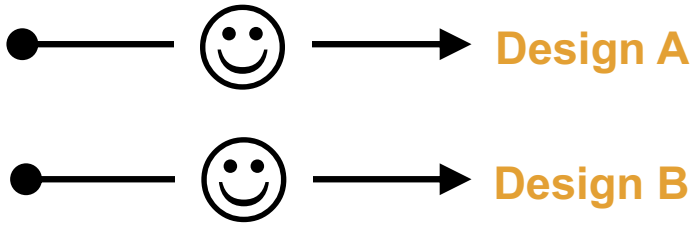


Pros and cons of between- and within-subjects tests



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

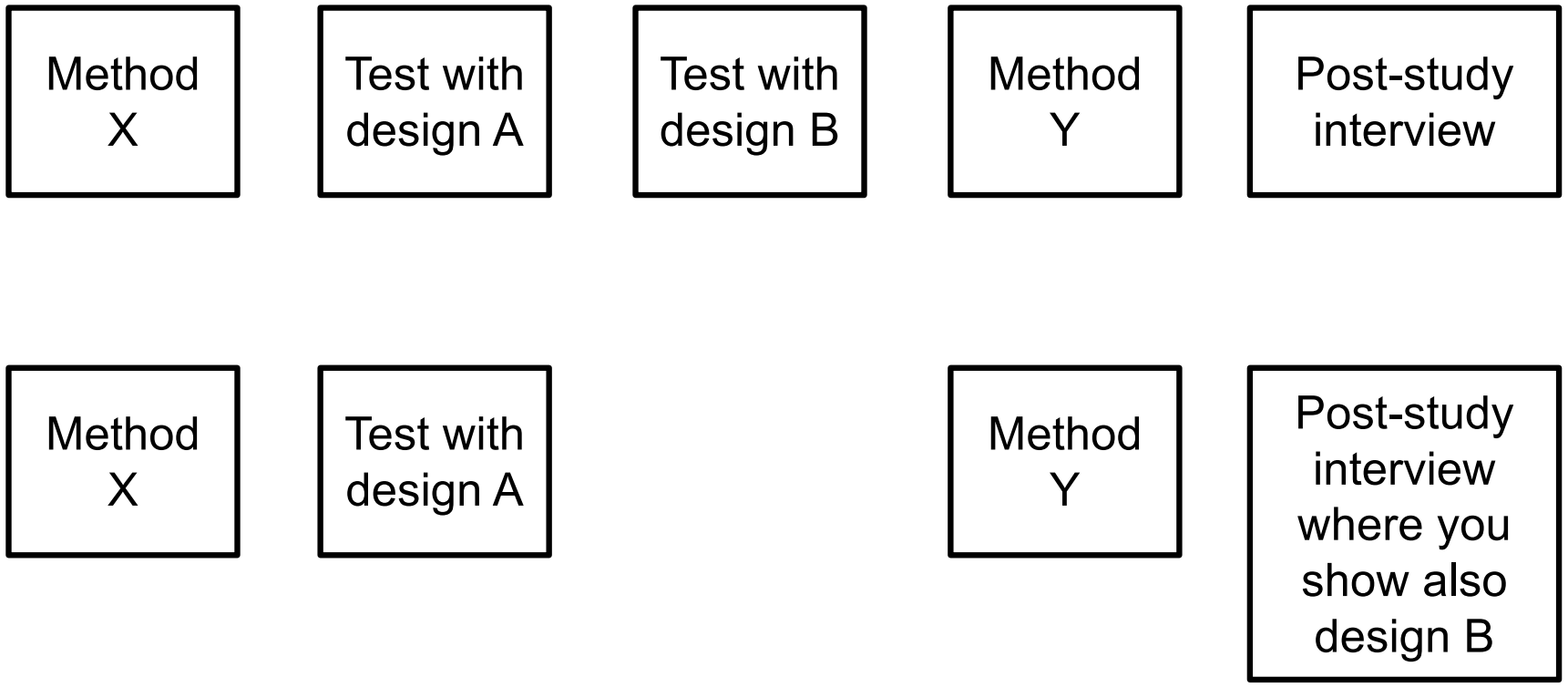


Between-subjects

- \+ No learning effects
- – Need more participants

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

Variants of within-subjects/between-subjects designs



Using same participants again?

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

Pros:

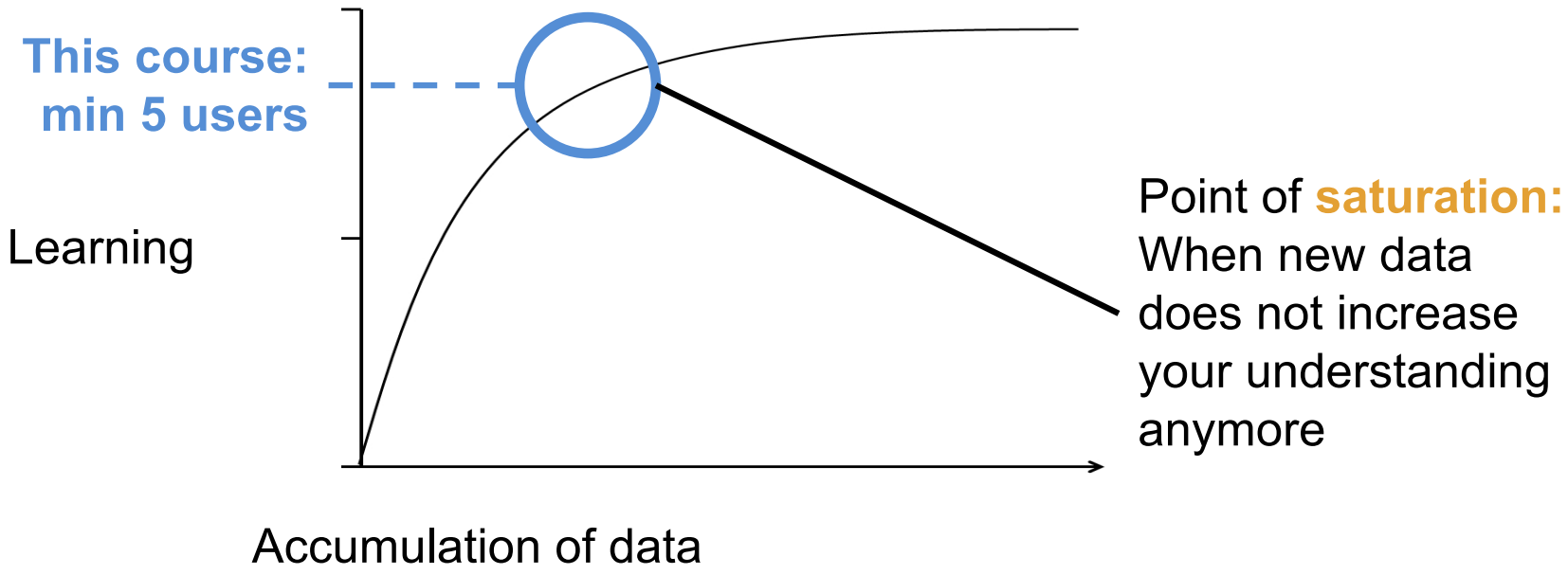
- More detailed feedback
- Easier recruitment
- No need to explain prototype in detail

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's students: Same issues

How much data is enough?



In usability evaluations, data can be 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

In a post-test interview:

- Use questionnaires (SUS, AttrakDiff, NASA TLX, your own questions…)
- Ask participants to explain their questionnaire answers

What evidence can 5 users give to you?

5 users do not prove…

- … that you have a good design
- It may be the 6th participant who spots a critical error!

5 users can prove…

- … that the design has a critical error
- Even a single user can show that there is something to fix

⇒ User evaluations should try to identify errors, not prove that the design is good

1. When you design: Make as good design as you can
2. When you evaluate: Do your best to prove that your design is not good
3. If you fail to find critical errors, even if you tried really hard, then your design may actually be quite good

Scientific research follows the same principle (“falsification”)

Break – Continuing at 14:30

Questionnaire-based methods

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.

SUS scale maximum is 100 points

- If you get 68 or more points, that is said to be above the average
- But without A/B test or control group, the plain values may not mean much

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

AttrakDiff

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

- Provides a web tool to carry out all the data gathering + analysis

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

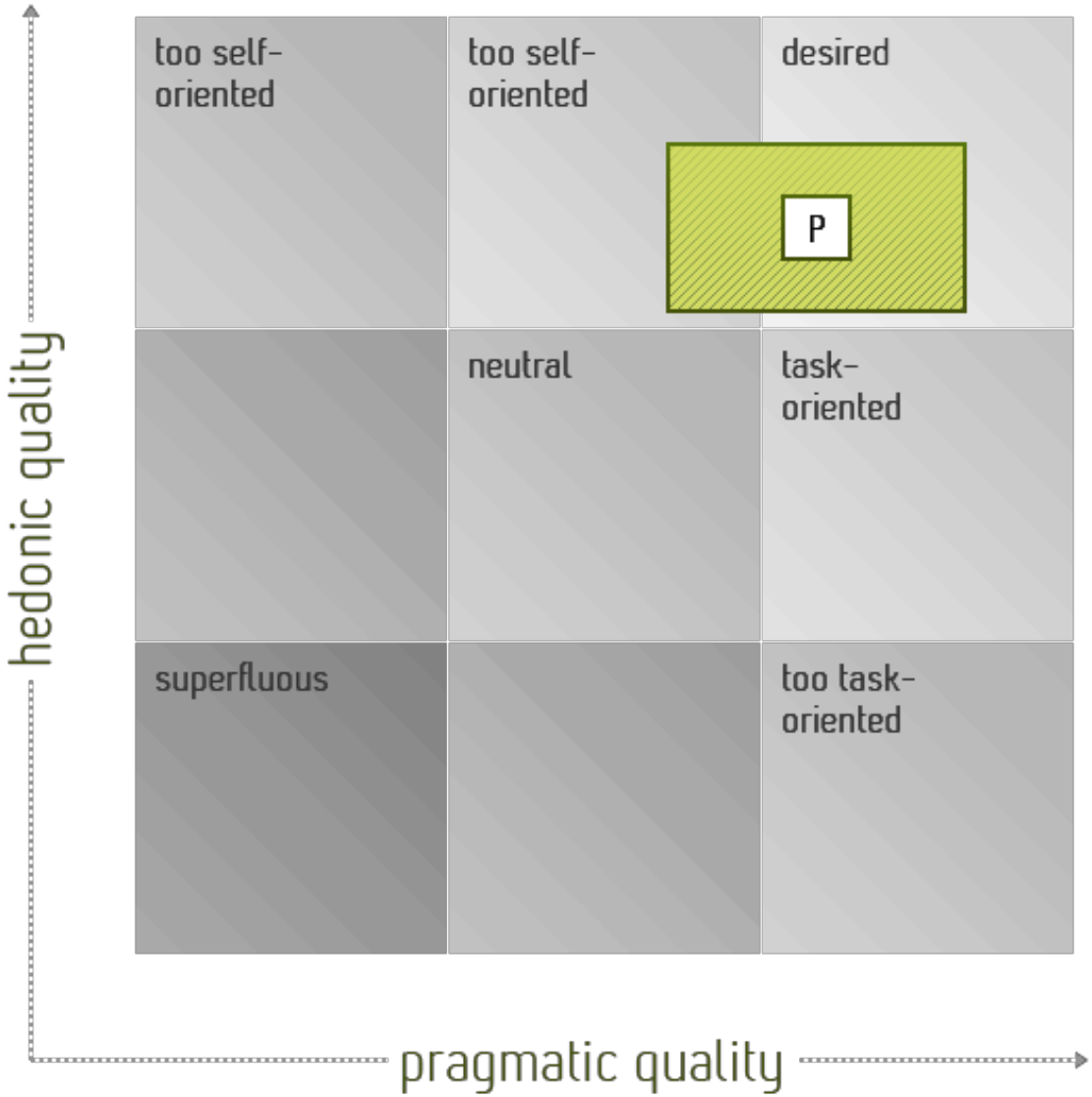
- Ugly — Beautiful
- Confusing — Clear
- …

Its result is 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



P — Medium value of the dimension with prototype P

Confidence rectangle

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

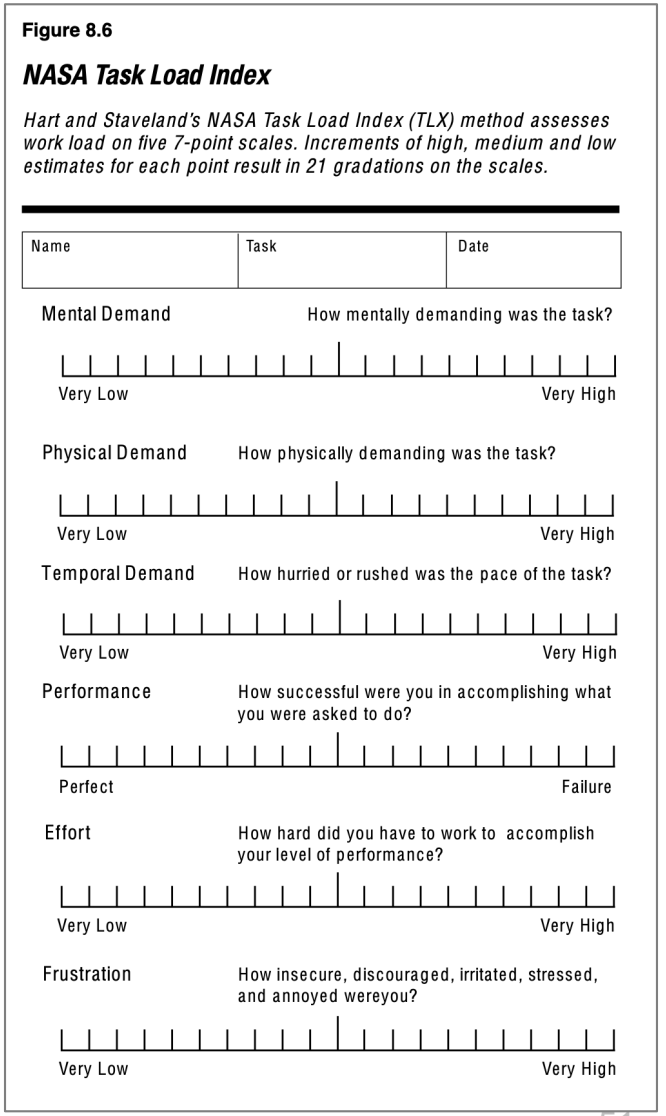


Figure 8.6

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

Tips:

- Use Likert statements (“Totally disagree – Totally agree”)
- Triangulate within your questionnaiore: Ask about the same topic several times, from different points of view
- Do a pilot study

How to visualize questionnaire data

Use bar charts to visualize answers

1. Calculate user-level averages:

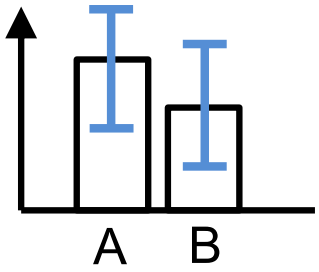
   Example: Average of all NASA-TLX answers from a user to Design A, and another average from answers to Design B

2. Calculate averages across all the users

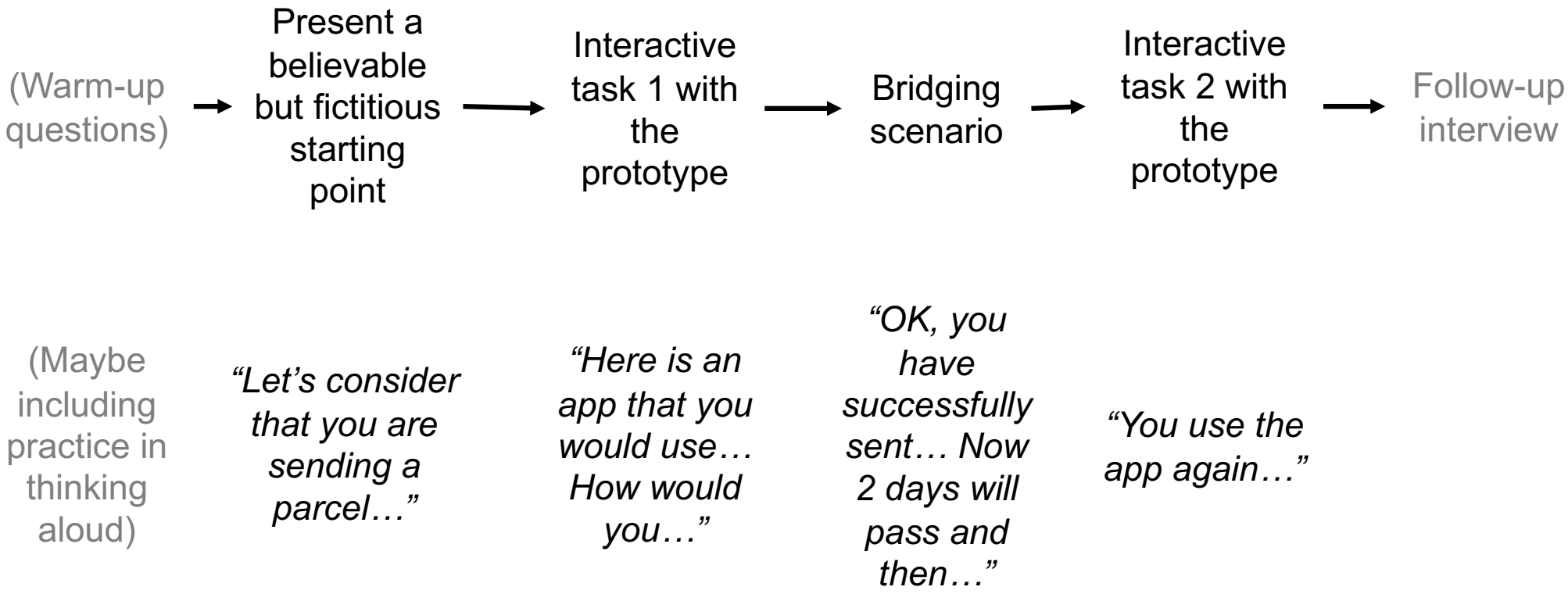
   Example: Average of all the user-level averages for Design A, and the similar average for Design B

3. Present the two bars in a diagram
4. Draw confidence intervals

   See e.g., YouTube tutorials on how to do that



Putting the methods together: The evaluation should have a story



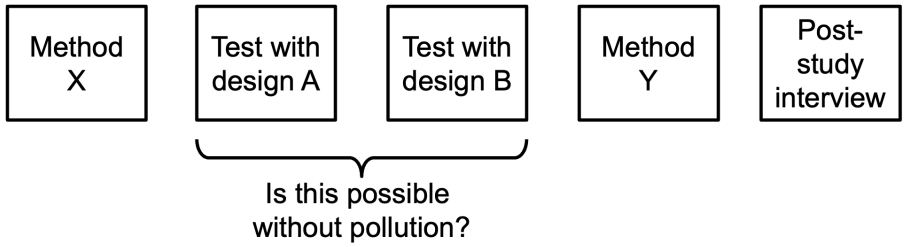
Group exercise (20 mins)

Invent a scenario for the evaluation

- The scenario should enable you to use your evaluation methods

Guidelines for a good evaluation:

- Evaluation should start with an easy task
- Methods should not "pollute" each other: an earlier method 1 should not simplify/complicate a later method
- Good scenario involves naturally repeating tasks because that increases redundancy and creates more data



New concepts:
Saturation
Within-subjects and between-subjects

Group exercise (15 mins)

Evaluation scenario version 2:

Look at your earlier plan from the previous exercise

- How can you improve it now?
- How can you make it more detailed?

Break?

Practical arrangements + tips

Making users feel relaxed

Making the evaluation more believable: helping users to “suspend their disbelief”

Pilot test

Some tips for successful studies

Making the user feel relaxed

Explain the anonymity and confidentiality in the beginning

- No names or other identifiable information will be revealed to Suomi-Seura or other people in the course
- User is free to terminate the evaluation at any time, with no need to explain why
- The recording and notes from the evaluation will be destroyed after the course
- But those contents that are relevant to the prototype's success will be kept and may also be used in presentations

Express interest in what user does

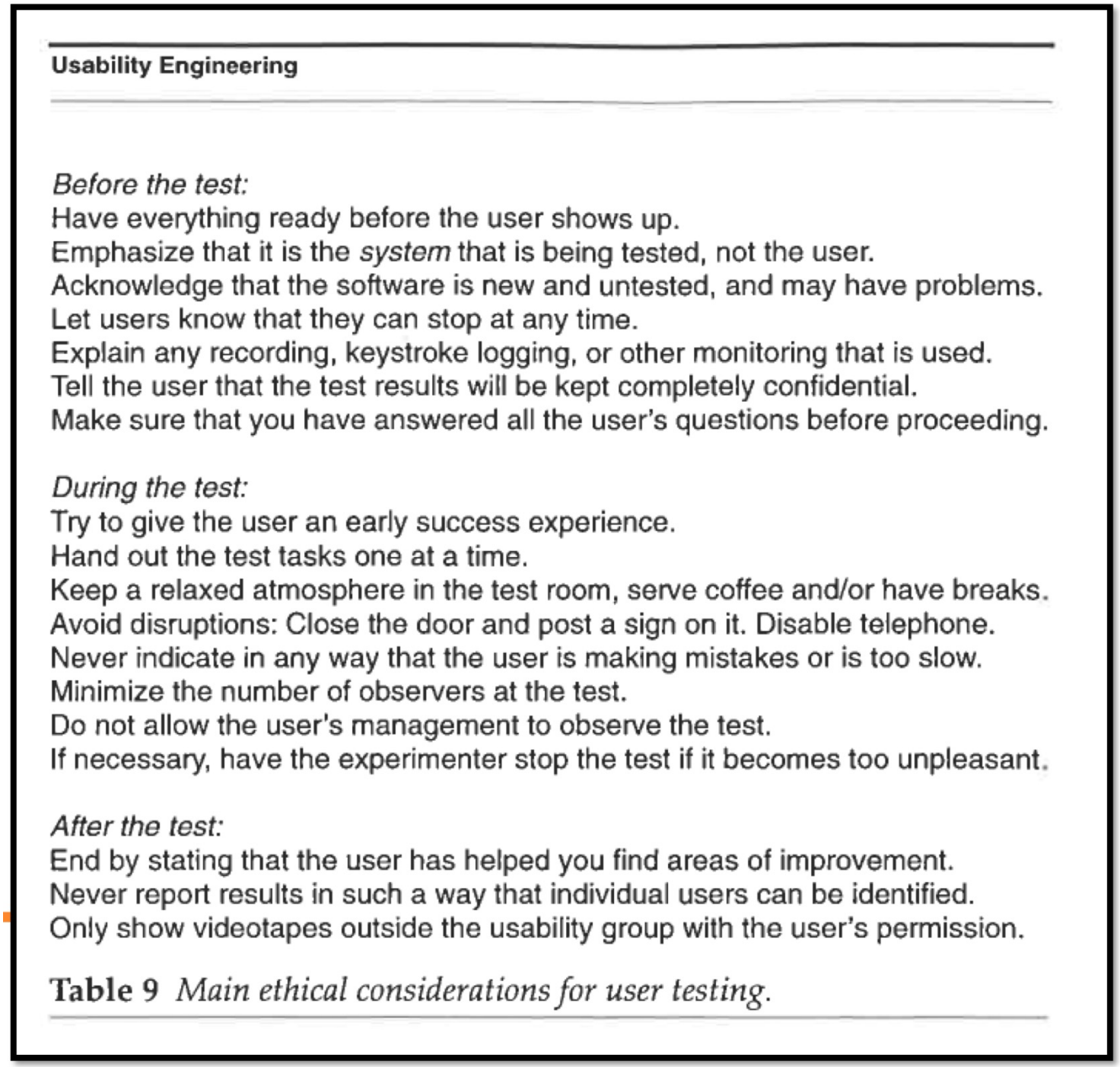
- Good also for gathering detailed data: if you ask for clarifications you both express interest and also don't leave unexplained user behaviours in your data

Don't:

- Don't sigh or yawn
- Don't express anxiety if user struggles
- Don't try to speed up the user if s/he is slow – Instead prepare the tasks so that some elements can be skipped without user noticing it

Do:

- First task has to be easy
- Present the tasks both verbally and visually on text => improves user's comprehension



Usability Engineering

Before the test:
Have everything ready before the user shows up.
Emphasize that it is the *system* that is being tested, not the user.
Acknowledge that the software is new and untested, and may have problems.
Let users know that they can stop at any time.
Explain any recording, keystroke logging, or other monitoring that is used.
Tell the user that the test results will be kept completely confidential.
Make sure that you have answered all the user's questions before proceeding.

During the test:
Try to give the user an early success experience.
Hand out the test tasks one at a time.
Keep a relaxed atmosphere in the test room, serve coffee and/or have breaks.
Avoid disruptions: Close the door and post a sign on it. Disable telephone.
Never indicate in any way that the user is making mistakes or is too slow.
Minimize the number of observers at the test.
Do not allow the user's management to observe the test.
If necessary, have the experimenter stop the test if it becomes too unpleasant.

After the test:
End by stating that the user has helped you find areas of improvement.
Never report results in such a way that individual users can be identified.
Only show videotapes outside the usability group with the user's permission.

Table 9 *Main ethical considerations for user testing.*

Nielsen, J. (1993). *Usability Engineering*. Boston, MA: Academic Press.

Suspending the disbelief: Mockups, propping 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

Pilot test

= “Dry run” of your evaluation

- Carry out everything in the way that you plan to do in the actual interview
- Recording method, tasks, mockup material, …
- No shortcutting! You also need to test the evaluation’s length!

Carry out one pilot test

- At least 1 day before the first actual interview
- One of team members pretends to be a user
- Make adjustment and fix problems

Compensations and costs

How to claim back costs:

Information is in MyCourses

100 eur/group

Research ethics

Which of these are ethically problematic actions?

Asking leading questions ("Don't you think that…")

Showing quotes from users to the project's customer

Using more time in an interview than was promised

Gathering a lot of background data about a user for the sake of completeness

Sharing user study data through Google Drive

Deceiving users by telling them in the beginning that the study is about one topic, but actually measuring something else

Sighing and yawning during an interview

Informed consent

User has to know what they are going to participate in, and give their permission:*

- Who are the members of the research team that organize this study
- That the purpose is not to evaluate the participant, but to investigate a research question
- That the participant may opt out any time during the study
- That the **relevant** material created by participants may be used in reports and publications (we'll return to this later)
- Confidentiality of the data: who will see it and in what form

These are explained in an **informed consent form** which the user can sign if they agree

* https://tenk.fi/en/advice-and-materials/guidelines-ethical-review-human-sciences

Informed consent form, part 1

The consent itself

INSIGHTFUL INTERFACES PROJECT

CONSENT FORM

I……..……agree to participate in the user interface experiment conducted by the Strategic Usability research group.

I have read and understood the study information sheet given to me. I have understood that the material and research data is gathered for scientific purposes only.

The purpose and nature of the study has been explained to me in writing. I have sufficient information on the process of the study. I understand that my participation in the study is completely voluntary and that I have the right to discontinue my participation at any stage without any consequences.

I give permission for my data to be recorded in the described manner. I understand that I can ask to take a break at any time during the study. It has been explained to me that a designated researcher will, at my request, provide me with additional details of the general principles of the study and its progress or of the results concerning myself.

I understand that anonymity will be ensured by disguising my identity. I have been explained who are the different parties involved in the research that have access to my data. I understand the practices of storing, protecting, and using the data. I know that the collected data will not be presented to a third party without my written consent. I know that the research group may ask for a professional consultation on possible unexpected incidental findings without separate consent provided that the anonymity of the results has been ensured. Any type of commercial exploitation of the results is prohibited.

I understand that a fully anonymized subset of the data may be released to other research groups for the purposes mentioned above, if I give permission to it.

(Please tick one box:)

[ ] I agree to releasing anonymized extracts from my data.

[ ] I agree to releasing anonymized extracts from my data only if I am informed about the research groups in question. I have been told what that subset will be.

[ ] I do not agree to releasing extracts from my data.

I understand that extracts from possible interviews may be quoted in subsequent publications if I give permission below:

(Please tick one box:)

[ ] I agree to anonymized quotation/publication of extracts from audiotaped data

[ ] I do not agree to quotation/publication of extracts from audiotaped data.

By my signature, I confirm my participation in this study and agree to volunteer as a study subject.

Date……..……..

RESEARCH PARTICIPANT	PRINCIPAL INVESTIGATOR
Name……	Name ……..……..……
Signature……..	Signature……
City……..	City…Espoo……

Informed consent form, part 2

INFORMATION SHEET FOR RESEARCH PARTICIPANTS

Name and topic of the research project: Insightful interfaces

General description of study method: This is mostly a quantitative system evaluation. The focus is on the times that text editing tasks with the word editor take from the participant in several different tasks.

Purpose of the study. The purpose of the study is to evaluate differences in task completion times between different versions of the system. Each participant will see only one of the system types.

Research group's experience of the method: The principal investigator of the study has experience of over 20 user studies in human-computer interaction. There are no reported incidents of ethical misconduct. The summer trainee who conducts the study has completed the course on usability evaluation methods with a high grade this spring (2019).

Funding and responsible researcher: This work is funded by the Department of Computer Science, Aalto University. Principal investigator is Dr. XXX XXX (tel XXX-XXXXXX, email xx.xxxxxx@aalto.fi).

Time commitment: Participation in this study will take appr. 40 minutes (max 1 hour).

Suitability for the study: Legally competent adults are allowed as participants. In particular we require good command of English (due to the language of our system) and experience of using word processing software.

Compensation: The compensation is one movie ticket per hour.

Voluntary participation: Participation in the study is voluntary. You have the right to discontinue participation at any time without obligation to disclose any specific reasons.

The rights of the study participant: As a research participant you have the following rights: the right to access stored personal information, to correct inaccurate personal information, to oppose the processing of your personal information and to delete your information. It might be necessary to depart from the participant's rights if the research is conducted for the purpose of public interest and if the participant's rights prevent or greatly hinder achieving this purpose. If, however, it is possible to achieve the aims of the study and the achievement of the purpose is not greatly hindered, Aalto University will actualize your rights as defined in the GDPR. The extent of your rights is related to the legal basis of processing of your personal data and exercising your rights required proof of identity.

Communication with the research staff during testing: You can stop the task and ask the experimenter at any time if you have questions about the study or your participation.

Description of study situation: The study starts with a short paper-based questionnaire about familiarity of symbols in computer software. After this, the experiment with our system will start. There will be one practice task and 10 actual tasks. The system will log the interactions in these tasks.

Collection of data: 1) Questionnaire data: computer literacy scale; 2) Screen recording data: mouse movements, text editing in the tasks; 3) Audio recording data: the conversations with the experimenter and the participant during the tasks; 4) Personal information: name, email address, gender, age. Personal information is collected to enable communication with the subject and for statistical information about the participants.

Transferring data outside the EU: Your data will not be transferred outside the EU, except for the United Kingdom.

Anonymity, secure storage, confidentiality: The data will be used for scientific purposes only and are confidential. All data will be anonymized. No explicit clues of your identity will be left to the rest of the stored data. All data will be stored securely and accessible only to the following members of Aalto University: XXX XXXXXXX and XXX. The questionnaire data, screen recording data, audio recording data, gender and age are accessible also to XXX and XXX, both from XXX. The personal information will be deleted when it is no longer needed.

Insurance coverage: You are covered by Aalto-level insurances for accidents and damages during the study.

Contact details: Aalto University is the data controller in this research.

In questions regarding research you can contact the responsible researcher: XXX, xxx.xxxxxxxx@aalto.fi, You can contact the Aalto University data protection officer if you have questions about data processing and protection: Xxxx Xxxxxxxxx, xxx@aalto.fi, tel. XXXXXXX If you notice a violation in the data protection legislation, you can contact the Data Protection Ombudsman (http://www.tietosuoja.fi/en).

If you agree to take part in the study, please sign the consent form overleaf.

GDPR (General Data Protection Regulation)

General advice:

- Be specific in the informed consent
- Collect only the data you need
- Define when the data will be deleted
- Specify where the data is stored securely
- Do not reveal the identities of users to each other or outsiders
- Keep a record of the consents
- ⇒ Include these in the informed consent

Special considerations:

- Do you plant to gather data from which participant can be indirectly identified?
- Does your interview deal with intimate personal experiences?

Useful links:

- Aalto: https://www.aalto.fi/en/services/aalto-university-data-protection-policy
- From UK Government: https://www.gov.uk/service-manual/user-research/managing-user-research-data-participant-privacy

Prohibited methods in this course

Do not plan these kinds of methods in your project:

1. Intervention in the **physical integrity** of subject
2. The study deviates from the principle of **informed consent** (excluding archival data)
3. The subjects are **children** under the age of 15
4. **Exceptionally strong stimuli** whose harmfulness needs to be evaluated by an expert
5. Possible long-term **mental harm** (trauma, depression, sleeplessness)
6. Possible **security risk** to subjects

How to fit all the ethics in a user evaluation

Before you meet in the interview/observation/test:

- Send the informed consent document(s) to the participant **in advance**
- => They will have more time to investigate them

Plan the timings of your meeting carefully

- Reserve 5–10 minutes of the beginning to informed consent, GDPR, confdentiality and anonymity principles & making sure that the participant is assure of good practices
- Decide which parts must be included and which can be dropped

Run a pilot study

In the evaluation:

- If you notice that you will run overtime, ask if user can stay longer.

Exercise (10 mins)

Analyse your previous test scenario plan

- Write how many minutes each step can take

Do you need to change anything?

Friday's program (heuristic evaluations)

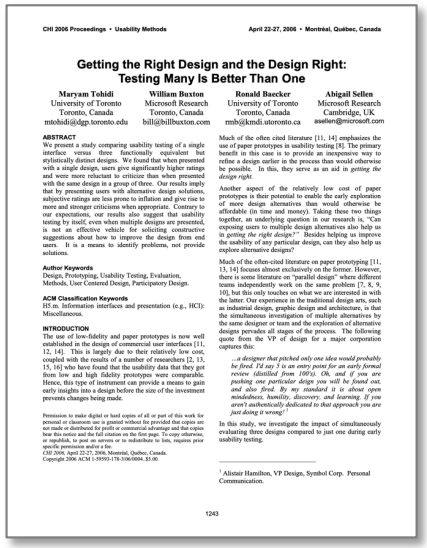
Heuristic evaluations on Friday

Instead of presentations, our program is:

13:00 – 14:00:	Discussion of reading materials
14:15 – 14:45:	Preparing the prototypes for heuristic evaluation
15:00 – ? :	In 2–3 person teams, heuristic evaluations on another group's prototype(s)

Reading materials + Quiz for week 5

Reading materials for week 5



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 got 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 3 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.

1243

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

273

Goodman, Elizabeth, et al. Observing the User Experience : A Practitioner's Guide to User Research, Elsevier Science & Technology, 2012. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/aalto-ebooks/detail.action?docID=878450.
Created from aalto-ebooks on 2022-01-06 21:13:31.

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

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

See you in the tutorings tomorrow!