Move so that you sit with your project group members

Lecture 3: Empirical Usability Evaluations

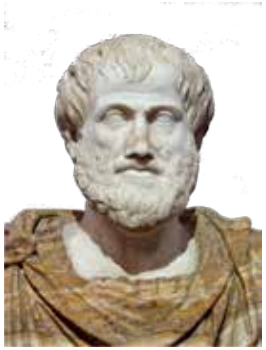
Antti Salovaara

With thanks to Aqdas for some of the contents

What this lecture is about

How you can plan a successful *empirical* usability evaluation

(empirical = with users)



Periechómena

(Table of contents)

1st act. Usability specialist confronts a challenge: the deceptive simplicity of usability evaluation with users

2nd act. The specialist is given the keys to the wisdom of research design

3rd act. The specialist is empowered and starts practicing the wisdom

https://upload.wikimedia.org/wikipedia/commons/thumb/a/ae/Aristotle_Altemps_Inv8575.jpg/896px-Aristotle_Altemps_Inv8575.jpg

1. The deceptive simplicity of usability evaluation with users

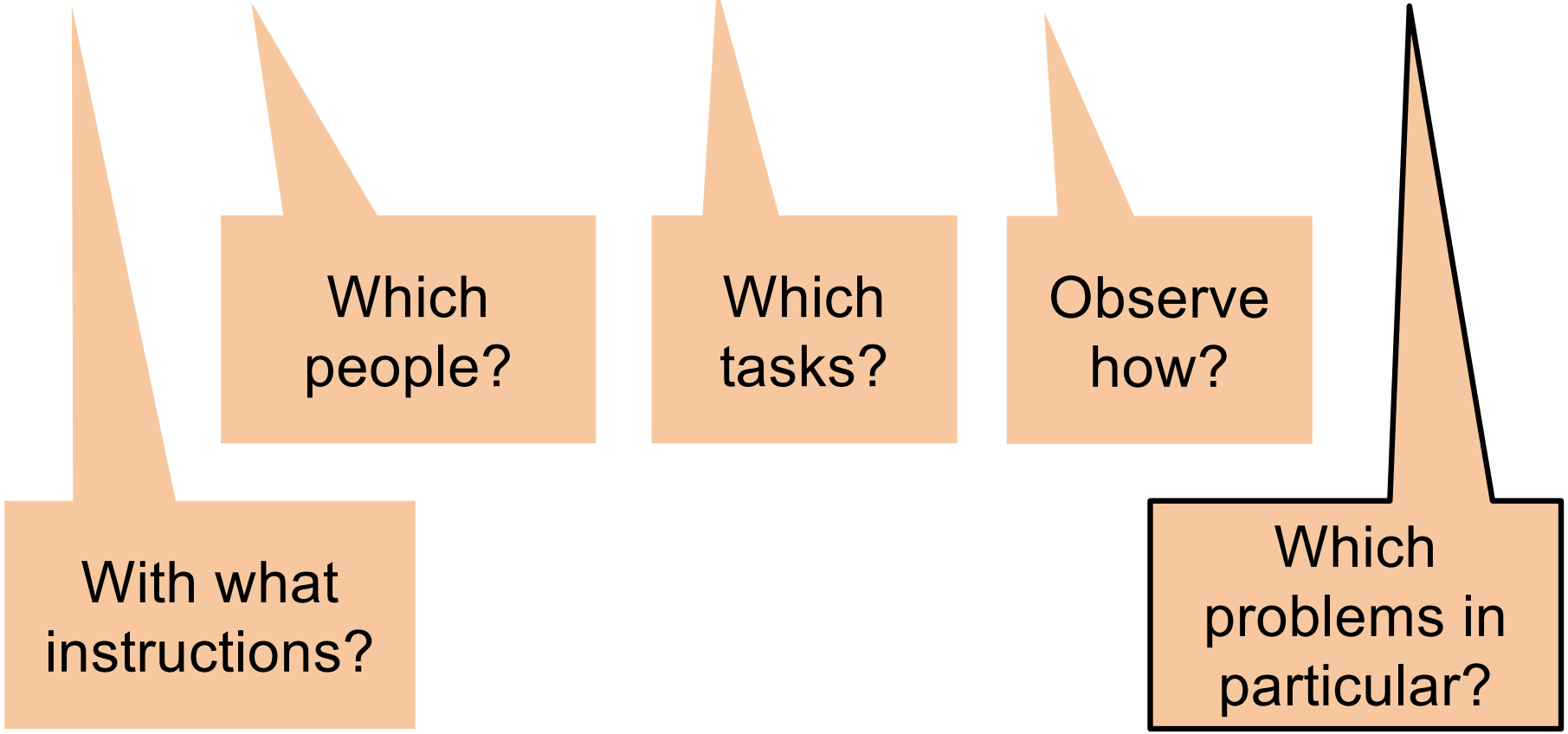
Usability evaluation (shown in Lecture 1)

Controlled 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

A simplistic view on usability tests + its problematizations

"Ask people to carry out tasks and observe their problems"



Evaluating usability comprehensively

Which problems in particular?

The standard 3-part characterisation:

- Effectiveness (the outcome)
- Efficiency (the interaction process)
- Satisfaction (the user experience)

Other possible characteristics:

- Safety
- Shock resistance
- Long battery time
- …etc

Which characteristics are of priority?

- This needs to be decided case by case
- Also your project needs to think about and prioritize usability characteristics

Case-specific measures of usability

Which problems in particular?

Which tasks?

Evaluations' purpose is usually to identify needs for improvement

=> "Objective" quantification of usability is not always of much interest

However check "System Usability Scale" (SUS): 10 Likert statements ("I think that I would like to use this system frequently.") *

Often measurements depend on 1) usability problems and 2) tasks that are of special interest

E.g., Hornbæk's review (2006) found 6 different classes of measures used only for evaluating effectiveness (+ the "other" category)

* https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html

Table 2
Measures of effectiveness

Measure	N	%	Explanation
Binary task completion	24	13	Number or percentage of tasks that users successfully complete
Accuracy	55	31	The accuracy with which users complete tasks, that is some quantification of error
Error rates	46	26	Errors made by the user during the process of completing a task or in the solution to the task
Spatial accuracy	7	4	Users' accuracy in pointing to or manipulating user interface objects
Precision	3	2	The ratio between correct information and total amount of retrieved information
Recall	11	6	Users' ability to recall information from the interface
Completeness	11	6	The extent or completeness of users' solutions to tasks
Quality of outcome	28	16	Measures of the quality of the outcome of the interaction
Understanding	18	10	Understanding or learning of information in the interface
Experts' assessment	8	4	Experts' assessment of outcomes of the interaction
Users' assessment	3	2	Users' assessment of the outcome of interaction
Other	6	3	Other measures of effectiveness

The result

"Ask people to carry out tasks and observe their problems"

The simplicity of usability evaluation methodology is deceptive

There are countless methods, each with its pros and cons, to carry out a usability study

Next up: Dealing with this challenge

EXERCISE

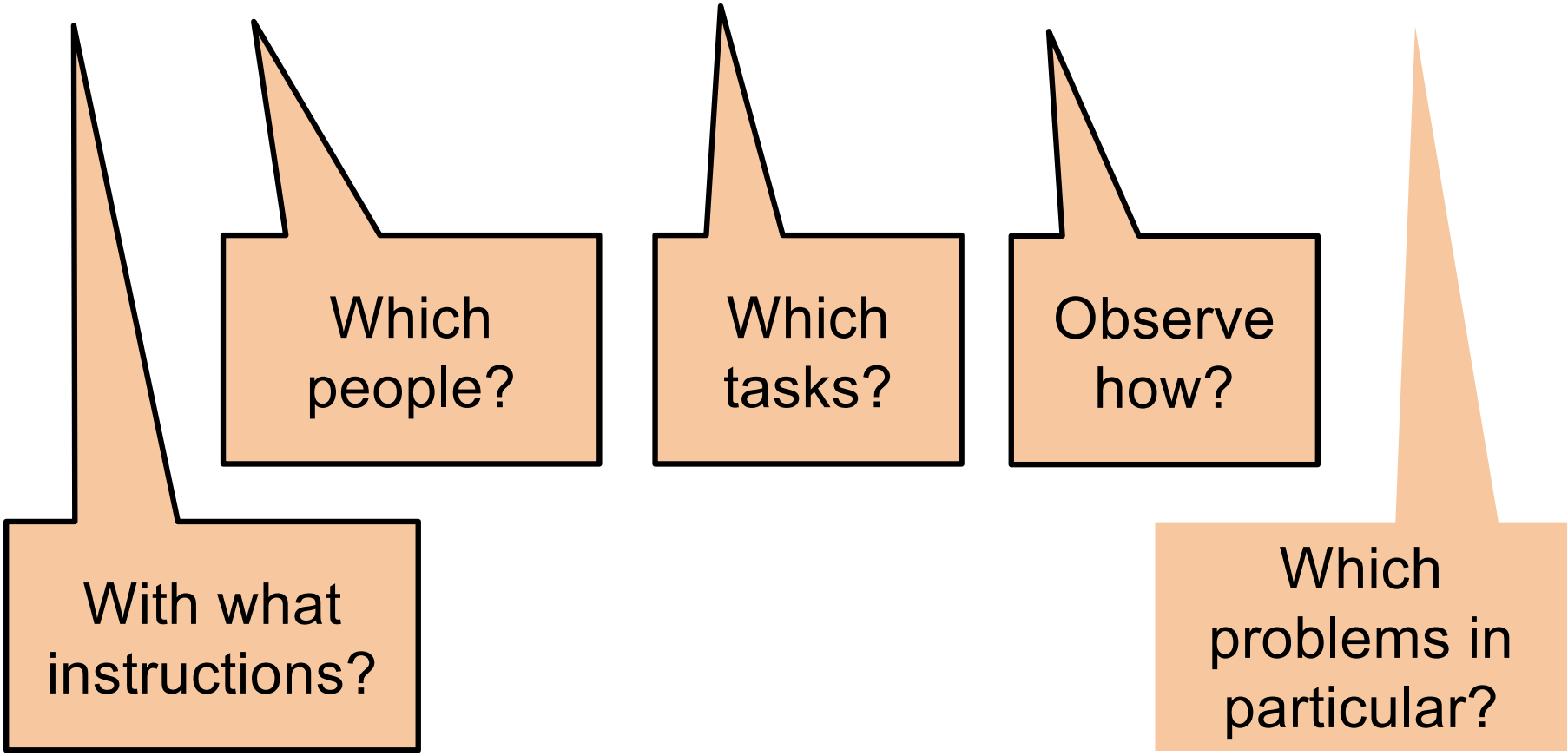
Contextualize the usability characteristics for your product

Usability characteristic	**Task where the characteristic is relevant**		
Effectiveness (outcome)			
Efficiency (interaction process)			
Satisfaction (user experience)			
?			

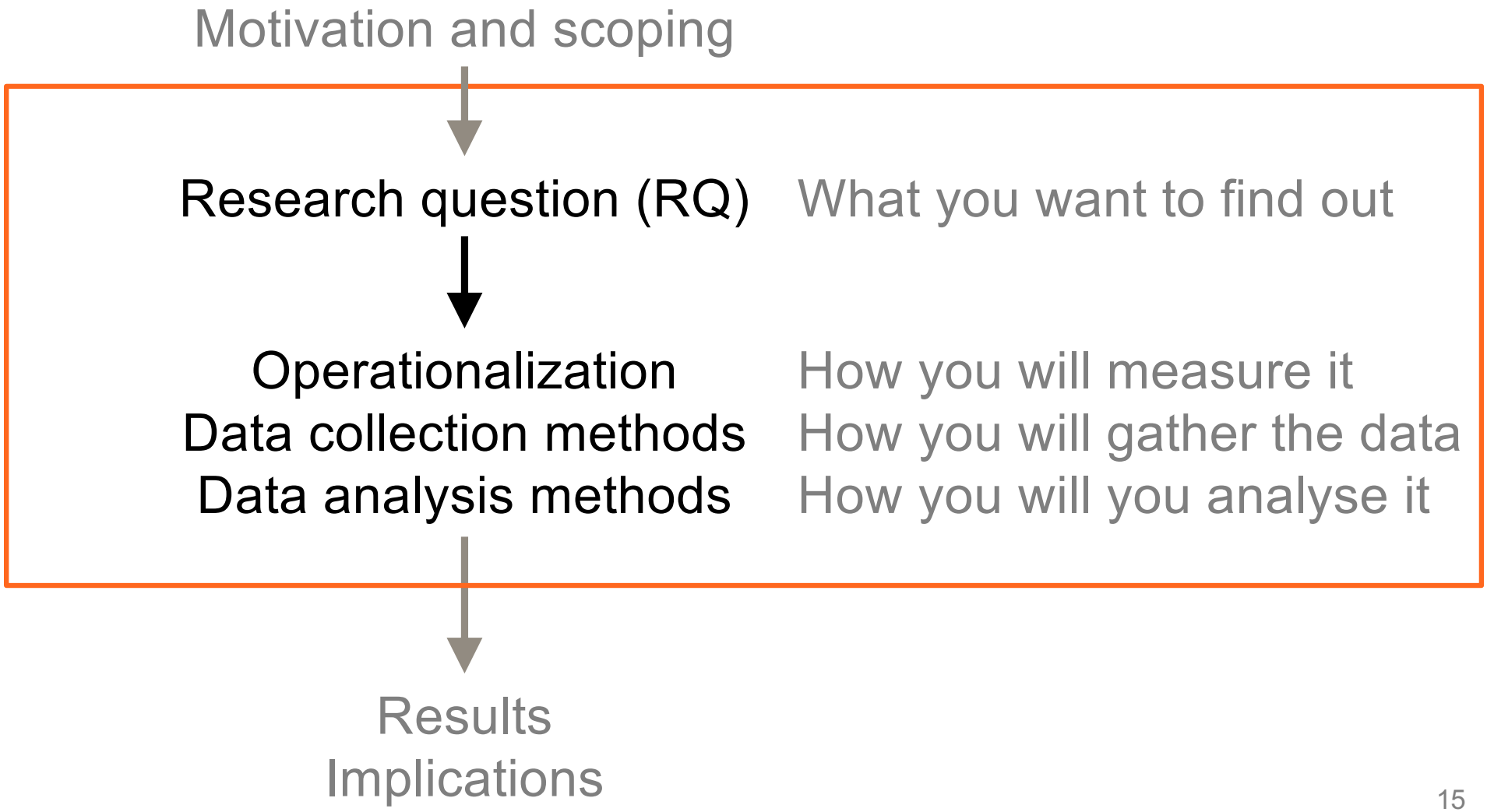
2. Research design

Turning the usability characteristics into research methods

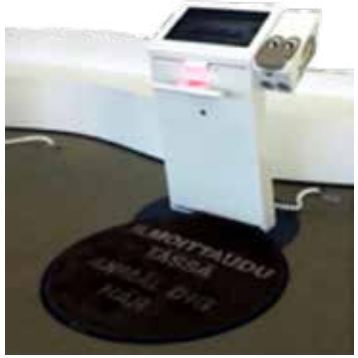
"Ask people to carry out tasks and observe their problems"



Research design =
how you design your research process

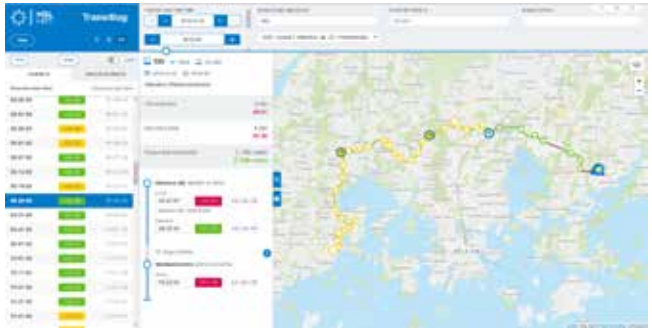


Examples of research questions



Can a 5-year old child create an avatar for herself using the New Children's Hospital self-service registration kiosk?

Yes / No



What kind of delay types can be best spotted using HSL's TransitLog?

Close-ended question

What kind of user experiences does Elisa Viihde elicit in users?

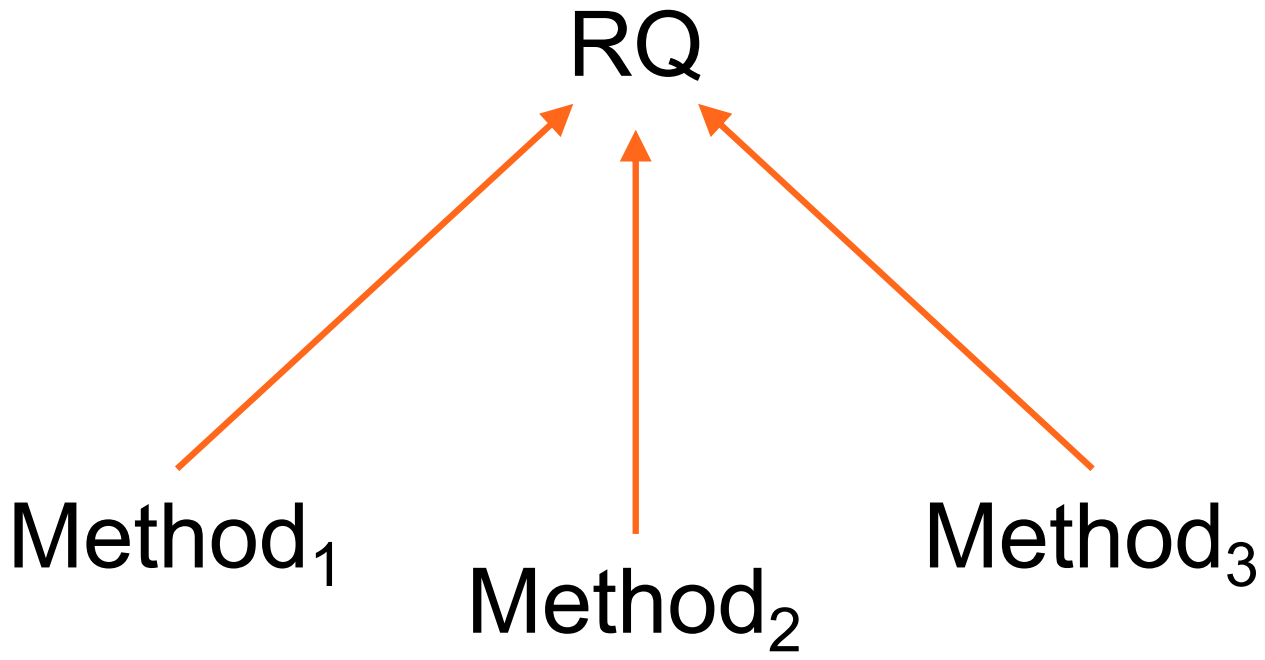
Open-ended question

Operationalization: turning RQ into methods

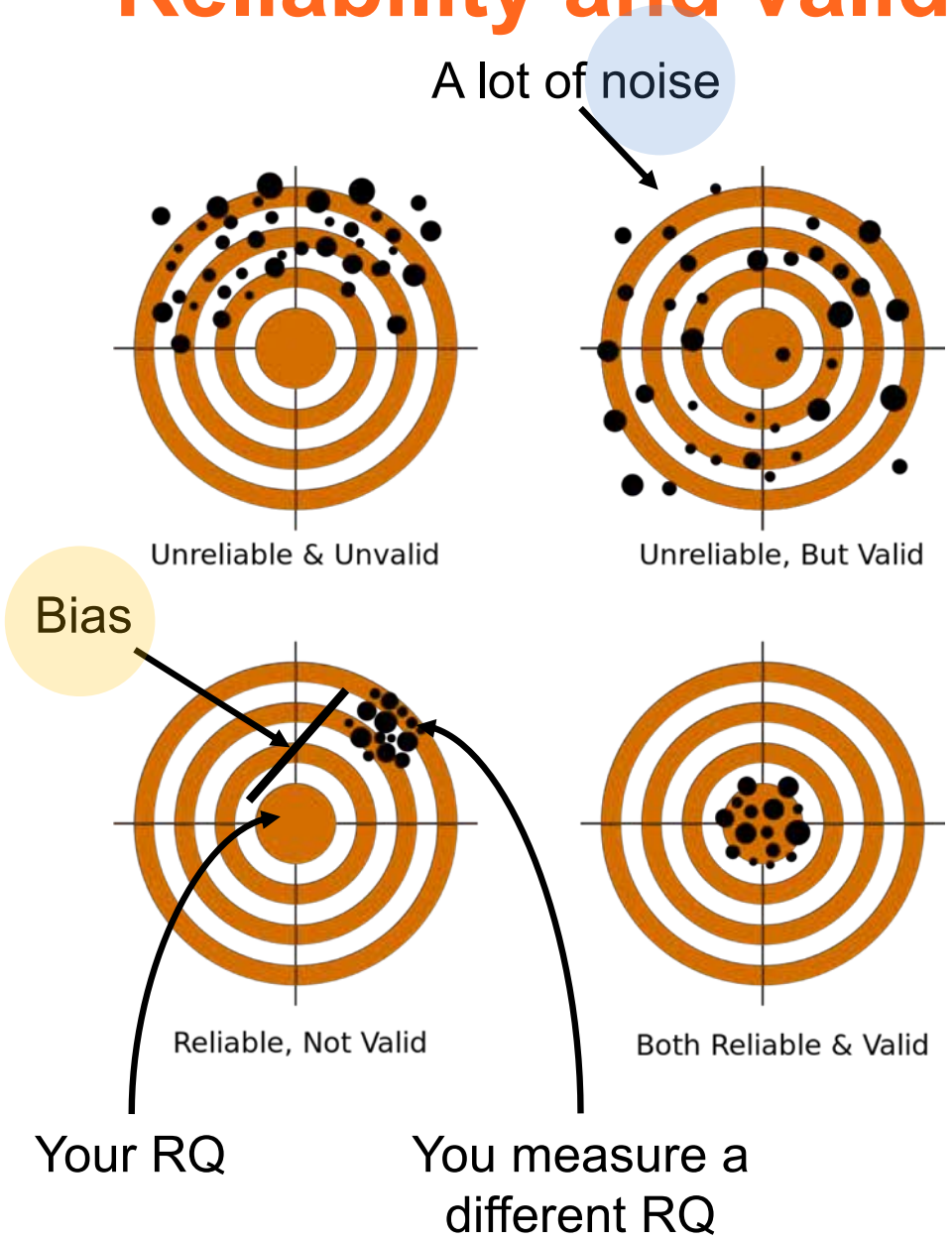
The challenge:

All RQs can be studied in several ways.

Which method(s) should one choose?



Reliability and validity of a method



Validity:

= method measures the intended RQ

~ Bias

Reliability:

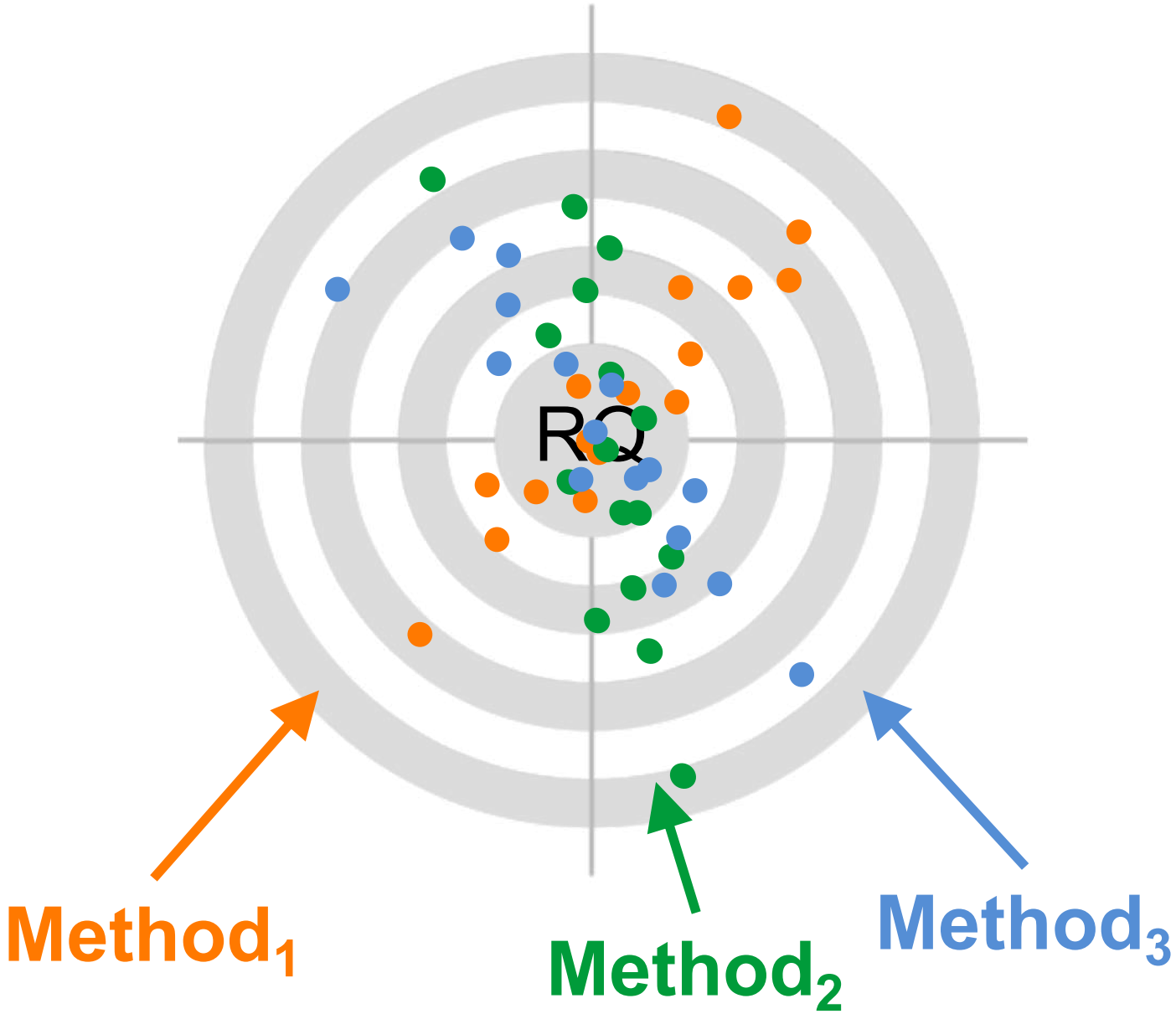
= method measures the RQ with good detail

~ Noise

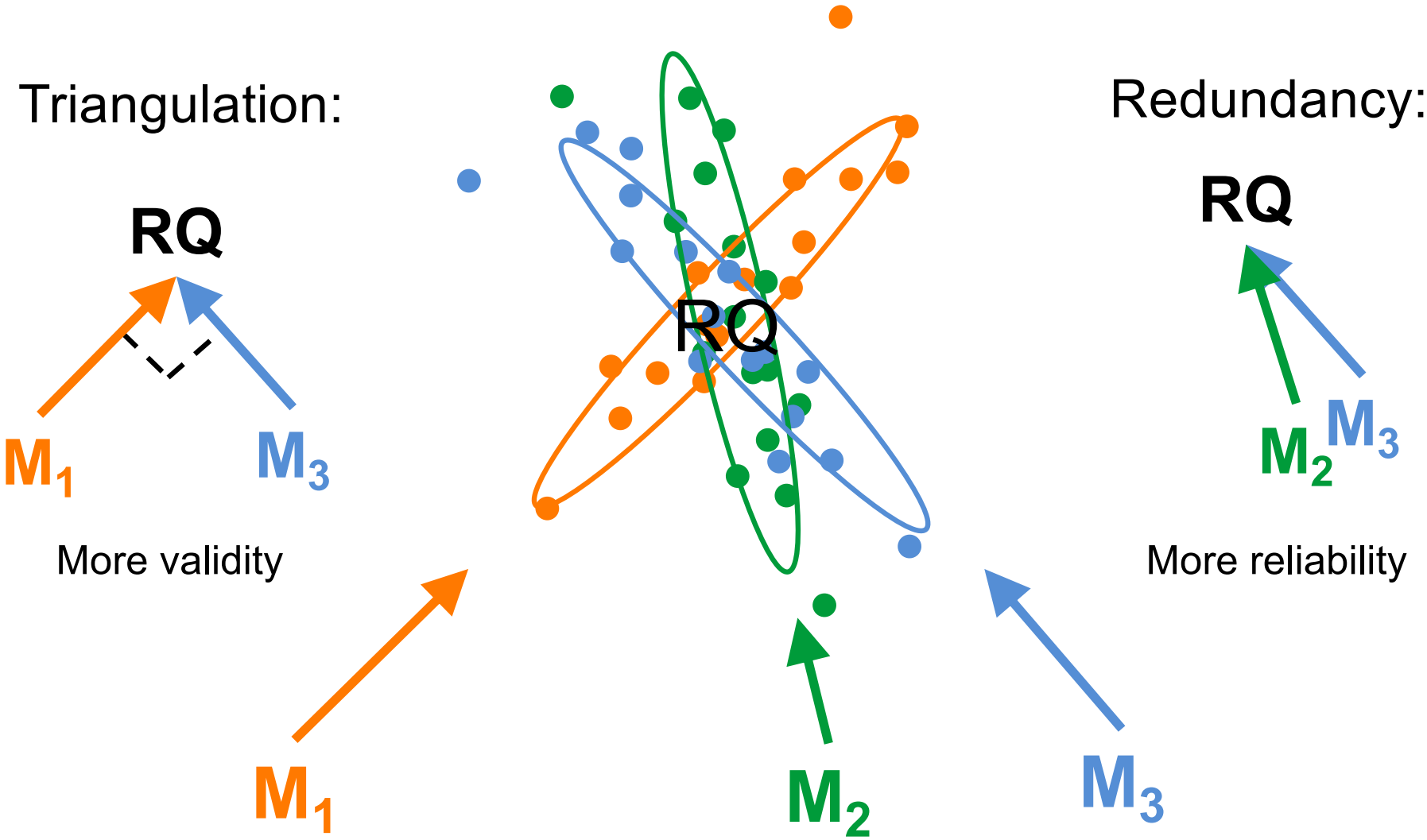
Both criteria should be met in a good method

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

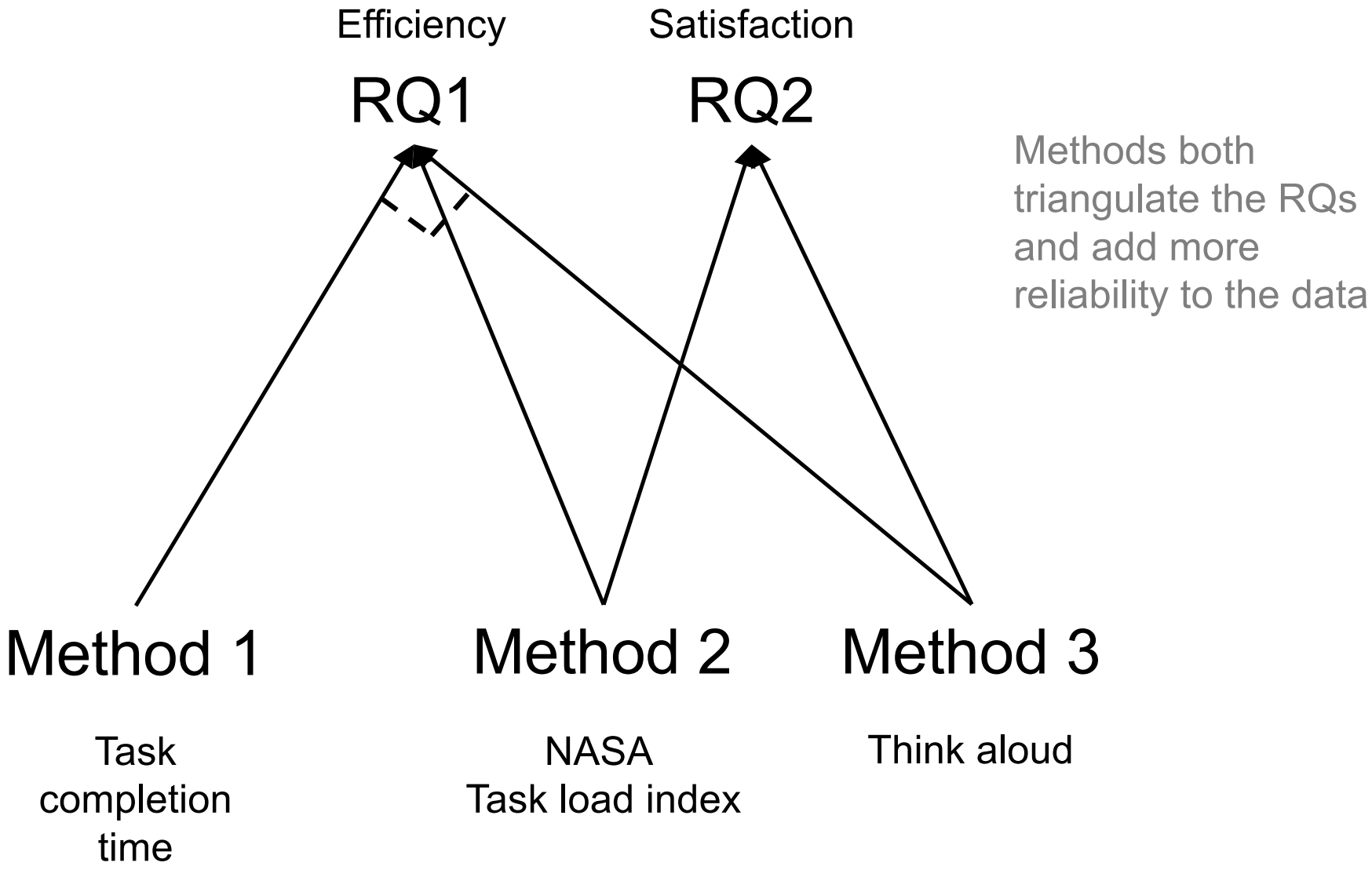
Operationalization with several methods



Triangulation and redundancy



Example of a desirable research design



Methods both triangulate the RQs and add more reliability to the data

Special challenge: Small sample sizes

Usability evaluations often combine experimental research with qualitative methodology

Experimental research:

Controlled lab setting

- Think: “psychological study in a lab”
- Better reliability (less noise)

Traditionally combined with:

- Preference for quantitative measures
- Preference for statistical analyses

Qualitative research:

Small number of participants

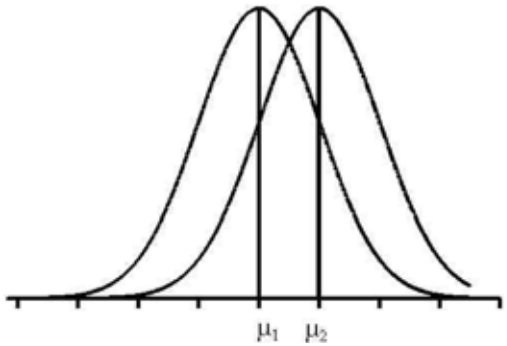
- Typically $N \leq 10$

Reasons for small N:

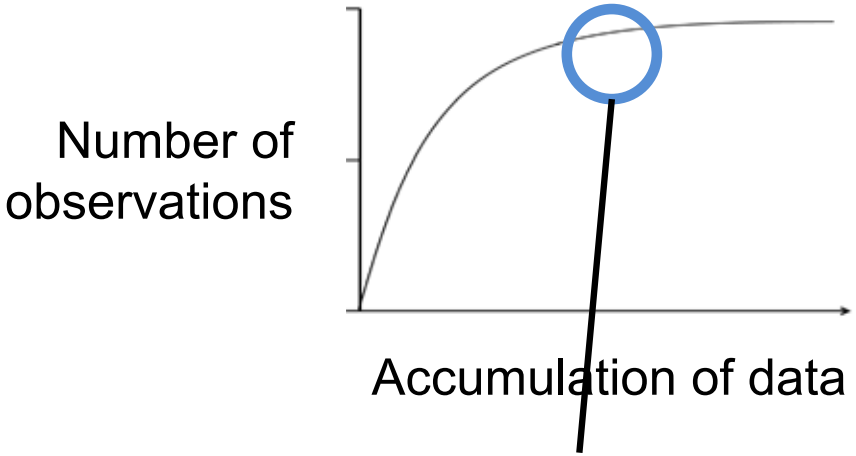
- Need for rapid results
- Interest in identifying things that need improvement, not in testing hypotheses
- Limited availability of representative users

Differences between the methodologies

Statistically significant differences between conditions



Saturation of observations



Point of saturation: When new data does not increase your understanding

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

What can qualitative methodology tell? – Two research approaches

User-centred concept development

User research

Inductive research:

- Develops new research hypotheses
- ”This data seems to suggest that users like these things”
- A large amount of evidence allows for detailed hypotheses but can never prove that they are ultimately true

Usability evaluations

Falsifying research:

- Seeks to identify false beliefs
- “This data shows that users cannot use the system without problems”
- Even a small amount of evidence proves that there is a usability problem

EXERCISE

Narrow down your study to RQs and operationalize them

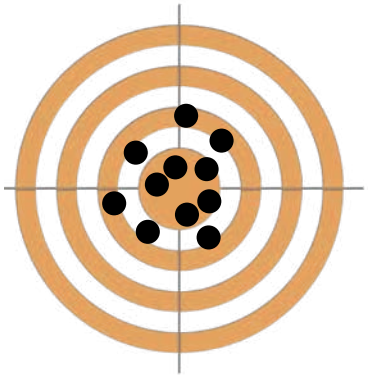
Usability characteristic	**Task where the characteristic is relevant**	**RQ**	**Operationalization (how RQ can be measured?)**
Effectiveness (outcome)			
Efficiency (interaction process)			
Satisfaction (user experience)			
?			

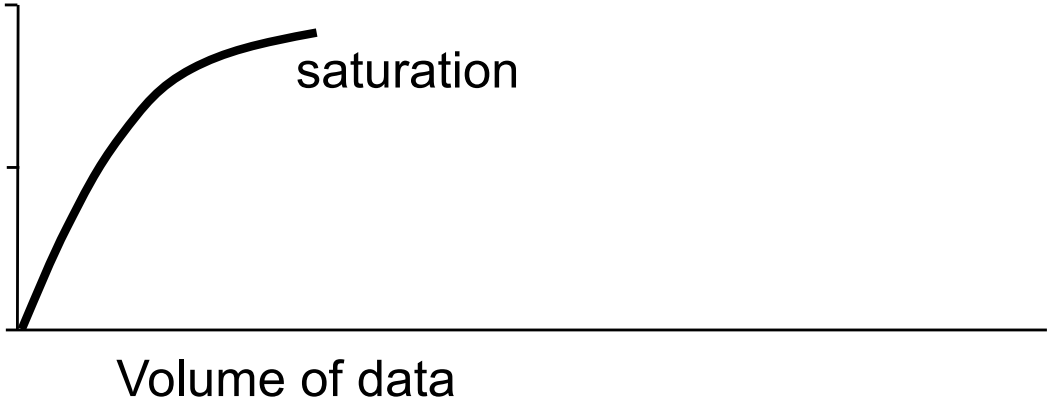
3. Practicing the usability evaluation methods

A collection of tips that improve reliability and validity

Minimize open-endedness

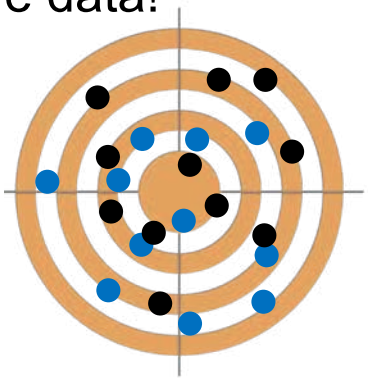
Tightly scoped research:

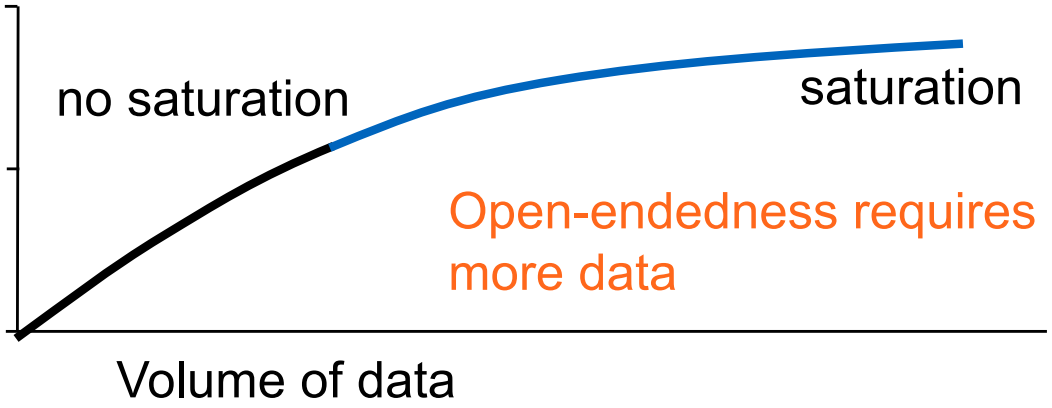




more data!

Loosely scoped research:





Decide your sampling (recruitment) strategy

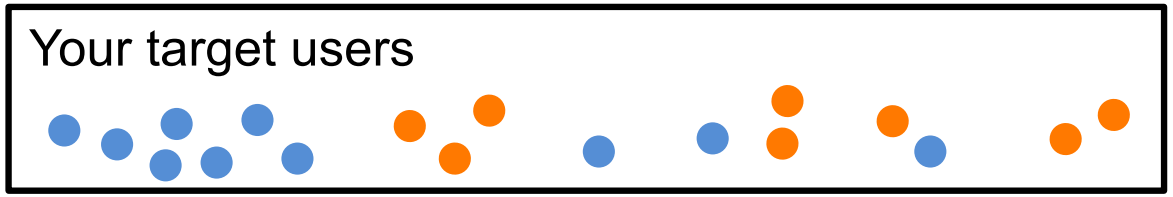
Random sampling

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

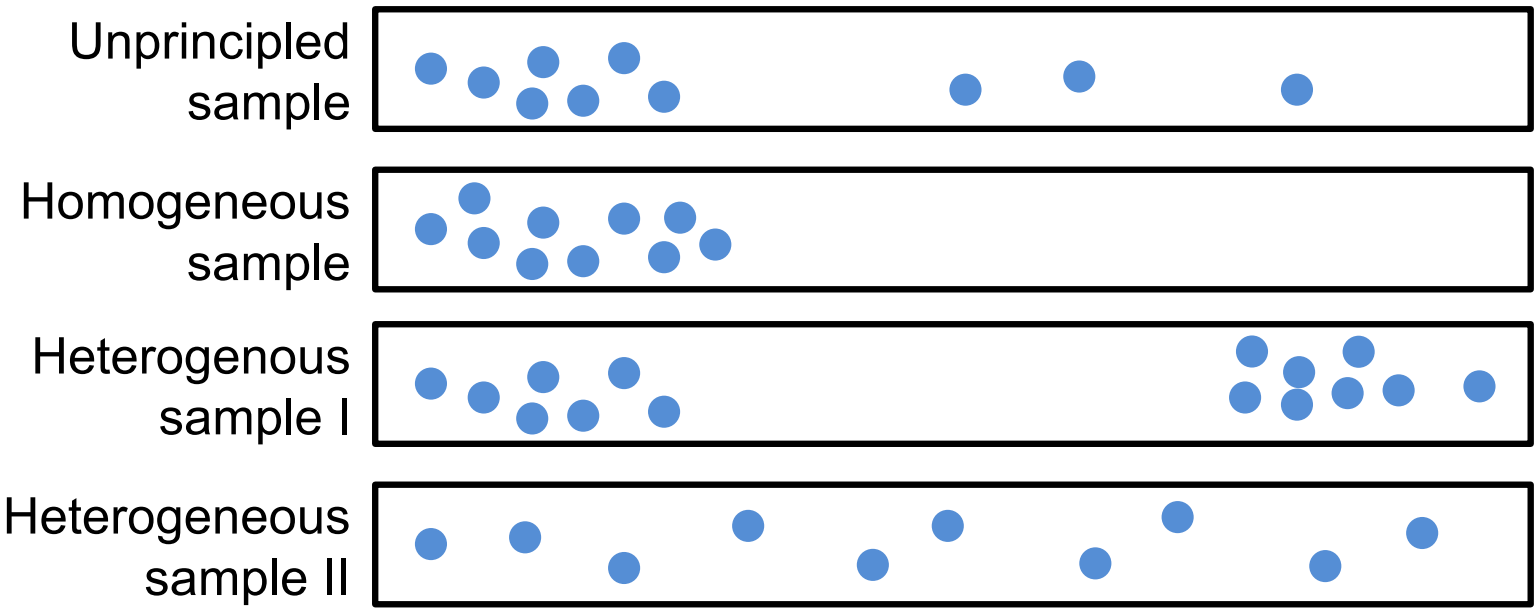
Choose between heterogeneous vs homogeneous samples

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



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.

Making the user feel relaxed

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:

- Create a simple first task
- Present the tasks both verbally and printed on paper, one paper slip at a time (in order to allow skipping)

Preparation of mockup (stimulus) material

Every usability evaluation is an **intervention** in the normal life of a user

- User is asked to carry out an artificially constructed task
- Even the system may be an unfinished prototype => This makes usability evaluations “time machines” that investigate possible futures*

Yet the task should feel natural and believable

- ⇒ You need to prepare authentic-feeling task material for the evaluation
- E.g., for an evaluation of CAD software, an unfinished 3D design

* Salovaara, A., Oulasvirta, A., & Jacucci, G. (2017). Evaluation of prototypes and the problem of possible futures. In Proceedings of the SIGCHI Conference on Human Factors in Computing (CHI 2017). New York, NY: ACM Press.

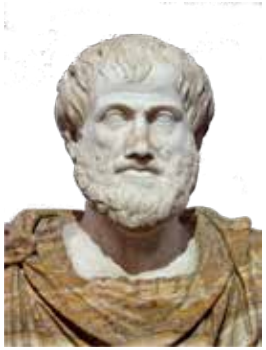
Yardstick of rigour: methodology of psychological experiments

If you have resources, make your evaluation as close as psychological experiments as possible

- Add comparisons between several alternatives (different interaction paths, alternative interactions, competing products) *
- Add repetition to increase reliability
- Use between-subjects and within-subject designs in the comparisons
- Use quantitative measures and a large N
- Use statistical testing methods
- Measure confounding variables that you cannot “control away”

* Comparisons are not allowed in this course because our customer companies want to carry out comparative evaluations confidentially by themselves.

Epilogue



Periechómena

(Table of contents)

1st act. Usability specialist confronts a challenge: the deceptive simplicity of usability evaluation with users

2nd act. The specialist is given the keys to the wisdom of research design

3rd act. The specialist is empowered and starts practicing the wisdom

https://upload.wikimedia.org/wikipedia/commons/thumb/a/ae/Aristotle_Altemps_Inv8575.jpg/896px-Aristotle_Altemps_Inv8575.jpg

What this lecture taught you

You must confront the challenge that there are many ways to carry out an evaluations with user

The lecture introduced several research design concepts the help you decide the details of your research method

- You can use your judgment in deciding what to focus on
- Considering all three usability characteristics increases comprehensiveness of your evaluation

Lecture also offered several practical recommendations for important details

EXERCISE

Continue working on the chart

Usability characteristic	**Task where the characteristic is relevant**	**RQ**	**Operationalization (how RQ can be measured?)**
Effectiveness (outcome)			
Efficiency (interaction process)			
Satisfaction (user experience)			
?			