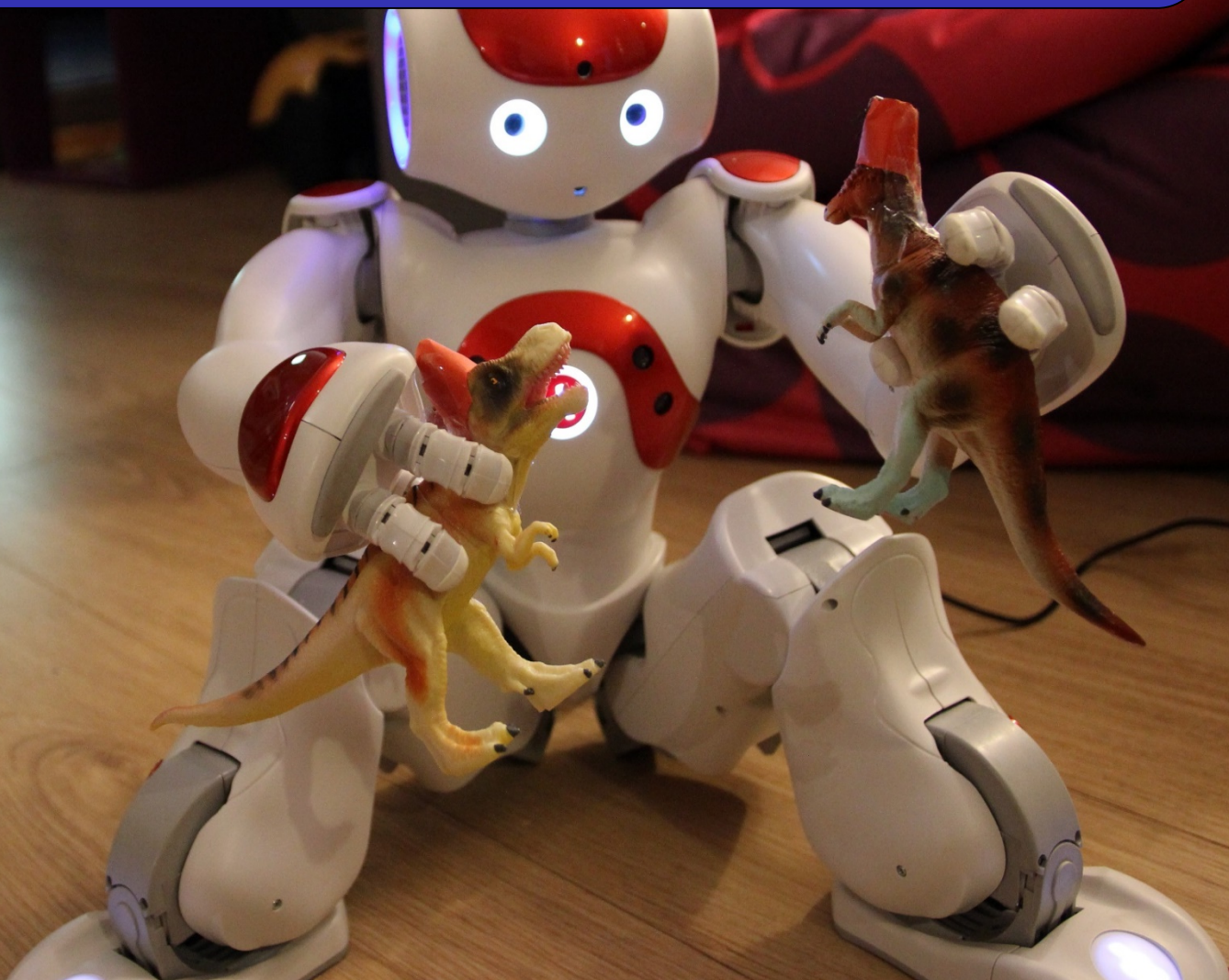


How Did I Do?

Evaluating Quality of Interaction

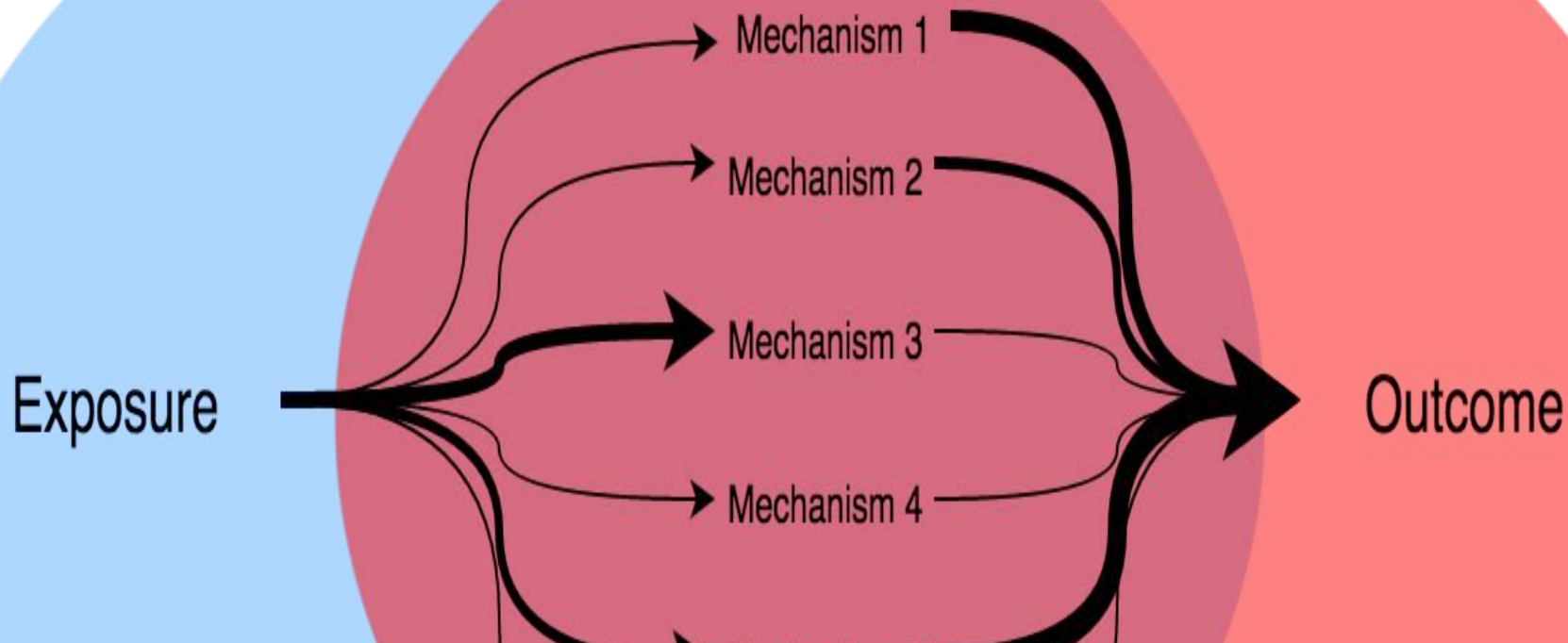


Goal & Learning Objectives

The goal of this lecture is to discuss methods for evaluating a social robot design

You are able to:

- create an evaluation procedure
- execute a procedure
- write up the results
- reflect on types of evaluation approaches and discuss their strengths and weaknesses



STUDY GOALS & TYPES

Study Goals

Formative Evaluation

- Exploratory
 - Informs the design process
 - Gives insight into design problem and solution
-

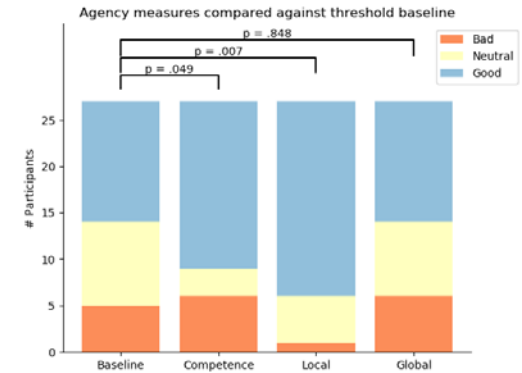
Summative Evaluation

- Conclusive
- Assesses the success and quality of the design

Study Goals

Quantitative

- About result of HRI
- Systematic
- Empirical / numerical data



Qualitative

- About process of HRI
- Meaning (why) / description of (how)
- Non-numerical data

"It's okay the robot couldn't always understand me. Other people don't always understand me either."

Study Goals

Simulation

- Evaluate
 - Technical performance
 - System behavior
- Types
 - Artificial interactions
 - Mock-up interactions



User study

- Real people / end-users
- Essential for user-centered design



Lab study

- More control
- Lower ecological validity



Image: Personal Robots Group, MIT Media Lab

“In the wild”



- Less control → quasi-experimental design
- More ecological validity

Study Type: Pilot

- **Goals**

- Evaluate proof of concept
 - Evaluate research set-up (mini-version of other study)
-

- **Characteristics**

- Formative (for evaluation procedure)
- Less strict procedure
- Low # participants

Study Type: Exploratory

Goals

- Gain insights into processes.
 - Generating research question / hypotheses.
-

Characteristics

- Primarily formative
- Primarily qualitative
- Often one condition
- Less strict / restrictive procedure
- Low-medium # participants


Study Type: Comparison

Goals

- Validate (added value of) your design.
- Measure effects of your design.

Characteristics

- Often compares robot with and without designed behavior
- Summative
- Primarily quantitative
- Strict procedure
- Medium-high # participants



Make sure humans can perceive the difference in your different conditions

Subtle differences are NOT detected by humans (definitely not “in the wild”)

Study Type: RCT

Goals

- Validate your solution.
 - Measure effectiveness of a (very specific) solution.
-

Characteristics

- Compare solution to a control group (common intervention / no intervention / placebo).
- Summative
- Quantitative
- Very strict procedure.
- High # participants

**RCT = Randomized
Controlled Trial**

Study Type: Hybrid / Mixed-Methods

Goals

- Multiple goals
 - Save time and resources
-

Characteristics

- Mix of other study types

Study with “Autonomous” Robot

robot not controlled by human, experimenter only starts and stops the robot

robot behavior has been *fully automated*



Single or Repeated



how will people interact with robots **in day-to-day life** and what are the technical, societal and psychological consequences

a single short exposure of a user to a robot may yield a result due to the **novelty effect** (a user's unfamiliarity with robots)

Participants

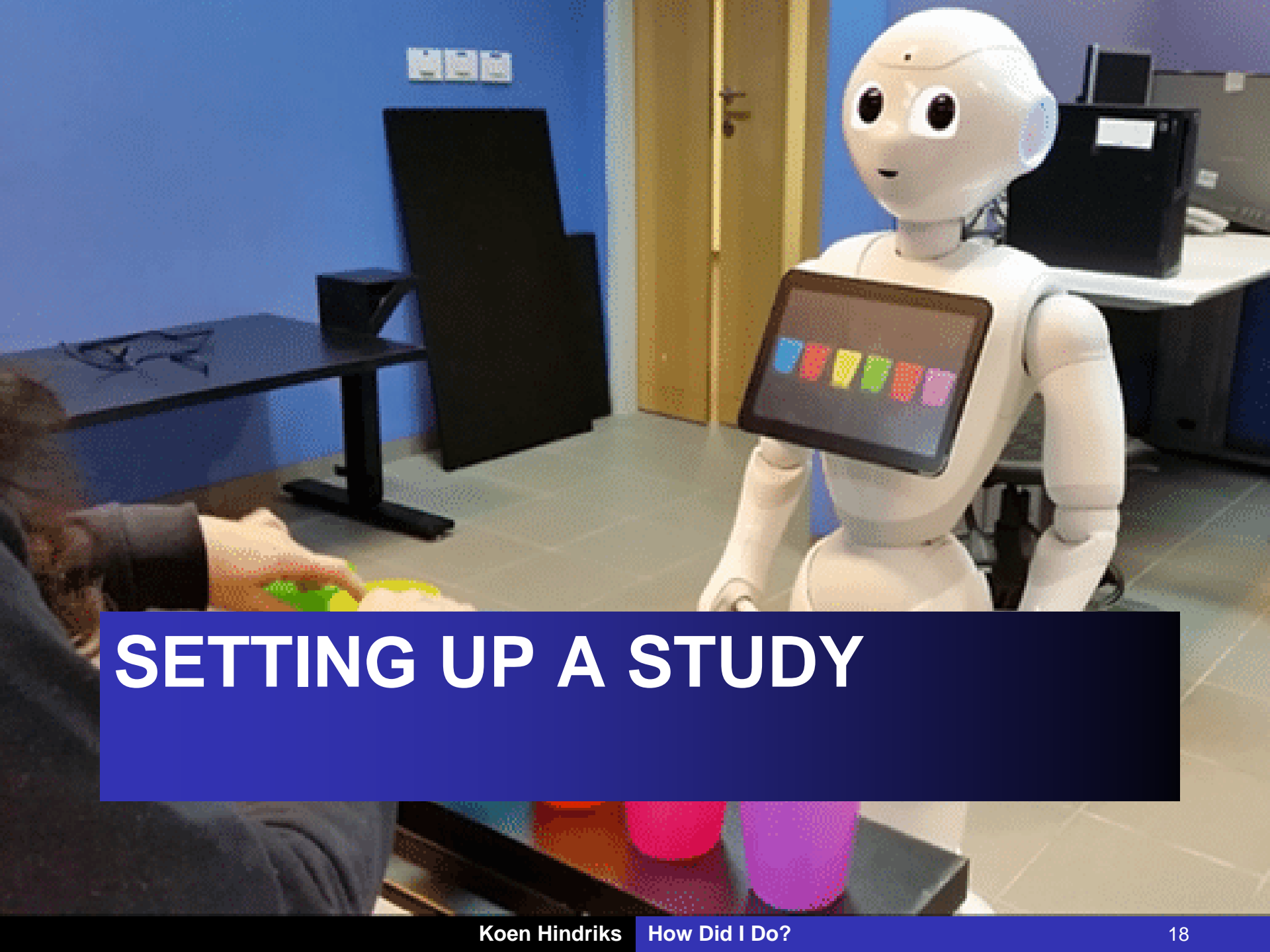
- Who, why, where, when
 - Informed consent
 - Recruitment and preparation
 - Amount
-

This course:

- *convenience sample*: members of other group

This Course

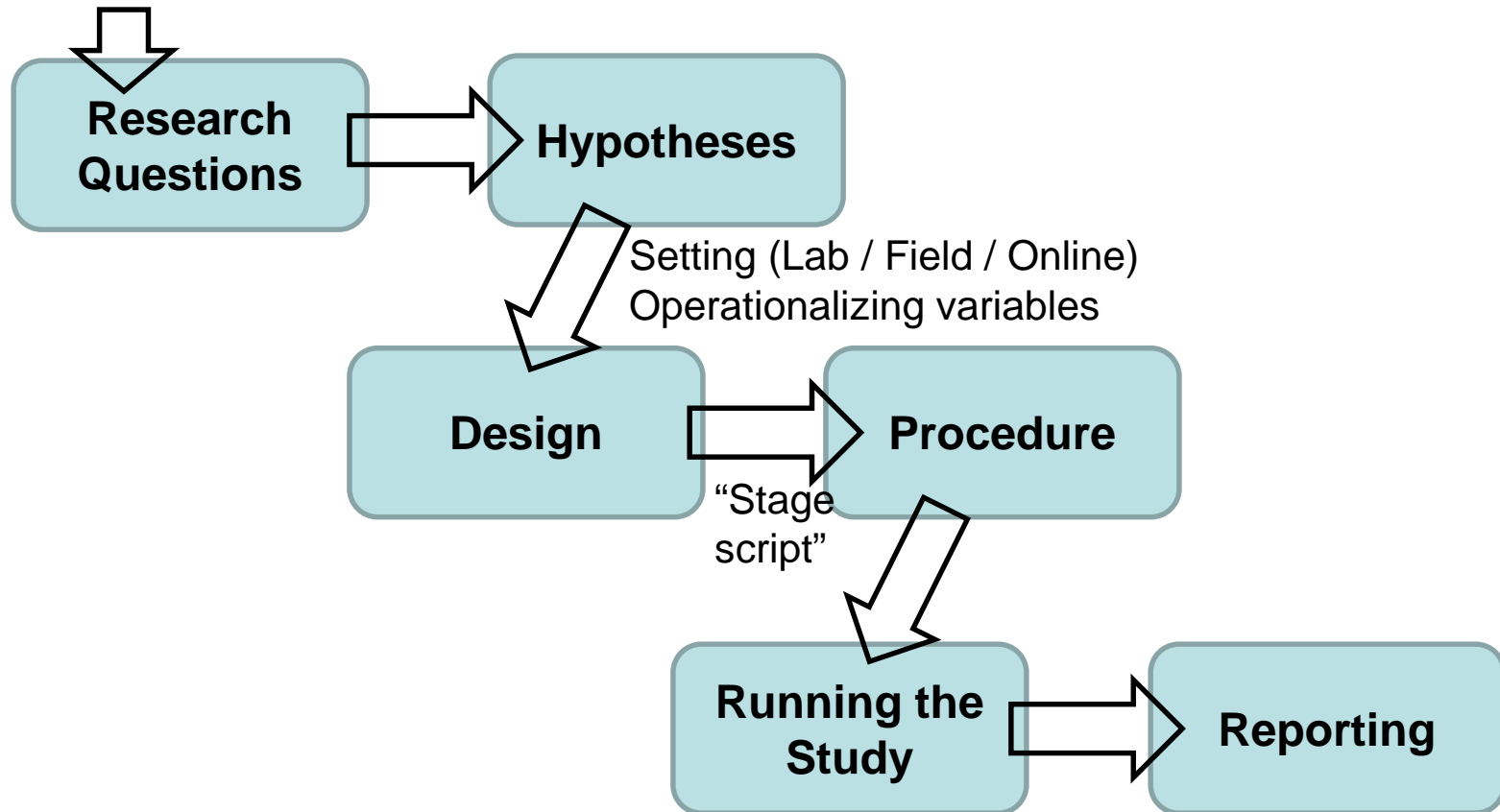
- *Qualitative*: You'll only have a few participants
- *Lab Study*: Your robot will be evaluated by another group during a session on campus
- *Pilot / Exploratory / Comparison*: Few participants
- *Fully automated*: your code should control the robot
- *Single*: experiment will be performed once (week 7).



SETTING UP A STUDY

Overview

Foundation:
Theory / literature



Research Question

Relate to:

- Problem statement
- Design objectives
- Use case objectives
- “kill your darlings”
- SPECIFIC (should be feasible to evaluate)

Formulate a **hypothesis** based on:

- Claims
- Theory (human factors knowledge)

A hypothesis is a **statement**.

Needs a *fair baseline*.

Method: Design

Independent variables

- Variables you control
- One, two, or more?
- How many levels?

Dependent variables

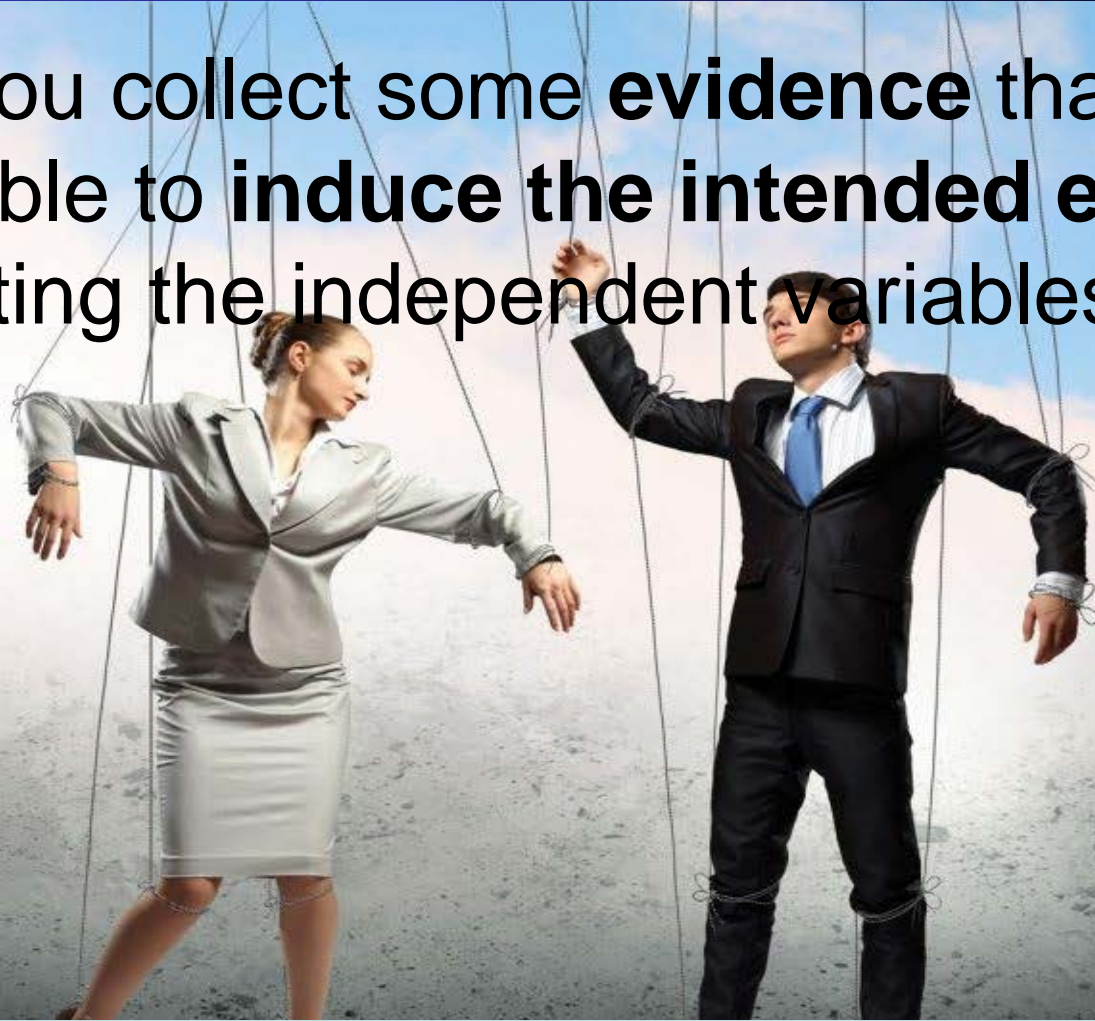
- Things you measure
- One, two, or more?
- Ratio / ordinal / nominal?

Settings

- Between- vs within-subject
- One vs. repeated measures
- Randomization
- Counter balancing
- Balancing user characteristics (e.g. gender, age)

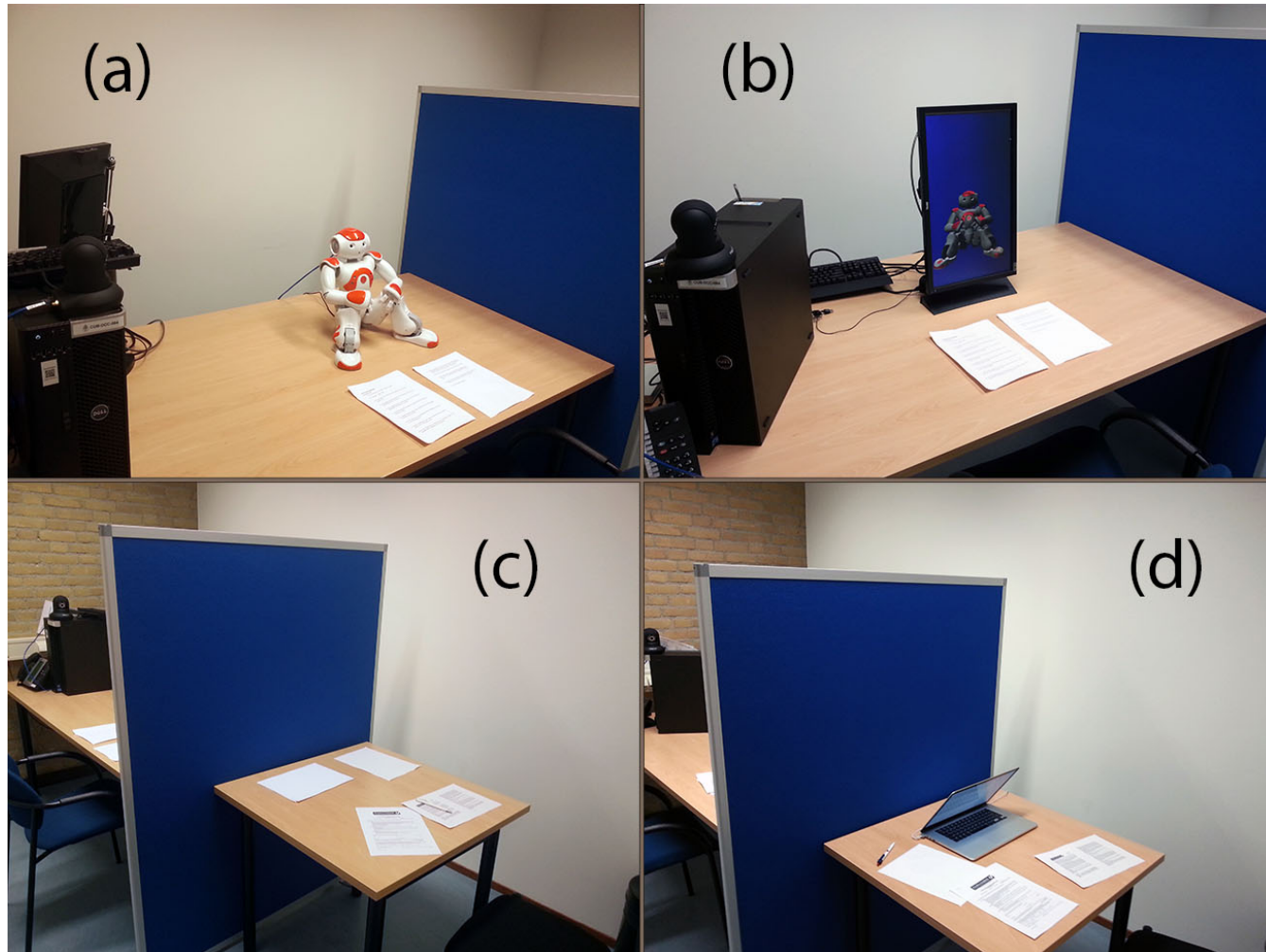
Manipulation Check

Make sure you collect some **evidence** that you have been able to **induce the intended effect** by manipulating the independent variables.



- Simple in case of objective manipulations (e.g. height of a robot, etc.)
- Important when you aim to manipulate psychological states or styles, etc.

Method: Materials & Setup



Method: Procedure

Procedure

=



a **detailed** description of all the **steps** to be performed during the experiment.

From the moment that a participant is collected to the moment he/she is exiting the experiment.

be **explicit** about every aspect of your procedure
→ two experimenters should know how to run the same procedure

Method: Procedure

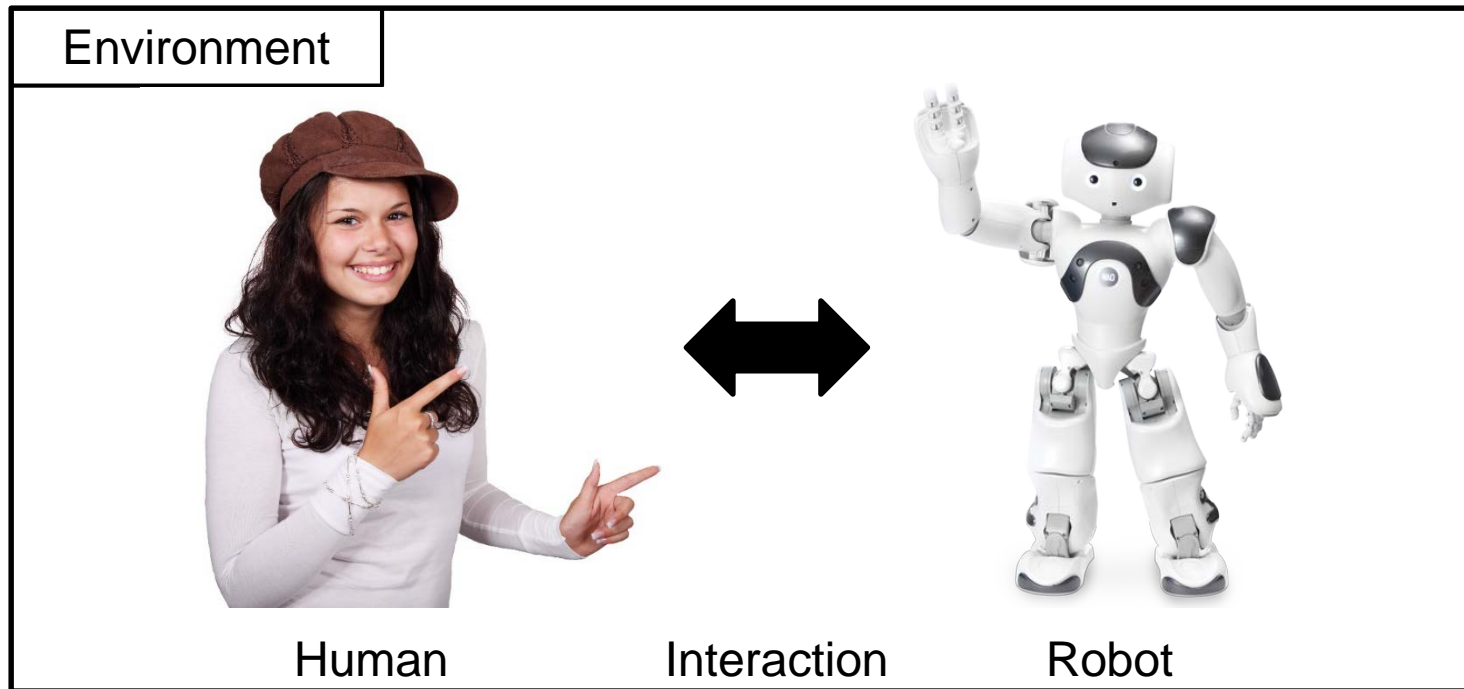
Should also provide a **protocol** for:

- Obtaining **consent**
- **Introducing** the task

Instructions for the experimenter

- How to observe / act during experiment?
- (Securely) **storing participants' data** after they complete your study
- Handling situations that deviate from your plan (e.g. an uncooperative participant, or a malfunctioning device). What is your **contingency plan**?

Measures & instruments – What



- Demographical data
- User experience
 - Effectiveness
 - Efficiency
 - Satisfaction
- Psychological constructs
 - Social aspects
 - Behavioral aspects
 - Cognitive aspects
 - Perceptual aspects
 - Attitudinal aspects

Measures & instruments – How

Observations

- Human / automatic
- Present / remote
- Direct / delayed
- Instruct/train observers

System logs (e.g. text from STT)

Questionnaires

- Self-report
- Others (e.g. personality)

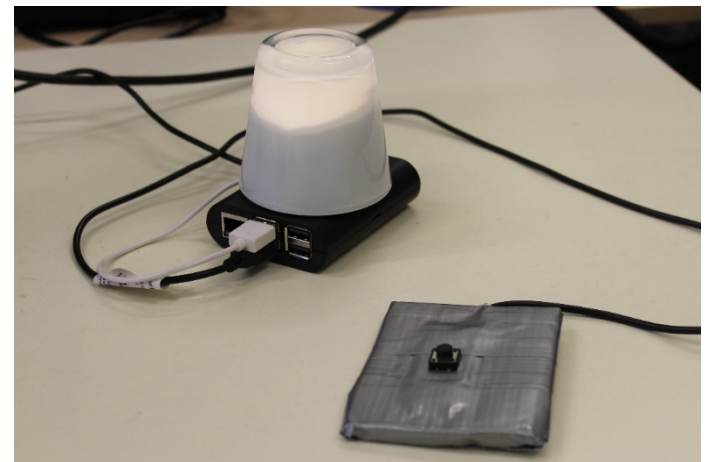
Interview (typically post)

Other

objective



subjective



Using Observation Sheet

Types of observation

- **Exploratory** (subjectivity): no scheme, just observe
- **Systematic** (blindness): coding scheme for classification of behavior

Coding schemes organize behavior, e.g., speech rate (high, low, normal) or type of interaction (listen, talk, ignore). Items should be **mutually exclusive** (only one applicable) and **exhaustive** (always one applicable)

Make a coding scheme (or use an existing one...):

- Focused: include only the necessary
- Objective: require as little inference as possible
- Explicitly defined: clear what does and does not fall within a category
- Easy to record (feasible): draw as little as possible on memory

Describe **detailed protocol** for making observations

Example Measures

- Use case specific performance measures:
 - Example EYE film museum: 2x2 condition; happy/neutral, video/questions first?
 - Which % of passersby interact with robot?
 - Do people rate happy Pepper better than neutral?
- Quality of interaction:
 - How did speech interaction go? *Count # failures*
 - How well did robot in detecting people? *Precision / recall*
 - How often was happy flow completed? *Percentage*
 - How did users rate the robot? *Multi-item Likert scale*

A RQ needs a measure

RQ: To what extent, if any, does the robot companionship improve the elderly user's morning routine quality?

Measures for:

- robot companionship,
- morning routine quality.

Reading assignment example

A RQ needs a measure

RQ: To what extent, if any, would a NAO robot be more beneficial for exercising at home than video tutorials for work outs?

Good idea to make comparison explicit!

Measure for: 'more beneficial'?

Reading assignment example

Measure for: 'more beneficial'?

Ideas on how to measure 'more beneficial' for an exercising robot



ask post exercise survey
adaptive to new scenarios expected outcome
number of times exercise
self report example more engagement
increase in exercise time
number of interactions
heart rate more motivated
increased happiness
meeting the goal
track progress over time

A RQ needs a measure

RQ: Will home-office workers exercise more (and more frequently) with NAO as a personal trainer?

Measures for:

- frequency and duration of exercises

How can you evaluate this in a pilot?

Reading assignment example

How evaluate use frequency in pilot?

Ideas for evaluating frequency of use in a pilot



number of times distracted

reminders by nao

send report result

asking out motivation

small exercises

A RQ needs a measure

RQ: Is a robot better at achieving cooperation from the subject than a human?

Measure for: cooperation?

Context could be more **specific!**

Reading assignment example

Ideas for operationalizing cooperation?

From the lack of responses it seems this is very difficult without providing any context.



Ideas for operationalizing cooperation

increase interaction

number of responses

A RQ needs a measure

RQ: Can using a robot for delivery assistance reduce the incidences of misplacing and stolen items?

How do you evaluate this in a pilot?

Reading assignment example

Evaluating 'reduce the incidences'

Seems very difficult without again this time because role of robot was not clear.



Proxy measure for 'reduce the incidences' that can be used in a pilot

persuasive messages

decrease number of lost pack

survey after every delivery

Note for this Course

Your main goal may be to establish a main effect (learning effect, robot is trusted, etc.).

However, this typically is very hard to measure (requires an elaborate experimental design)!

THEREFORE, consider using:

- qualitative self-reporting tools
- at least one measure related to the interaction with the robot itself (e.g., efficiency, effectiveness, satisfaction) that can provide concrete insights on your interaction design

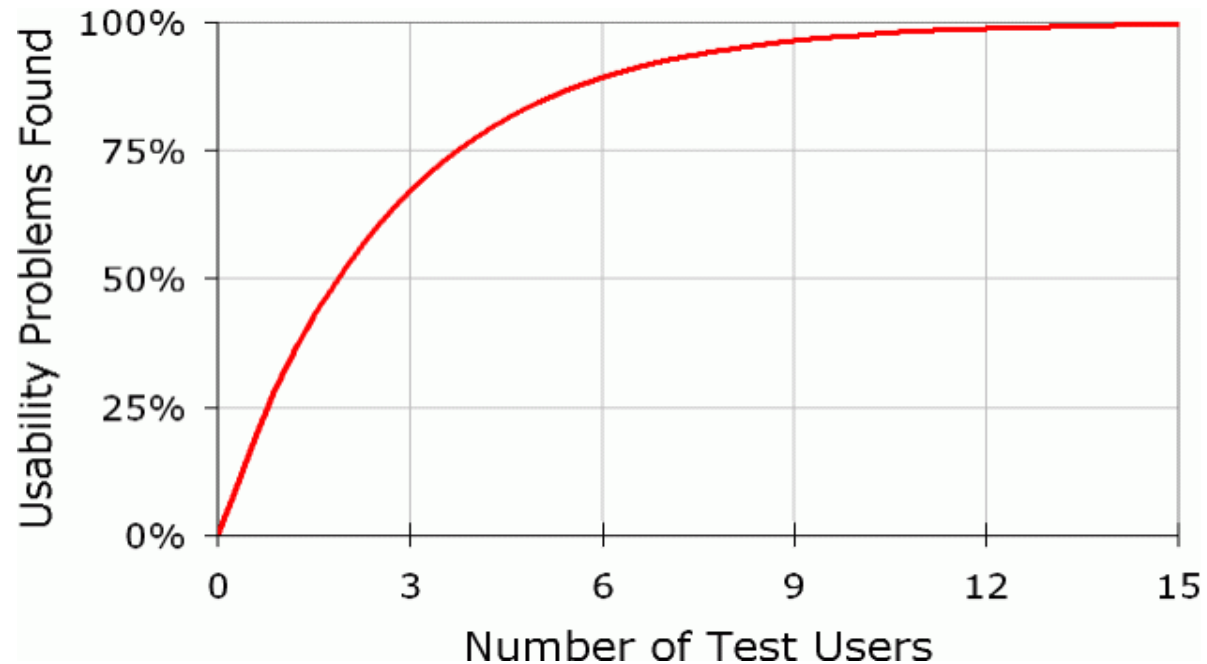
Usability Testing

As defined in ISO9241-11:

The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use

Aspects of usability:

- Usefulness
- Efficiency
- Effectiveness
- Learnability
- Satisfaction



Jakob Nielsen

Data about Humans

- We're collecting data about humans.
- Key characteristics of humans:
 - You can never predict them.
 - They always do things you didn't anticipate.
- Interesting to learn from but “messy” data
- You can ask people to follow a procedure (by upfront instructing them), but that's not collecting data “in the wild”



REPORTING YOUR STUDY

Result Section in Design Doc

Factual.

- No or very little interpretation.
 - No speculation(!)
-

Results and analysis method

Metrics and (descriptive) statistics

- Cover your assumptions (for statistical tests).
-

Tables and figures.

Discussion Section in Design Doc

- Argue convincingly for interpretation of results
- Lessons learnt
- Limitations

A Child and a Robot Getting Acquainted

Interaction Design for Eliciting
Self-Disclosure



FIRST STEPS IN SOCIAL INTERACTION GETTING ACQUAINTED

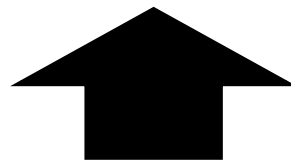
Based on: Mike Lighthart, Timo Fernhout, Mark A. Neerincx, Kelly L. A. van Bindsbergen, Martha A. Grootenhuis, and Koen V. Hindriks. 2019. A Child and a Robot Getting Acquainted – Interaction Design for Eliciting Self-Disclosure. In Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019, 10 pages.

Motivation

Long-term = repeated social interaction



Engaging → Relationship formation / bonding



First step = Getting acquainted

Getting Acquainted Interaction

1. Child gets acquainted with robot
 - a. The child *learns how to communicate* with the robot effectively
 - b. The child gets *familiar* with the robot
2. Robot gets acquainted with child
3. Relationship formation / bonding is initiated

How do humans get acquainted?



- Unstructured dyadic interaction
- Social norms:
 - Mutual self-disclosure

Literature: Intro- / Extraversion

“Personality similarity resulted in relatively good initial interactions for dyads composed of 2 extraverts or 2 introverts, when compared with dissimilar (extravert-introvert) pairs”

Ronen Cuperman and William Ickes. 2009. *Big Five predictors of behavior and perceptions in initial dyadic interactions: Personality similarity helps extraverts and introverts, but hurts “disagreeables”*. Journal of personality and social psychology 97, 4 (2009), 667

Socio-psychological influences of personality dimension suggests we need to **match extraversion dimension**.

Effect related RQ

To what extent, if any, will extraversion matching of a robot with a child improve initial interactions?

A RQ needs a measure

A number of very good suggestions 😊



Measure for improving initial interactions

time spend with the nao

length of the answers

translate the persona of

personalized answers

jokes

friendly introduction

asking questions

number of responses

uses name

Measure operationalized

“The amount of self-disclosure is operationalized as the total count of unique statements related to oneself within all the responses made by a participant. The annotators marked and counted the unique statements per response. Summing these statements resulted in the total amount of self-disclosure per participant. To summarize the instruction set, **every part of the response that is or could syntactically be separated by either a comma or an ‘and’ should be counted as a unique statement.** For example, “I always wanted to have a cat” counts as one and “I like to play football and tennis” counts as two. An exception however is when two parts of a statement belong to the same concept. For example, “My favorite TV-show is Tom & Jerry” counts as one.”

Use measure in hypothesis

- H1a: extraverts self-disclose more to an extravert robot
- H1b: introverts self-disclose more an introvert robot

Interaction Design

Design with a focus on **self-disclosure elicitation**

- *[Unstructured conversation vs. autonomous CRI]*
Design choice: structured dyadic interaction
Meaning in our case that robot will be driving the conversation.
- *[Social norms: reciprocation]*
Design choice: robot “self-disclosures”
- *[Psychology: Extraversion matching]*
Design choice: Behavior adaptation to extraversion trait

Interaction flow / script

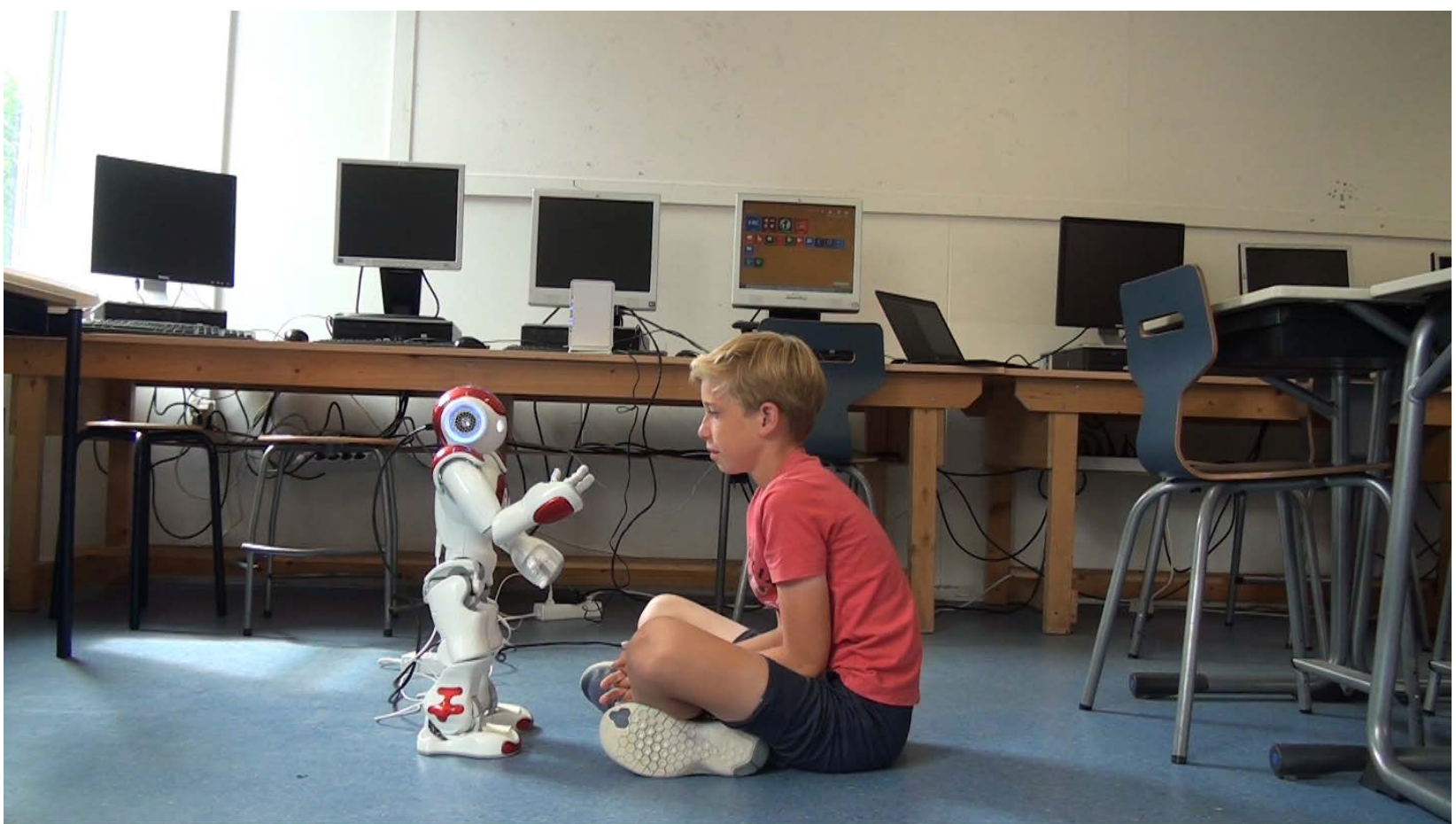
1. Robot takes initiative and asks closed-ended / pseudo-open-ended question
2. Child answers
3. Robot responds to child's answer
 - a. Backchannel and/or
 - b. Robot disclosure
4. Robot asks open question
5. Child answers
6. Robot acknowledges answer

Literature: Child ASR

“Using the data collected we demonstrate that there is still much work to be done in ASR for child speech, with interactions relying solely on this modality still out of reach. However, we also make recommendations for childrobot interaction design in order to maximise the capability that does currently exist.”

James Kennedy, Séverin Lemaignan, Caroline Montassier, Pauline Lavalade, Bahar Irfan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2017. Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations. In Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17). ACM, 82–90.

We need to design for ASR failures, i.e. **repair**.



TOUCH-BASED RECOGNITION AND REPAIR PIPELINE

Interaction Design related RQ

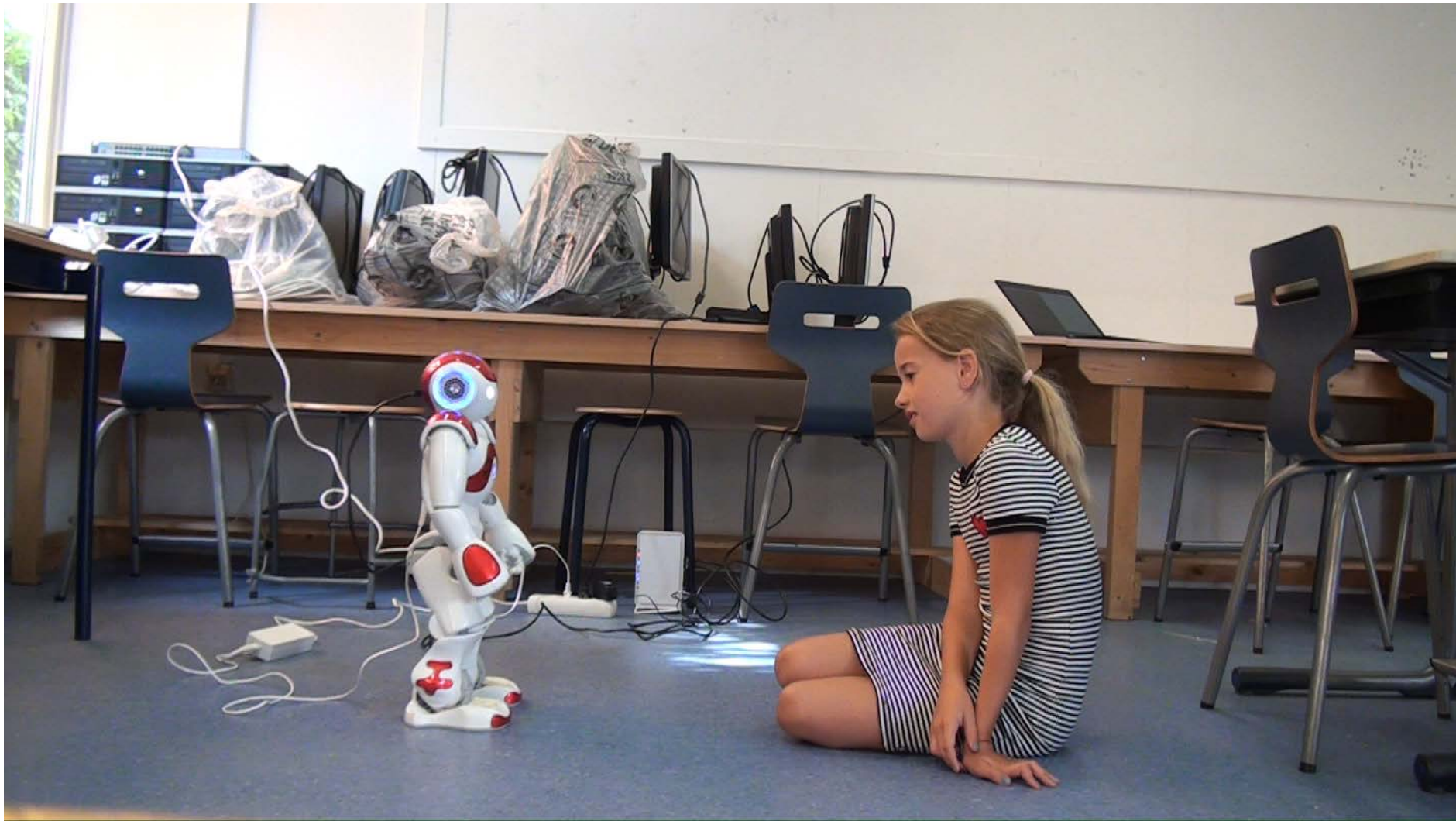
How successful is the recognition and repair pipeline and is the touch-based mechanism an effective alternative?

A RQ needs a measure

- Measure: recognition performance.

Extraversion adaption

Behavior setting	Extravert	Introvert
Speech speed	100%	90%
Speech volume	49	40.5
Language style	Directive	interrogative
Emotion words	Strong	weak
Speech activity detection interval	2-3s (100%)	2.5 -3.75s (125%)
Gestures amplitude	100%	60%
Gesture speed	100%	50%
Head movement speed	100%	75%
Breathing animation	20 bpm	10 bpm
Activity order	Dance – game	Game - dance



Evaluation

To what extent, if any, will extraversion matching of a robot with a child improve initial interactions?

How successful is the recognition and repair pipeline and is the touch-based mechanism an effective alternative?

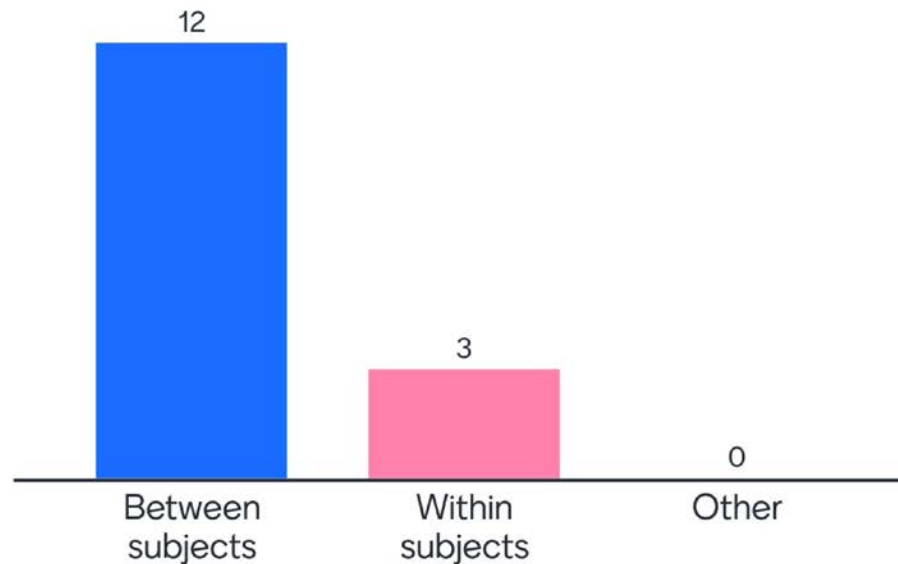
User study: Participants

- N = 75
- 8 – 11 y.o.
- 45 girl – 30 boys
- 4 classes from 2 Dutch primary schools

User Study: Design



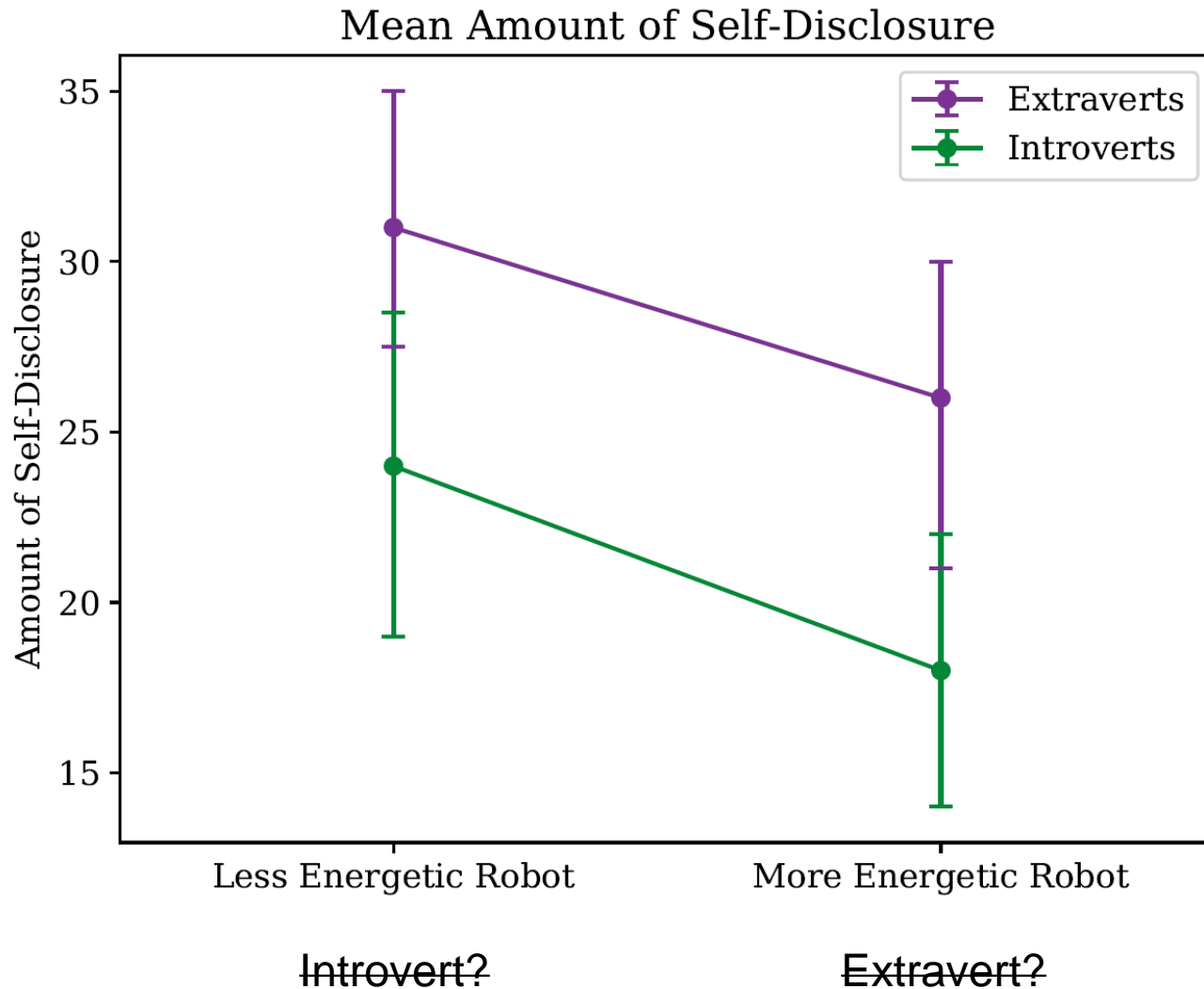
What study design would you choose for the getting acquainted study?



Study design we used

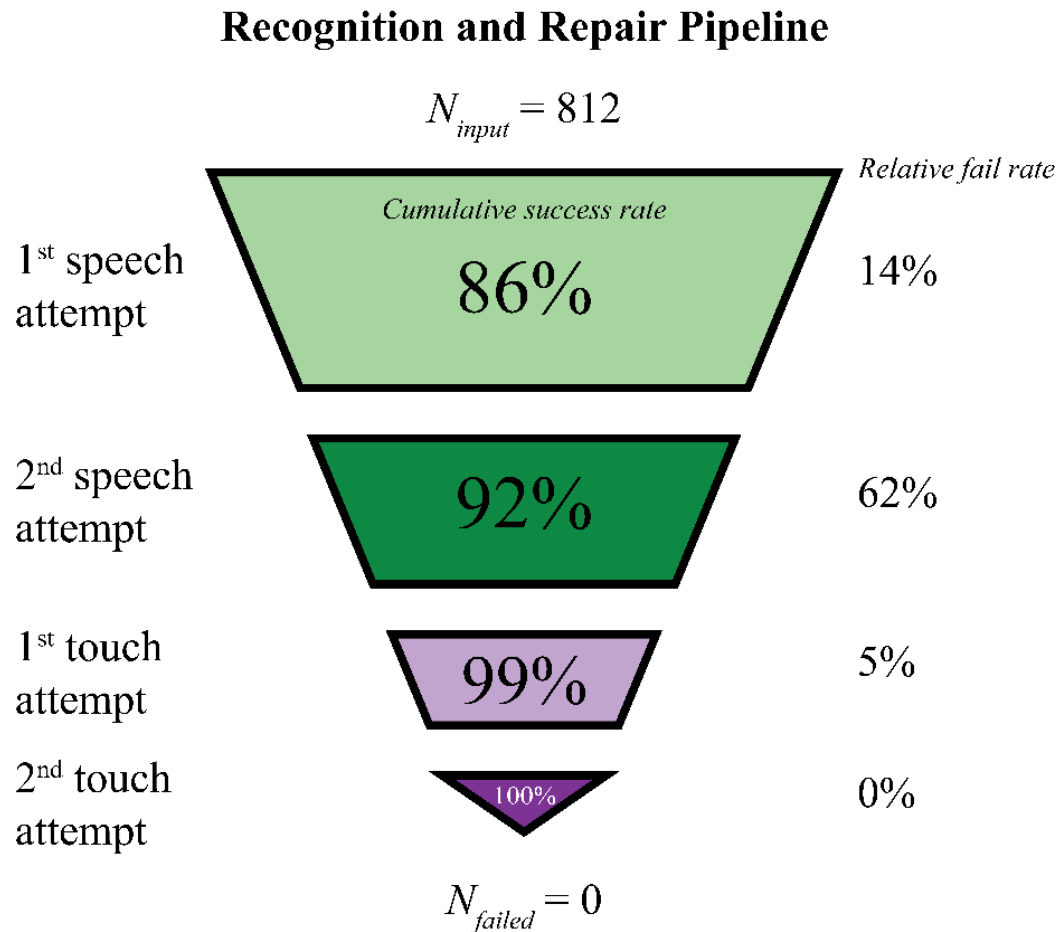
- 2x2 between-subject design
- Variables:
 - Independent: child's intro / extraversion
 - Independent: robot's intro / extraversion adaptation
 - Dependent: amount of self-disclosures
- Balanced age, sex.

Extraversion and self-disclosure



Structured Dyadic Interaction

How successful is the recognition and repair pipeline and is the touch-based mechanism an effective alternative?



Structured Dyadic Interaction

How successful are the different questions in eliciting self-disclosure?

Type	#	Response rate
Closed-ended	54 2	98%
Pseudo-open-ended	28 5	99%
Open-ended	53 3	88%

Do children give valid (i.e. pre-specified) answers to the pseudo-open-ended (and closed-ended) questions?

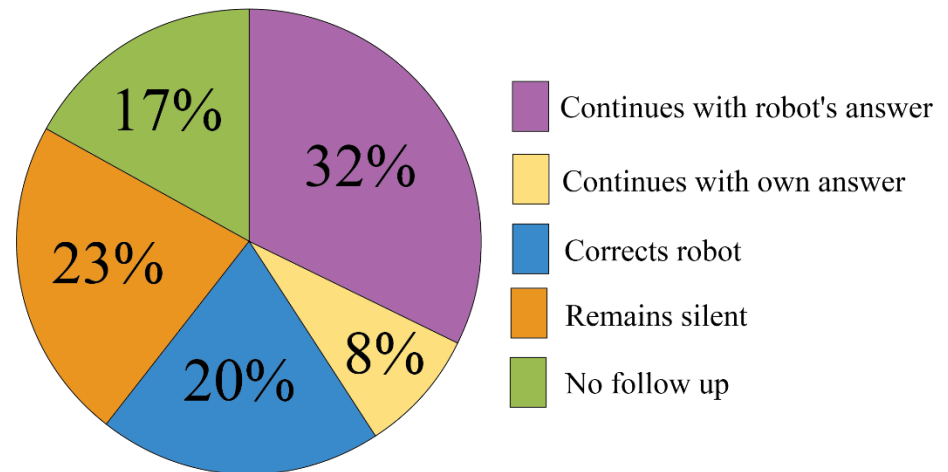
Type	#	Response rate	Valid
Closed-ended	542	98%	97%
Pseudo-open-ended	285	99%	95%
Open-ended	533	88%	n/a

Structured Dyadic Interaction

How often is speech incorrectly recognized and how do children respond to those mistakes?

8.7%
*Incorrectly recognized
speech*

Responses to incorrect speech recognition



Conclusion HERO Study

- First steps towards an autonomous social robot that can repeatedly engage with children.
- Recommend focusing on lower-level behavior aspects of the interaction than high-level and convoluted psychological constructs.

Summary

- *Qualitative*: You'll only have a few participants
- *Lab Study*: Your robot will be evaluated by another group during a session on campus
- *Pilot / Exploratory / Comparison*: Few participants
- *Fully automated*: your code should control the robot
- *Single*: experiment will be performed once (week 7)