



# Evaluation methods in HAI

Joost Broekens

# In this lecture we cover



## Common evaluation goals

What do we  
want to *know*  
regarding the  
interaction with a  
social robot?



## Common analysis methods

How do we  
*analyze* our study  
outcome data



## The most commonly used evaluation methods

How do we  
shape the  
evaluation  
*procedure* to get  
to know this?

# A guiding example from practice

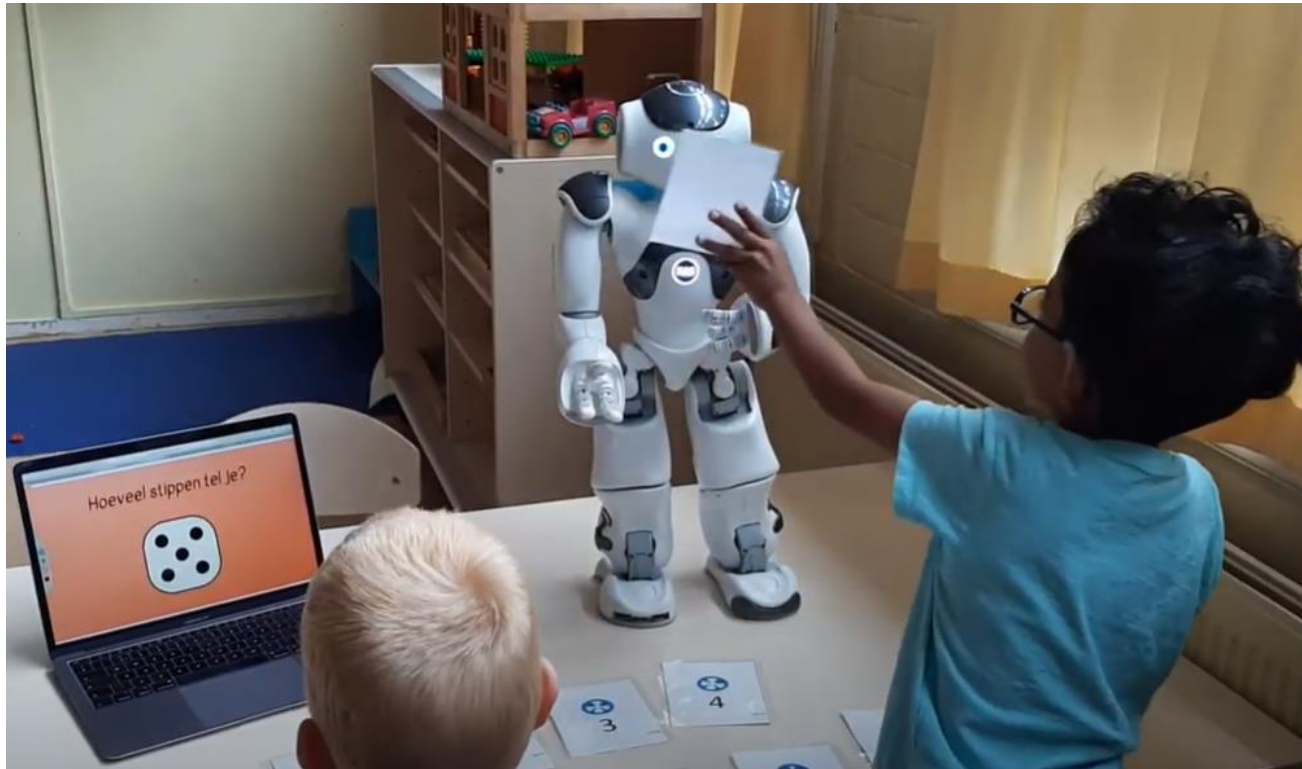
- Design goal:
  - A math robot for children aged 5-6 to learn counting and numbers until 10.
- <https://youtu.be/YGYloaGvW-s>
- Watch the video



# Evaluation goals: things we want to know

- Robustness:
  - Does the robot work from a technical point of view?
  - In the book called: robot centered evaluation, system study
- Usability:
  - Do the users interact with the robot as intended?
- Perception:
  - (How) do users perceive (behavior of/interaction with) robot as intended?
- Impact:
  - (affective) How is the user's emotional/attitude/mood impacted?
  - (cognitive) How is the user's thinking influenced?
  - (behavioral) Does the interaction change the user's behavior?

# Evaluation goals: robustness



- Is the robot stable enough to not fall during the interactions?
- How often does the robot fail to detect the face of the child?
- Is the math application running without bugs?
- Are the answer card detected fast and robust enough?

# Evaluation goals: usability



- Do the children understand how to hold the card?
- Are the children afraid to hurt or damage the robot?
- Is the flow of interaction between robot and screen clear for the children?
- Do they know how to “correct” the robot when it fails to understand the input?

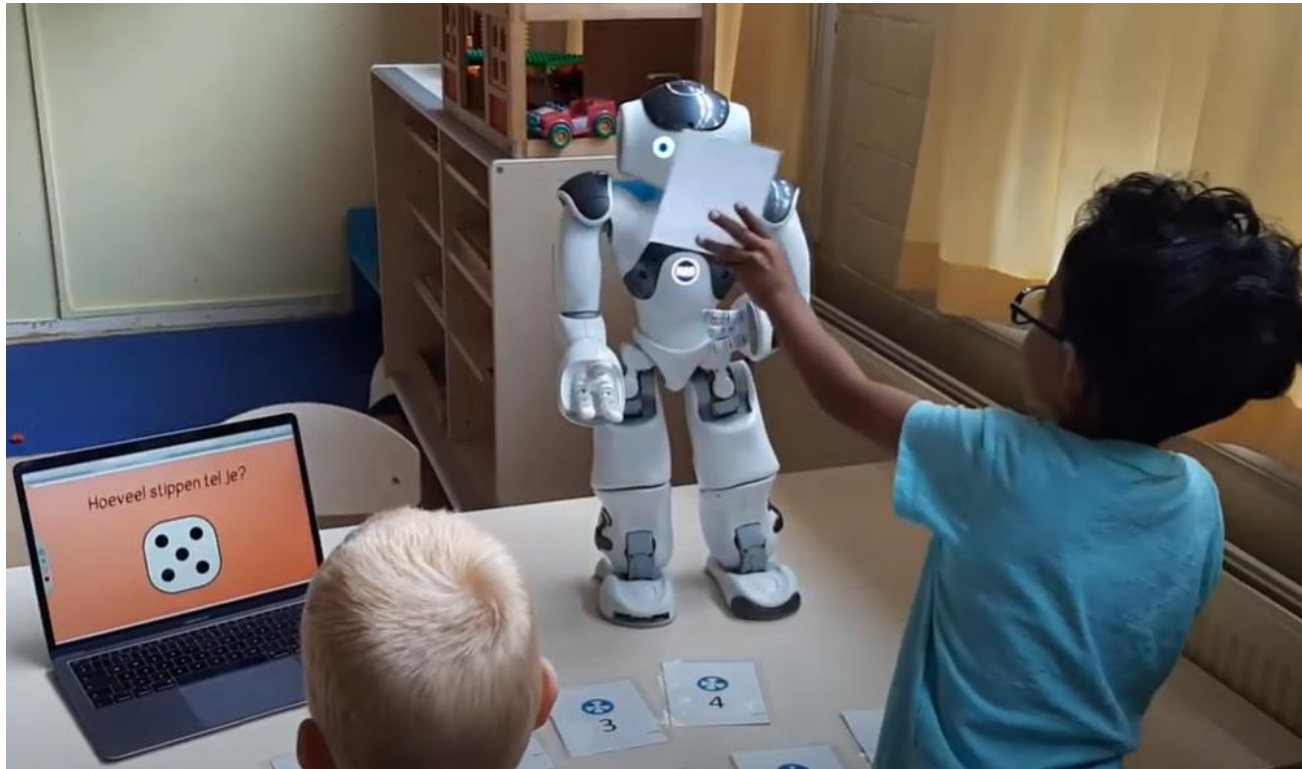


# Evaluation goals: perception



- Do the children perceive the robot as a friend or as a teacher?
- Do they perceive the robot as alive or inanimate?
- Do they like the looks and behavior?
- Is the robot perceived as smart or stupid?
- How much do they like the task?

# Evaluation goals: impact (effects)



- Do children learn to better count after a session with the robot?
- Does their attitude towards robots change?
- Are they more motivated to work on math problems after using the robot?
- How is their attention distributed (robot, screen, cards, other kid)?



# Evaluation methods: main dimensions

- Type of research question
- Type of outcome data
- Study type (not the same as study design in the book!)
- Mode of interaction
- Location context
- Temporal context
- User involvement

# Evaluation methods: research question type

- Exploratory

- Open questions
  - We do not know enough to formulate hypotheses...
- *How do children react to the math robot after repeated sessions?*



- Confirmatory

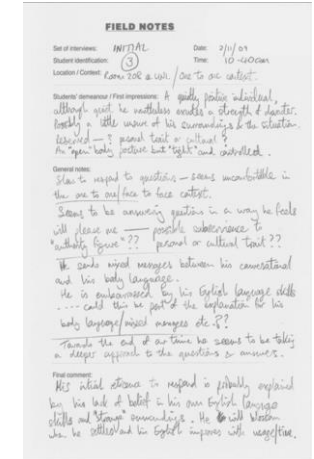
- Closed question hypothesis
  - Need background knowledge to make the assumption!
- *Children get better at counting after one session with the math robot!*



# Evaluation methods: outcome

- Qualitative measures

- Interviews, observation notes, opinions.
  - "we observed that the child hesitated answering the question when the robot looked away erroneously."
  - Sometimes you can *count* or *classify* the data, allowing some form of quantization.



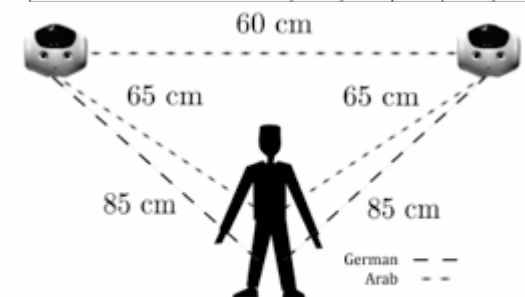
- Quantitative measures

- Subjective data, such as ..
  - A survey that measures perception...
  - ...the robot is your friend (1=not at all --- 5=very much), average rating = 2.3
- Objective data, such as...
  - Distance between kid and math robot in cm, avg distance = 50cm
  - Or, math skill as measured by the performance on a post test.

1 Access

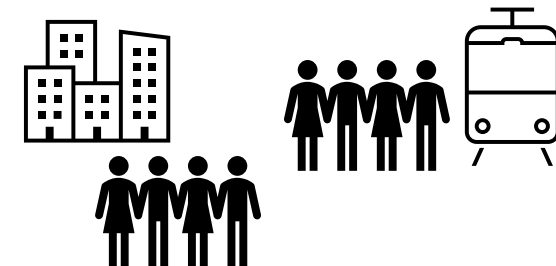
How do you personally evaluate the importance of the following aspects of coordinated care?

How important is.....?	very important	important	so-so	less important	not important
The surgery hours of the doctor/service provider are flexible.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The doctor/service provider takes a proactive approach with me (far-sighted, preventative) and agrees check-up appointments or reminds me that an appointment is due.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My doctor/service provider is available around the clock in case of emergencies.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The health insurer actively supports me.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The interfaces between GP, specialist and hospital are not perceptible to me. There are no problems at these interfaces in the case of referral from GP to specialist or admission to hospital (loss of information)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



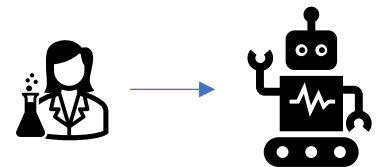
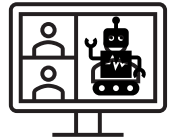
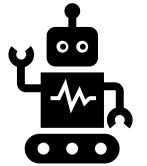
# Evaluation methods: study type

- Case study (book: single-subject study)
  - You investigate one (or several) individual(s) and analyze/report your findings in detail, usually the reason is because the individual is of particular interest.
  - Example: you want to know how a typical boy with ADHD reacts to the math robot.
- Classical experiment (book: user study)
  - You investigate how variation in factors (independent variables) influences your outcome (dependent variables)
  - Example: what is the effect of support gestures (yes/no) on learning outcome in a counting education robot, 30 participants (pp) with, 30 pp without.
- Observational/ethnographic study
  - You “unleash” a robot in a setting and observe and report what you see.
  - Example: you want to know how people at a train station react to a station-guide pepper.



# Evaluation methods: mode of interaction

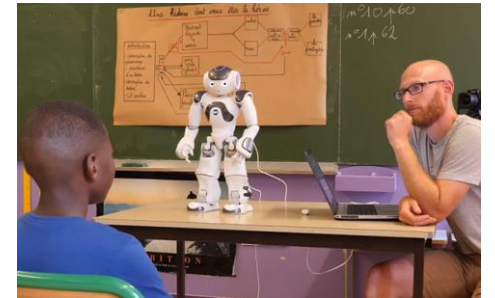
- Autonomous robot interaction
  - The interaction is with a real robot including the intended functionality
- Simulation study
  - The interaction is not with a real robot but with either a virtual robot or videos of the interaction (images should be avoided)
- Wizard of oz study
  - The interaction is (partly) controlled by a human operator, unknown to the participant.





# Evaluation methods: location context

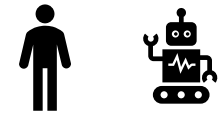
- Lab study
  - When the interaction is performed in a lab setting.
- Field study
  - When the interaction is in the real world, in the intended usage setting.
- Crowd sourced study
  - When the interaction is at home using e.g. Amazon MTurk, while the intended usage typically is not (e.g. in a simulation study with video's requiring feedback from Mturkers as participants).



# Evaluation methods: temporal context

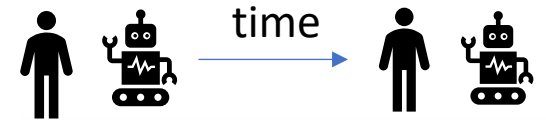
- Single interaction

- Where the interaction consists of a single session



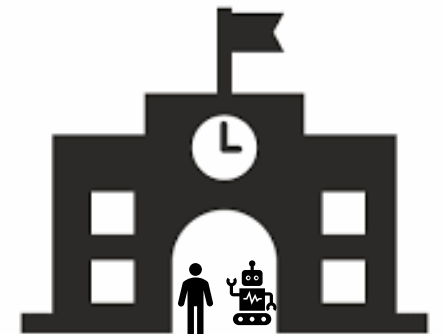
- Repeated interaction

- Where the interaction consists of several repeated sessions, potentially with some time in between.



- Long-term interaction

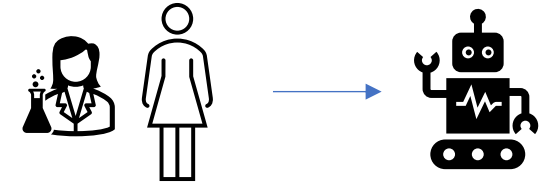
- Where the interaction can continue for a longer time (weeks, months) and sometimes is at the initiative of the user.



# Evaluation methods: user involvement

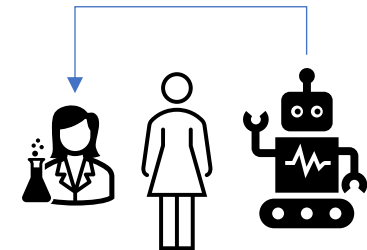
- Before the design

- Users provide requirements (see slides user centered design as well)



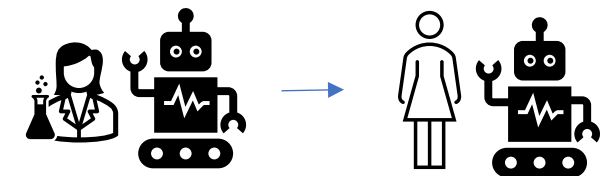
- During the design

- Users shape the design in iterations or through participation (see slides participatory design as well), or evaluate a prototype or do a pilot.



- After the design

- Users evaluate the robot system or are used as participants in a research study



# Exercise: method dimensions



- Assuming we want to observe, as a pilot study, how several children react to the robot...
- Can you identify choices made for dimensions in our example child – robot setting?
  - Type of research question
  - Type of outcome
  - Study type
  - Mode of interaction
  - Location context
  - Temporal context
  - User involvement

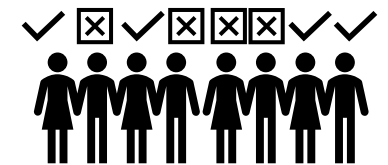
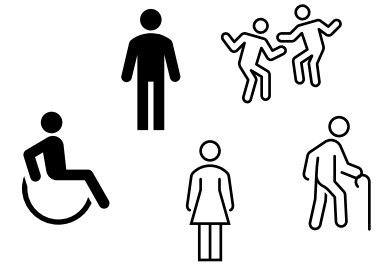
# Analysis methods: main aspects

- Participants
- Quantitative analysis: statistics
- Qualitative analysis: coding and annotation



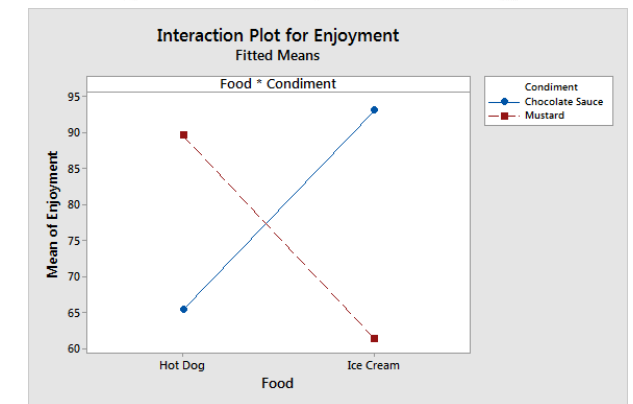
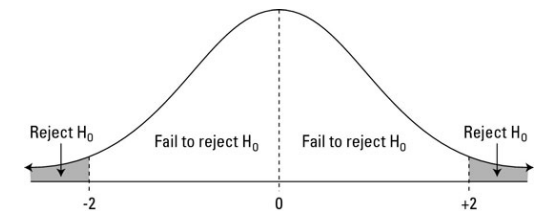
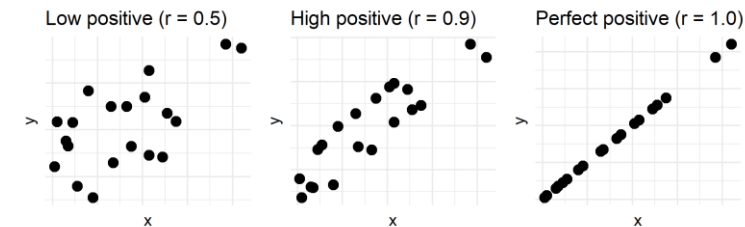
# Analysis: Participants

- The demographics of your participants/subjects
  - Such as age, gender, culture, education level.
- The level of familiarity with robots/HRI
  - Novelty and robot-fear are important factors to consider
- Generalization and participant selection
  - For case studies you need prototypical users.
  - For experimental effect studies you need larger numbers of representative users, so that you can generalize.

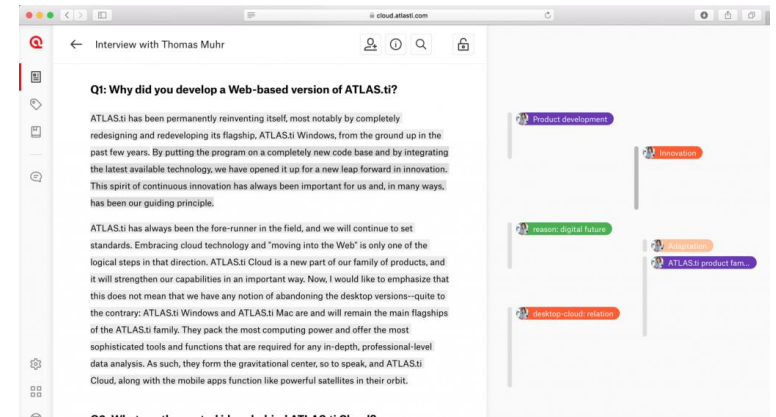


# Analysis: statistics

- Basis are your outcome variables as numbers and conditions as classes
- Correlations: to show co-occurrence of phenomena
  - E.g.: children like social robots better than adults
  - Correlate age with perception of likability.
- T-tests and Anova's: to show causality in classical experiments
  - E.g.: supportive gestures increase math learning
  - T-test with independent variable "gesture type" and dependent (outcome) variable "counting accuracy".
- Interaction effects: to show how the effect of one variable depends on another
  - E.g.: A teacher robot increases cognitive outcome more than a peer robot, but only for children with high IQ
  - Two-way Anova with two IQ bins and robot role as independent variables and cognitive outcome (some post test) as dependent (outcome) variable.
- And of course: box-plots, scatter plots, interaction plots, etc...



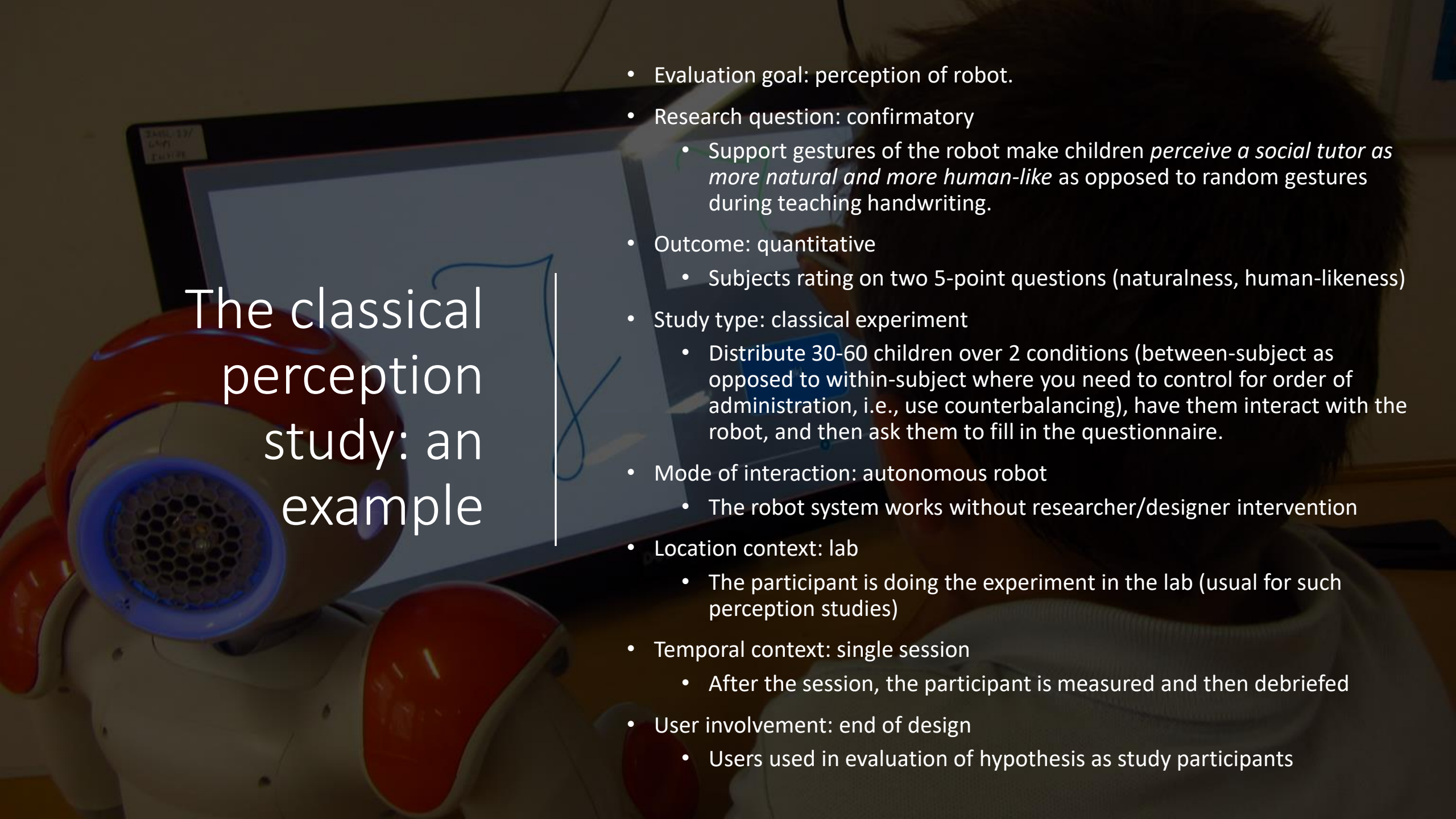
# Analysis: qualitative



- Basis is recorded video and audio and first-person observation and interview notes.
- Quotes from users
  - Used as evidence that participants expressed an opinion, concern or requirement
  - E.g. child2 said “I don’t like the way the robot looks at me, it makes me scared”, as evidence for the claim that “Robot attention to the robot can be a source of fear for children”
- Coding schema’s
  - Used as structure to annotate qualitative data
  - E.g. “I don’t like the why the robot looks at me, it makes me scared”, can be coded as “I don’t like [attitude] the way the robot looks at me [attention], it makes me scared [emotion]”
- Counting of occurrence of coded concepts
  - Used as evidence of importance or difference of occurrence in different conditions.
  - E.g. children preferred the non-staring over the staring robot, as evidence by a significant higher number of reported negative emotions for the staring condition.
- Conversation analysis.
  - Used to gain detailed insight into the multimodal interaction flow, including turn-taking, attention, verbal and non-verbal utterances, usually through annotation of interactions
  - E.g. children’s eye and head movements always follow that of the robot’s gaze.

# Prototypical HAI experiments

- The classical perception study
  - Where a form or behavior factor of the agent is varied, and participants provide a subjective measure of some outcome variable to measure the effect of the variation.
- The pilot
  - Where a (or several) prototypical users are observed to either pretest a more elaborate study, or, to gain insight into the usability of a HAI system.
- The impact (effect) study
  - Where a form or behavior factor of the robot is varied, and participants perform a task to measure the effect of the variation.
- The participatory design study
  - Where a robot is designed together with intended users in design cycles. (see slides on design)

A child with dark hair is sitting at a desk, interacting with a robot. The robot has a white body, red accents, and a large blue circular sensor on its head. It is positioned in front of a laptop. The laptop screen shows a blue background with a white line drawing of a person. The child's hand is visible, pointing at the screen. The background is slightly blurred, showing a lab environment.

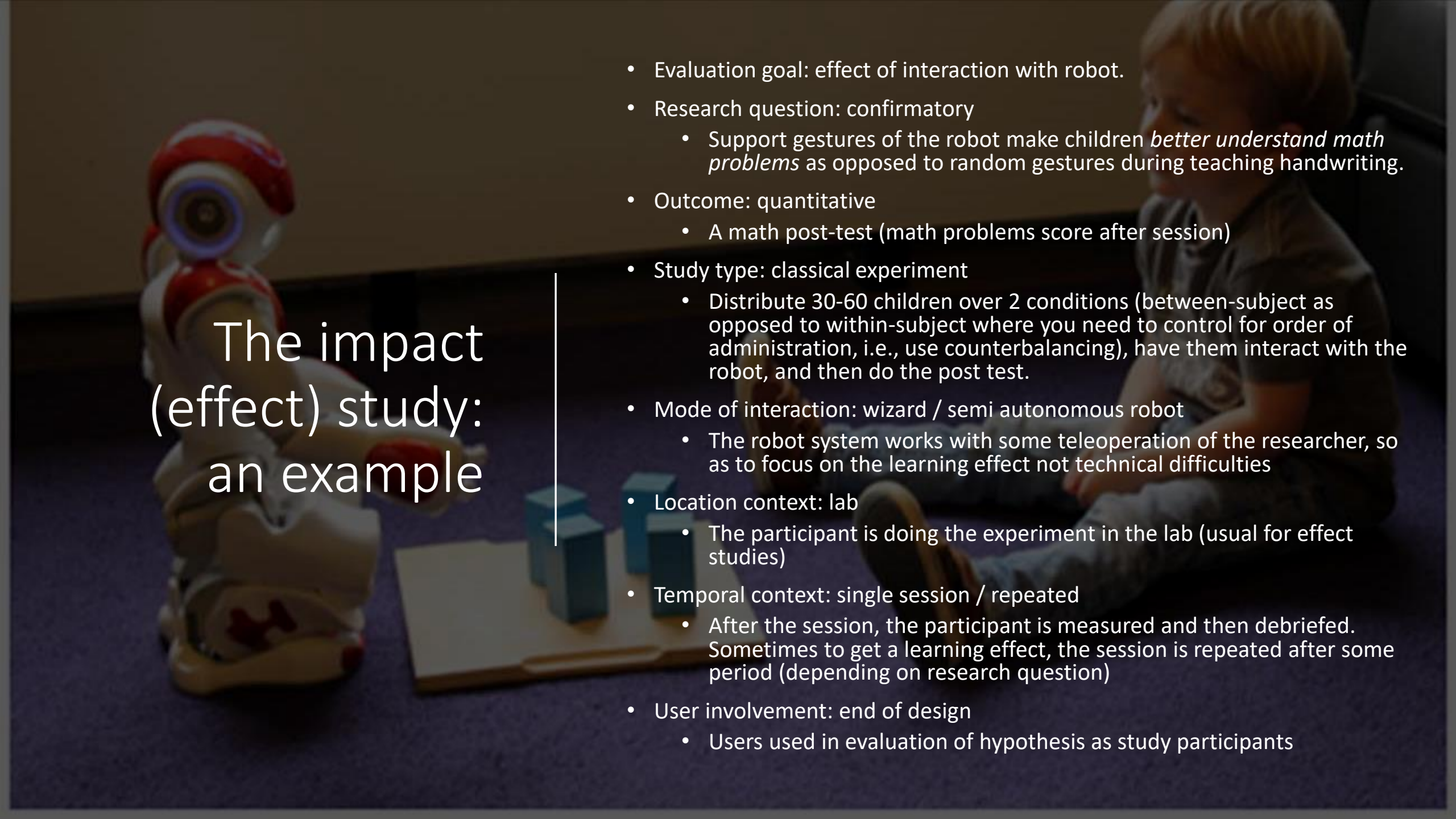
# The classical perception study: an example

- Evaluation goal: perception of robot.
- Research question: confirmatory
  - Support gestures of the robot make children *perceive a social tutor as more natural and more human-like* as opposed to random gestures during teaching handwriting.
- Outcome: quantitative
  - Subjects rating on two 5-point questions (naturalness, human-likeness)
- Study type: classical experiment
  - Distribute 30-60 children over 2 conditions (between-subject as opposed to within-subject where you need to control for order of administration, i.e., use counterbalancing), have them interact with the robot, and then ask them to fill in the questionnaire.
- Mode of interaction: autonomous robot
  - The robot system works without researcher/designer intervention
- Location context: lab
  - The participant is doing the experiment in the lab (usual for such perception studies)
- Temporal context: single session
  - After the session, the participant is measured and then debriefed
- User involvement: end of design
  - Users used in evaluation of hypothesis as study participants



# The pilot: an example

- Evaluation goal: usability, technical robustness
- Research question: open
  - How do children react to a math teaching robot, does the system work as intended.
- Outcome: qualitative
  - Observations notes, interviews and conversational analysis
- Study type: case study
  - Several “typical” kids are invited in a lab to play with the math robot and give their feedback after the session.
- Mode of interaction: autonomous robot
  - The robot system works without researcher/designer intervention
- Location context: lab/field
  - The participant is doing the experiment in the lab, or the researcher goes to the classroom (field)
- Temporal context: single session
  - After the session, the participant is measured and then debriefed
- User involvement: end of design
  - Users not used in design but to pilot the interactions



## The impact (effect) study: an example

- Evaluation goal: effect of interaction with robot.
- Research question: confirmatory
  - Support gestures of the robot make children *better understand math problems* as opposed to random gestures during teaching handwriting.
- Outcome: quantitative
  - A math post-test (math problems score after session)
- Study type: classical experiment
  - Distribute 30-60 children over 2 conditions (between-subject as opposed to within-subject where you need to control for order of administration, i.e., use counterbalancing), have them interact with the robot, and then do the post test.
- Mode of interaction: wizard / semi autonomous robot
  - The robot system works with some teleoperation of the researcher, so as to focus on the learning effect not technical difficulties
- Location context: lab
  - The participant is doing the experiment in the lab (usual for effect studies)
- Temporal context: single session / repeated
  - After the session, the participant is measured and then debriefed. Sometimes to get a learning effect, the session is repeated after some period (depending on research question)
- User involvement: end of design
  - Users used in evaluation of hypothesis as study participants

# Conclusion

- We learned about...
- Evaluation goals
  - Robustness, usability, perception, effect
- Evaluation method dimensions
  - Type of research question, outcome data, study type, mode of interaction, place and time context, user involvement
- Analysis methods
  - Participants, qualitative, quantitative
- Prototypical HRI studies
  - Perception study, pilot study, effect study, participatory design study

# Next week

- Multimodal interaction, social signal processing
- Gestures faces and posture