Human–Robot Interaction

Module 4: Research Methods in HRI

Detailed Study Notes | 9 hours

Source. Notes follow the Module 4 syllabus topics and are prepared from Bartneck et al., "Human–Robot Interaction: An Introduction" (Cambridge University Press), **Chapter 10 "Research Methods", Sections 10.2 to 10.8** (the second-edition numbering of the syllabus's "9.2 to 9.8"). Figures are reproduced from the chapter with their original numbers. Section 10.1 is included as background because 10.2 builds on it.

Syllabus coverage map

Syllabus topic	Section	Part in notes
Qualitative, quantitative and mixed methods	10.2 (with 10.1 background)	Part 1 and 2
Research participants and study designs	10.3	Part 3
Context of interaction	10.4	Part 4
Selection of robot for study	10.5	Part 5
Mode of interaction	10.6	Part 6
HRI measures	10.7	Part 7
Research standards	10.8	Part 8

Overview of the chapter

HRI studies answer questions such as: What do people think of the robot's appearance? How do they react to its behaviour? Will they accept it? What are the short- and long-term effects? How does it perform technically? Each needs a different research approach.

- **Two essential components:** the human and the robot. Humans without robots is social science; robots without humans is robotics/AI. The **unit of analysis in HRI is always some form of interaction** between the two.
- **Context matters:** lab, school, hospital, different cultures or application domains all influence results and must be defined explicitly.
- **Robot-centred work** develops technical capabilities or tests functionality/design. **User-centred work** studies human behaviour or cognition that affects HRI success (e.g. extroverts may prefer direct robot communication, introverts indirect).
- **Balancing both:** *iterative design* (prototype, test, analyse, refine) and *behavioural models* (study human behaviour, build a model, test it with users).
- Methods are borrowed from sociology, anthropology and human factors, and some are HRI-specific (e.g. Wizard-of-Oz). An **experimental approach is now standard**, and there is a push to integrate qualitative and quantitative approaches.

Part 1: Defining the research question (10.1, background)

1.1 Exploratory vs confirmatory research

	Exploratory	Confirmatory
Aim	Find the "lay of the land" in a new, little-studied area	Test hypotheses
Prior knowledge	Not enough to formulate testable expectations	Enough to predict outcomes
Example	"How do people adopt and use a robot vacuum cleaner in their home over one month?"	"Do teenagers engage in more cleaning with a robotic vacuum than a conventional one?"
Classic study	Forlizzi and DiSalvo (2006): Roomba in real homes. People treated it as a social actor, it inspired teenagers to clean, and pets interacted with it	Asch conformity experiment run with robots: adults do not conform to robots, children do



Figure 10.1 A cat riding on a Roomba robot (2002–present). (Source: Eirik Newth)

- **Pre-registration:** register the hypothesis before the study (Center for Open Science, AsPredicted, ClinicalTrials.gov). It shows you did not adjust the hypothesis to fit the data or report only selected results.
- **CASA (Computers as Social Actors, Reeves and Nass, 1996):** people treat computers like human partners, automatically (e.g. they rate a computer higher when it asks about its own performance, which indicates politeness). Extended to HRI through gender attribution, mental models, social presence and agency studies.

1.2 Correlation vs causation

- **Correlational study:** shows variables change together but not *why*. Example: Roomba ownership vs teenage cleaning time. Teenagers with a Roomba might simply be tidier, or their parents ask more.
- **Experimental study (causal):** randomly divide a sample into groups, so no pre-existing differences. Treat groups identically except for the manipulated variable, then measure the outcome in both. Any major difference is due to the manipulation.
- **Spurious correlation:** a correlation with no causal link. The chapter's example is US cheese consumption vs deaths by tangled bedsheets ($\rho = 0.97$, $r^2 = 0.896$). The more tests you run, the more

likely at least one spurious correlation appears, so do not "test everything", even in exploratory work.

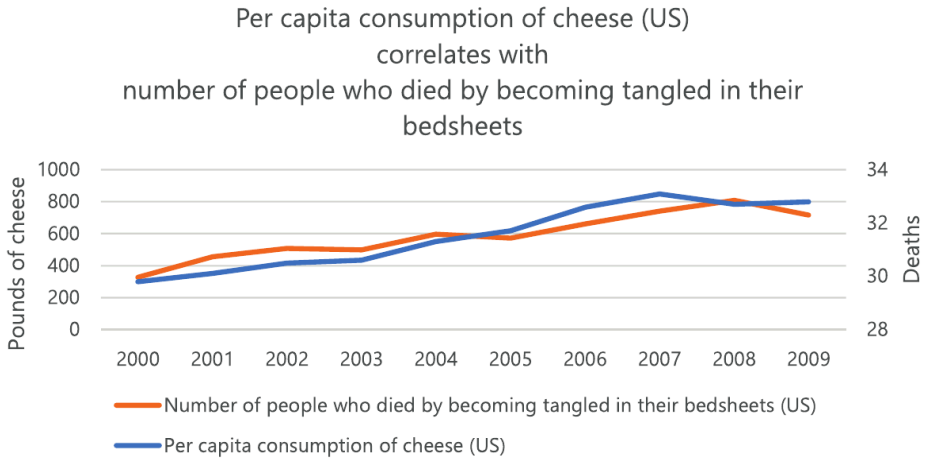


Figure 10.2 A strong correlation that has no causal relationship.

Part 2: Qualitative, quantitative and mixed methods (10.2)

Approach	What it gives	Limits / use
Qualitative	Qualities of interaction that are hard to put in numbers: underlying meaning and thematic patterns. Rich data	Data cannot be expressed numerically, so it cannot establish correlation or causation. Best for generating new hypotheses or theories
Quantitative	Surveys and controlled experiments; numeric data analysed statistically	Allows prediction and establishing cause and effect
Observational	Can produce both qualitative and quantitative data	Finds common patterns and correlations among people, robots and context. E.g. times adolescents clean vs conscientiousness; people call the Roomba "he/she"
Mixed methods	Exploratory work (interviews, focus groups, naturalistic observation) to find emergent factors, then experiments to confirm	E.g. interviews suggest autonomy makes the Roomba seem social, so test an autonomous vs game-controller-steered Roomba and compare ascribed sociality

2.1 User studies

- Experiments where people interact with a robot. **Not all HRI research needs one** (e.g. testing a receptionist robot's navigation).
- Critical feature: **random assignment** of a large enough sample to **experimental conditions** (variations of the robot, interaction or context).
- **Independent variable (IV):** what you control or manipulate (e.g. robot gender: male vs female, so two levels).

- **Quasi-experimental factor:** an IV you cannot assign randomly (e.g. participant gender). Adding it gives a **2 × 2 design** with four cells: male/female participants × male/female robot.
- **Dependent variable (DV):** what you measure (e.g. perceived warmth and assertiveness, since women are commonly seen as communal/warm and men as assertive). Eyssel and Hegel (2012) found robot gender cues produce stereotypical trait perception.
- **Manipulation check:** verifies the treatment worked (did participants perceive the "male" robot as male?). Done via a post-interaction question or by how they refer to the robot.
- **Operationalisation:** translating a theoretical construct into a measurement or manipulation. Validity needs both a well-designed DV and a validated IV.

2.2 Survey studies

- A list of questions answered via multiple choice or rating scales; easy to analyse but needs careful design.
- **Likert scale:** rate agreement with a statement (e.g. "I found the robot friendly", 1 to 5). Pronounced "lick-ert".
- **Semantic differential scale:** rate between opposing terms (scary–friendly, competent–incompetent).
- Use existing validated scales when possible: **Big Five** (personality), **Robot Social Attributes Scale** (robot sociality). Open-ended questions let respondents answer on their own terms.
- Surveys test correlations in a broad population, often with hundreds of participants. Some aim at a **representative sample** (matching gender, age, ethnicity proportions, or weighting data).

2.3 System studies

- Evaluate the **technical capabilities** of the robot rather than attitudes. May or may not involve users, but need the same rigour: verifiable hypotheses and performance claims, a study protocol and clear metrics.
- Example: does speech recognition work for children (higher pitch, more disfluencies)? Better to run **recorded child speech** through the software than to test live: it is **repeatable** across parameter settings and engines.
- Perceptual capabilities (face recognition, facial emotion classification, voice sentiment) are best assessed with consistent **test data sets** (e.g. IMDB-WIKI for faces) and established metrics so results are comparable.
- Classification metrics: **accuracy** (correct ÷ total), **precision** and **recall**.
- **Word error rate (WER)** = (substitutions + deletions + insertions) ÷ number of words in the actual sentence. Example: "Can you bring me a drink please" recognised as "Can bring me a pink sneeze" gives (2 + 1 + 0)/7 = 0.43.

2.4 Observational studies

- Made viable as robots became robust, reliable and cheap. Researchers observe naturalistic interaction **without intervening**.

- **Exploratory example:** Chang and Šabanović (2015) placed the Paro seal robot in a nursing home public space; they counted interactions and found social factors (gender, social mediation) affecting duration. Nothing was manipulated.
- **Field experiment / iterative example:** ATR (Japan) Robovie in a mall. Observe human behaviour, identify patterns, build behavioural models, deploy the robot, then evaluate whether the models had the expected effect.

Data source	Strengths	Weaknesses
In-person observation	Better grasp of broader context; can amend the protocol; notes aid interpretation	Limited by observer senses; no replay; interrater reliability needs multiple coders present, which can distract people
Video coding	Review as often as needed; revise coding scheme; easy to give a second coder for interrater reliability	Limited camera view (define focus beforehand); some places (nursing homes, hospitals, schools) forbid recording; labour intensive
Robot logs	Record robot state/actions and the human actions it perceived at the same time	Limited by the robot's ability to sense and categorise human actions

- **Coding scheme:** developed top-down (theory/practical interests) or bottom-up (from a sample of data). **Pilot test** it to remove missing, overlapping or unclear codes. Decide how fine-grained it needs to be.
- **Interrater reliability:** the extent to which different people agree on interpreting an observation (e.g. was a passer-by stepping aside "social behaviour"?).
- Outputs: frequency counts, qualities and patterns, and **temporal patterns** (e.g. a robot gaze cue followed by a person's joint-attention behaviour).

2.5 Ethnographic studies

- In-depth, usually **long-term** (months to years) observation that also seeks **what behaviours mean** to people and how they connect to organisational, social and cultural context (practices, values, goals, beliefs, discourse of all stakeholders).
- Behavioural observation is inspired by **ethology**; ethnography by **anthropology** and holistic understanding of sociocultural experience.
- Can be as an outside observer or through **participant observation** (researcher takes part). Often paired with **grounded theory**: data collection and interpretation continue throughout, with ongoing reflection and iteration.
- Still rare in HRI because of labour needed and the scarcity of robots capable of long-term interaction. Examples:
  - Mutlu and Forlizzi (2008): a year-long hospital service robot study. Patient type (oncology vs postnatal) determined whether nurses appreciated or hated it.
  - Forlizzi and DiSalvo (2006): families given a Roomba or a conventional vacuum for months. The robot (not the vacuum) was treated as a social agent and changed how the family cleaned, inspiring teenagers and men.

- Leite et al. (2012): empathic social robot in an elementary school. Task scenario and children's preferences shaped perceived empathy.
- Vertesi (2015): NASA Rover scientists. Team structure shaped robot use, and scientists acted out the robot's behaviours with their bodies, building team identity.

- Valuable because HRI is a young field still identifying which factors matter in robot design and implementation.

2.6 Conversational analysis (CA)

- Reports verbal and nonverbal aspects of interaction in **great detail**; applies to any interaction between people or people and technology.
- Process: record the interaction (video, several cameras), then make a very detailed **transcription** including pauses, turn-taking cues, laughter, behaviours; resolution can go down to video frame rate (blinks, gestures, posture). Participants may or may not know they are recorded.
- Attends to **turn-taking, back-channeling, overlap, repair statements, echo utterances, discourse markers**.
- Example: Fischer et al. (2013), contingent vs non-contingent feedback from iCub while participants taught it to stack shapes. Contingency changed verbosity, attention-getting tactics and word diversity, which matters for learning by demonstration.

2.7 Crowdsourcing participants

- Obtaining responses from many people online, paid or unpaid. Main tools: **Amazon Mechanical Turk (MTurk)** and **Prolific**. Typical job: view images/videos of robots and answer questions.

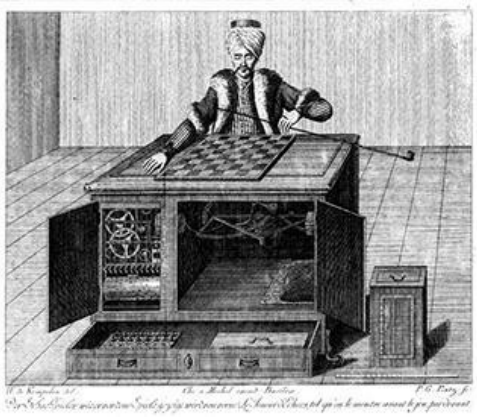


Figure 10.3 Amazon Mechanical Turk was named after a fake chess-playing machine called "The Turk" constructed in the late 18th century.

- **Pros:** large data, short time, modest cost (a few US dollars per participant), reach people otherwise hard to find. Good for **subjective evaluation** of behaviours with no formula (co-speech gestures, empathy in voice, gaze, back-channeling).
- **Cons:** little control over who participates and in what environment (distractions, wrong person on the account), and malicious or low-quality responses.

- **Quality control:** verification (attention-check) questions (e.g. enter a number shown briefly in the video, or "click the third option"); exclude responses completed far faster than the minimum time; withhold payment for poor quality (hurts the worker's rating); collect more data to offset variability.
- Use case by case. Unsuitable when the robot's physical presence matters, the effect is small, the population is scarce (elderly users, Swedish primary school teachers) or high language proficiency is needed. Age and technical affinity matter in computer-based studies, so use **plain language** and adapted questionnaires for children or people with mild cognitive impairment.

2.8 Case studies

- Qualitative design comparing an intervention's effect on a **single participant**: collect **baseline measures**, then compare behaviour during and after the intervention. Each participant is **their own control**.
- Used when participants are rare or individual differences are large and relevant. Common in medicine, education and HRI work on **autism**.
- Examples: Pop et al. (2013), three children and the Probo robot for identifying situation-based emotions. Tapus et al. (2012), four children with autism and Nao, with large variability. Averaging a group could **mask effects** because individual responses cancel out.

Part 3: Research participants and study designs (10.3)

3.1 Representativeness of the sample

- The "usual suspects" are **university students**: a **convenience sample** (accessible, available, close to labs).
- Ideal: a large, representative sample of end users for **external validity** (findings generalise beyond the study). Hard in experiments, easier in surveys.
- Students are treated as "close enough" for broad traits (stereotyping), cognition (memory) and attitudes (fear of robots), but balance gender or educational background where relevant. Computer science students would likely hold more positive attitudes and be more at ease with technology.
- Specific claims need specific samples. You cannot claim benefits for older adults with cognitive decline using students (or healthy older adults), nor study robots supporting young children's learning with students.
- Also plan access, recruitment, motivation, and whether to bring people to the lab, go to them, or run online.

3.2 Sample size

- Small samples cannot reliably detect effects because individual differences add **noise** (unless the effect is huge).

Between-subjects vs within-subjects designs

	Between-subjects	Within-subjects

Assignment	Participants randomly assigned to **one** condition	Each participant experiences **all** conditions
Example	One group sees the "male" robot, another the "female" robot; compare group means	Same people rate both robot versions
Participants needed	More	Fewer (the same person gives two ratings, cutting noise)
Limits	—	Not always possible (e.g. a robot assistant for broken-leg recovery). **Order effects** (e.g. always liking the first robot more) so **counterbalance**: half see female first, half male first

- Estimate sample size with tools such as **G*Power**, but resources (time, money, robots, participants) often constrain this.
- Special populations (e.g. older adults with depression) may need smaller samples. For autism studies, treat participants as **individual cases**.
- **Qualitative studies:** aim for **saturation**: stop when new data only repeat existing themes. Harder to operationalise and measure, so various definitions exist.

Part 4: Context of interaction (10.4)

4.1 Location of study: lab vs field

	Lab	Field ("in the wild")
Control	Strict control over introduction, task, environment and length; strict manipulation of variables	Flexible; participants decide how, when, whether and why to interact, and may ignore the robot
Strength	Isolates variables	Closer to day-to-day HRI; reveals emergent phenomena, new variables, and complex contextual effects (institutional culture, interactions among people)
Status	Majority of early HRI research	Still relatively rare, though platforms are now robust enough

4.2 Temporal context: short-term vs long-term

- Most lab studies capture **"the first 10 minutes of HRI"**, affected by the **novelty effect** (unfamiliarity with robots), so validity for long-term interaction is limited.
- They still show which person and robot characteristics affect the first encounter, and help set up a **positive feedback loop** for long-term interaction.
- Long-term studies (days, weeks, months, years) show how interaction develops, how robots are integrated into social contexts, and how human–human interaction changes because of the robot.

4.3 Social units of interaction

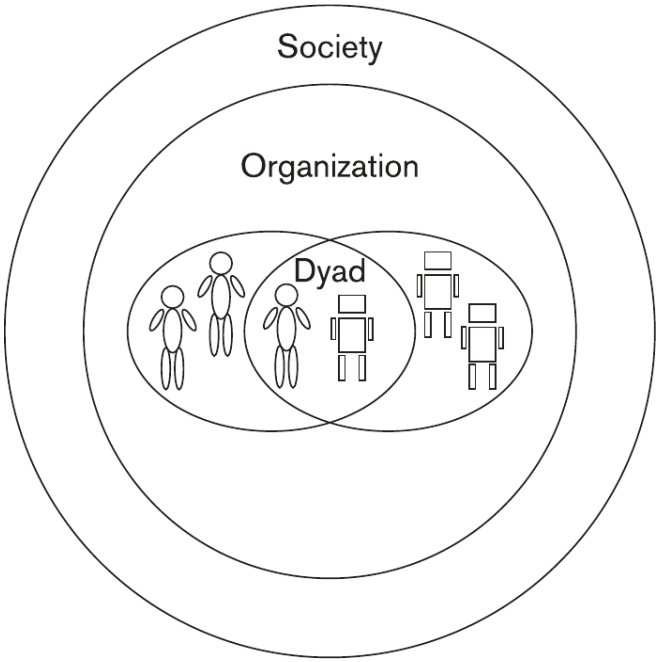


Figure 10.4 Units of analysis in HRI.

- **Dyad** (one person and one robot) is the most common unit, largely due to early constraints: robots were hard to procure, maintain and operate, so the typical study was a lab experiment with one participant and one robot.



Figure 10.5 Robovie in school.

Example. As early as 2006 Robovie supported group interaction at an elementary school: it taught English and tracked children's social networks, keeping them interested by unlocking secrets (Kanda et al., 2007b).

- **Groups:** people often interact with robots in groups ("in the wild"; Šabanović et al., 2006). Groups bring group effects, multi-party collaboration and team dynamics. Leite et al. (2015): children recalled a robot-told story better when interacting individually than in a group of three. Brščić et al. (2015): children abused a mall robot only when in groups.

- **Multiple robots** (teams, swarms, co-present actors): robots in teams are perceived as having more social agency. Carpenter (2016): soldiers saw bomb-disposal robots as team members and felt sadness when one was destroyed.
- **Organisations** (education, nursing, military): shows effects of the broader context such as labour distribution, roles, acceptance, adaptation to practice and institutional values. Mutlu and Forlizzi (2008): a robot reduced work for some and increased it for others. Chang and Šabanović (2015): one **advocate** for Paro created a positive feedback loop for long-term adoption. Vertesi (2015): the NASA team's egalitarian set-up suited sharing a scarce robot.
- Figure 10.4 nests these units: **Dyad** within **Organization** within **Society**.

Part 5: Selection of robot for the study (10.5)

- Decide the robot's **appearance, functionality and ease of use**. Some decisions are practical (availability, cost), others follow research interests.
- Robots as **research tools / stimuli**: manipulate appearance, behaviour, communication mode and style, and role in the interaction to see the effect on outcome variables. Central to experiments, also useful in exploratory work.
- Researchers use **off-the-shelf robots** or design their own prototypes. Consider required hardware and software capabilities and the **level of autonomy**.
- Commercial robots suited to HRI studies: **Nao** (Aldebaran), **Furhat** (Furhat Robotics), **QTrobot** (LuxAI), **Paro** (Intelligent System).
- Programming: Nao and QTrobot support **visual programming** for quick set-up. The **Robot Operating System (ROS)** needs more advanced skills but greatly extends behaviours and includes packages for sensory perception and visualisation.

Part 6: Mode of interaction (10.6)

People can meet a real robot or see pictures or videos; the robot can be autonomous or tele-operated; people can come to the lab or researchers can take robots to people; a study may need one data point or thousands.

6.1 Wizard of Oz (WoZ)

- Participants are led to believe the robot acts autonomously while a **member of the research team operates it**. Used when developing autonomy is not the research focus. Participants must be **debriefed** afterwards.
- Lets researchers "pretend" the robot has skills that need more development or programming time. Especially suitable when technology is almost usable (e.g. **speech recognition**): a wizard recognising utterances makes the interaction flow robust and believable.
- Problematic to fake a complete AI that sustains a serious prolonged conversation (an unrealistic level of capability).
- Can test perceptions of **advanced capabilities** (nuanced social context understanding; Kahn et al., 2012) and **collect data** to develop designs or autonomy.

- In experiments, **constrain the wizard's behaviour** so the robot is consistent across conditions and adds no confounding variation.

Name origin. From the film *The Wizard of Oz*: Toto pulls back the curtain and reveals Professor Marvel operating the machinery behind the giant wizard. In HRI the wizard hides in the background and gives the robot the semblance of more autonomy than it has.

6.2 Real vs simulated interaction

- Ideal: real-time, face-to-face interaction, yet many studies show only **photos or videos**.
- Debate: Dautenhahn et al. (2006) found the two broadly equivalent; Bainbridge et al. (2011) found a **more positive experience with physically present robots**; Powers et al. (2007) found large attitude differences between co-located and remote robots.
- Visual stimuli **limit generalisability**, but suit exploratory studies (e.g. perception of robot forms; DiSalvo et al., 2002), hard-to-access populations (cross-cultural samples), avoid the uncontrolled nature of live interaction, and fit **online participant pools** (MTurk, universities, word of mouth).

Part 7: HRI measures (10.7)

7.1 Direct measures

- Ask participants directly to rate the DV (e.g. rate the robot's warmth and authoritativeness).
- **Self-reports** (questionnaires) have high **face validity**: people know what is being measured. This makes it easy to answer to please the researcher, present oneself well or "be a good participant" (social desirability).
- **Interviews** give a more holistic picture. **Structured:** predetermined questions in set order. **Semistructured:** some planned, others arise spontaneously. Open responses need labour-intensive coding after transcription. Useful alongside questionnaires (de Graaf et al., 2017: long-term survey plus interviews on why people do not use a communication robot at home).

7.2 Indirect measures

- Needed when people are reluctant to report true attitudes or are unaware of unconscious beliefs.
- **Reaction times:** proxy for attention or engagement. **Eye tracking:** attentional focus and cognitive processing. **Physiological measures** (heart rate, skin conductance): stress level. **Computerised measures:** e.g. a variant of the **Implicit Association Test** for anthropomorphisation.
- Drawbacks: harder to administer, need equipment, and results are **ambiguous**. Skin conductance shows excitement but not whether it is fear or enjoyment; a Nao study (Kuchenbrandt et al., 2014) found readings inconclusive.

	Direct	Indirect
Examples	Questionnaires, rating scales, interviews	Reaction time, eye tracking, heart rate, skin conductance, Implicit Association Test

Pros	Easy; participants know what is measured; easy to analyse	Not affected by wish to please; can capture unconscious attitudes
Cons	Social desirability; limited to what people can and will report	Equipment, harder to administer, ambiguous interpretation

Best practice. Combine direct and indirect measures, or several indirect measures, and establish that every measure is **reliable and valid**: pilot test, formally validate new measures, or use widely accepted validated measures from the literature.

Part 8: Research standards: statistical analysis (10.8)

Fisher's warning. "To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of." Seek statistical advice **early**.

8.1 The three building blocks of statistics

Most tests rest on **tendency, variability and number of observations**, illustrated with restaurant ratings.

- **Tendency (difference in means):** 4.8 vs 3.2 stars, you are fairly sure A is better; the closer the averages, the less sure.
- **Sample size:** 5 stars from 3 reviewers vs 4.7 stars from 1,000: more responses give a better estimate of true quality. Coin example: 75% heads in 4 flips proves little; in 1,000 flips it is convincing.
- **Variability:** same 4.2 average, but reviews spread from 1 to 5 stars vs mostly 4 stars: the more variable the data, the less certain the mean reflects the true effect.

Descriptive statistics (first stage of analysis)

- Report **means** (tendency), **standard deviations** (variability for normal data; otherwise give the range) and **number of participants**.
- Report **demographics** (age, gender) and **excluded data points with reasons** for integrity and transparency.

Inferential statistics and the p-value

- Combine tendency, sample size and variability into a **test statistic**, then a **p-value**: the probability of getting the data if there were **no true effect**.
- **Null hypothesis significance testing (NHST):** if $p \leq$ threshold (typically **0.05**), the result is "significant" and the null hypothesis is rejected in favour of the alternative. Different designs need different tests (see Andy Field, 2018).

8.2 Problems with NHST

Issue	Explanation

Type I error (false positive)	At p ≤ .05, 5% of the time when the null is true the data will look like an effect. False positives look identical to true positives, so significance is not conclusive proof
Misreading the p-value	p is **not** "the chance of a Type I error". It only gives that chance **conditional on the null**; the overall chance results are a Type I error cannot be computed
p(A|B) ≠ p(B|A)	NHST gives P(data | no effect), but researchers want P(effect | data). Shark example: P(dead | shark bite) ≈ 1 but P(shark bite | dead) ≈ 0 (Cohen, 1994, "The Earth Is Round (p < .05)")
p-values are unstable	Replicating a significant study can give p anywhere in [0.00008, 0.44] for 80% of replications (Cumming, 2008). Unreliable as a measure of solidity
Significance ≠ effect size ≠ importance	**Effect size** (from tendency and variability) says how large a change is; **p** (also uses sample size) says how consistent the effect is in the sample. Importance depends on context: a rise in recovery from 4% to 6% is trivial for foot fungus but important for an aggressive cancer
p-hacking	At p ≤ .05, a false positive is expected 1 in 20 times under the null. Run enough tests and report only significant ones, and a "finding" appears. Especially risky in observational studies with many variables (same logic as the spurious correlation in 10.1.2)

- **Common effect sizes: Cohen's d** (t-test), **η^2p or ω^2** (ANOVA), **R^2adj** (regression). Guidelines classify small/medium/large effects.

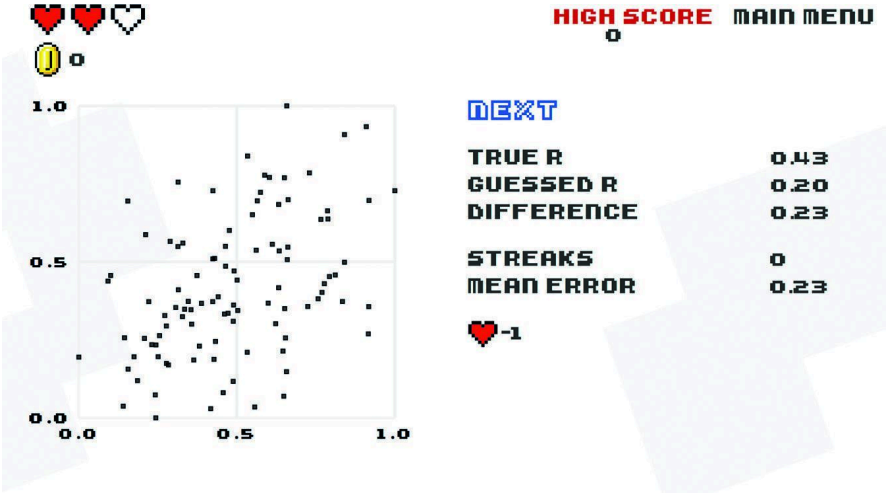


Figure 10.6 How strongly would you say the two variables are correlated? People find it very hard to infer correlation strength from plots (try www.guessthecorrelation.com). The correlation here is r = .43, considered a medium effect. (Source: Omar Wagih)

8.3 Good practices to overcome the issues

- **Report confidence intervals (CIs)** and **effect sizes** along with p-values. A CI does not test significance; it shows how confident we are the population mean lies between a minimum and maximum. For a

95% CI, a replication mean has an 83% chance of falling within the original CI. Together they convey the **magnitude** and **precision** of an effect.

- **Type II error (false negative):** concluding there is no effect when there is one. Avoid Type I and II errors with sufficient **statistical power**. Power depends on tendency, variability and sample size, but only **sample size** is under your control (more participants or more data points per participant). Needed numbers rise with design complexity and smaller effects. **G*Power** calculates power before and after a study.
- **Replication:** psychology faced a **replication crisis** (p-hacking partly responsible). HRI is now taking reproducibility more seriously (Irfan et al., 2018), helped by widely available platforms (Nao, Baxter) instead of bespoke robots.
- **Pre-registration** forces researchers to specify planned tests before data collection, aiding replication and preventing p-hacking.
- **Share code and procedures** so others can rerun the experiment and test **generalisability** across contexts. Methodology choice affects generalisation from lab to field; combining methods over time by the research community may be needed.
- **Bayesian inference:** a radical alternative that abandons NHST. NHST gives a dichotomous outcome (significant or not); Bayesian statistics starts from **prior information** and updates it with new data, yielding a **range of possible values** with confidence in each. Hard conclusions are rare; prior beliefs are strengthened or weakened.

Related standards the chapter adds

Ethics in HRI studies (10.9)

- **Informed consent** is required for any human-participant research (correlational or experimental, qualitative or quantitative, online or in person): nature of study, voluntary participation, risks and benefits, what is done with the data.
- **Deception / cover story** (e.g. WoZ, bogus feedback about personality or performance) requires a **debriefing** so participants do not leave believing robots are fully autonomous, protecting their psychological well-being.
- Qualitative studies: later share findings with participants, possibly co-develop interpretations and design guidelines.
- Consider physical and psychological safety, and **attachment**: an older adult attached to a home robot may be distressed when it is removed.
- Consult codes of conduct (APA, American Anthropological Association, ACM) and the **university ethics committee**; ethics approval is required by many journals, so obtain it **before** data collection. Also reflect on funding sources and societal consequences of robots.

Conclusion (10.10)

- HRI should adopt the same rigour as the social sciences. It shares their problems: experiments are almost never repeated and there is **publication bias** toward positive results.
- Opportunities: stored and shared data and video logs, open-access publishing, pre-registration. **Green open access** (self-archiving) increases citations. Beware flaky journals and conferences. HCI

offers methodological guidance, but robots are embodied and raise different technical and social challenges.

Quick revision: key terms

Term	Meaning
Independent variable	What the experimenter manipulates/controls
Dependent variable	What is measured
Quasi-experimental factor	IV that cannot be randomly assigned (e.g. participant gender)
Manipulation check	Confirms the manipulation was perceived as intended
Operationalisation	Translating a construct into a measure or manipulation
Between- / within-subjects	Different participants per condition / all participants in all conditions (counterbalance order)
Convenience sample	Participants chosen for accessibility (e.g. students)
External validity	Findings generalise beyond the study
Saturation	Qualitative stopping point when new data bring no new themes
Novelty effect	Initial reactions differ because robots are unfamiliar
Wizard of Oz	Human secretly operates the "autonomous" robot
Face validity	Participants can tell what a questionnaire measures
Interrater reliability	Agreement between coders
WER	(S + D + I) ÷ words in the actual sentence
p-value	Probability of the data if there were no true effect
Type I / II error	False positive / false negative
Effect size	How large a difference is (Cohen's d, η^2p, ω^2, R^2adj)
Statistical power	Ability to detect a true effect; controlled by sample size
Confidence interval	Range likely to contain the population mean