10

Research Methods

What is covered in this chapter:

- Methodological considerations and various decisions you need to make in setting up and performing a human–robot interaction (HRI) study.
- The strengths and weaknesses of different research methods and how to identify them for understanding and evaluating HRI.
- How the choice of robot, environment, and context matter for study results.
- The importance of looking at new ways of reporting data and insights befitting HRI, even though there is a tradition of reporting experimental work.

Now that you have a robot, you want to know with some certainty how it performs. What do people think about its appearance? How do they react to its behavior? Will people accept it? What will the effects of using the robot be in the short term or over a longer period of time? How does the robot perform technically? These are common questions in human–robot interaction (HRI), and they will require you to use different research approaches and methodologies to find the answers.

HRI research consists of at least two interrelated components: the human and the robot. These are essential to any HRI study; if you investigate humans without robots, you are engaging in social science research, whereas research on robots without humans involved would qualify as robotics or artificial intelligence (AI) research. The unit of analysis in HRI is always some form of interaction between the two. The context in which HRI happens is of high relevance and needs to be explicitly defined in studies. You might study HRI in the lab or in a school or hospital; you might study HRI in different cultures or in different application domains. The context in which the robot interacts with people is very likely to have a strong influence on your results, and you need to be aware of with whom and in what circumstances the interaction unfolds.

Although the focus of HRI is always on the interaction between humans and robots, there are different aspects of this relationship to study. In *robot-centered work*, the research focus might be on developing the technical capabilities that robots need to interact with people or testing different aspects of the robot's functionality or design to see which are most effective. In *user-centered work*, on the other hand, the focus of a study could be on understanding aspects of human behavior or cognition that will affect the success of HRI. For instance, an extroverted user might prefer more direct communication by the robot, whereas an introverted user might like indirect communication.

HRI research increasingly strives to strike a balance between these two approaches, coupling robot- and user-centered aspects in different ways. For example, in iterative design, the robot's design goes through a number of cycles of prototyping, testing, analyzing, and refining. Researchers come up with a series of robot design ideas, which they then test out with users. Based on the users' preferences, the researchers then further develop the robot's appearance and capabilities. Another mode of coupling user- and robot-centered aspects of HRI is through studying human behavior to develop behavioral models that can then be applied to HRI and testing those out with users to see if they produce the expected and desired results in interaction.

Studies in which users interact with the robot, tests of the robot's performance, and more open-ended explorations of ways in which people and robots interact in everyday life are all part of HRI research. Consequently, HRI researchers draw on and often mix a variety of research methods and techniques, some adapted from other disciplines (e.g., sociology, anthropology, or human factors research) and some developed for the HRI field itself (e.g., the "Wizard-of-Oz" technique, described in Section 10.6.1). To employ these methods successfully, HRI researchers need to be aware of their strengths and weaknesses, the kinds of data and insights they may produce, and the types of technical and human resources they require.

Taking an experimental approach has become standard in the HRI community (Hoffman and Zhao, 2020). This was not always the case, and a quick glance at older HRI research will show methods that would make current HRI researchers blush. There is a push to have current research meet criteria for methodological soundness that are applied in other empirical sciences (e.g., psychology), integrating qualitative and quantitative approaches (Baxter et al., 2016; Hoffman and Zhao, 2020; Fischer, 2021; Seibt et al., 2021).

This chapter discusses the kinds of decisions that HRI researchers make at different points in the research process, from defining the research questions (Section 10.1) to study design (Section 10.2) and statistics (Section 10.8), and explains the journey you will make when evaluating the interaction between robots and people. After walking through the steps to formulating a research question in Section 10.1, Section 10.2 provides examples of different uses of qualitative, quantitative, and mixed methods in user and system studies, observational and experimental studies, and other forms of HRI research. The selection of participants is the focus of Section 10.3, whereas Section 10.4 emphasizes the importance of defining the context of interaction as part of the initial study design. Sections 10.5 and 10.6 consider how to choose an appropriate robot and mode of interaction for your HRI studies. Sections 10.7 and 10.8 present various metrics and research standards to be taken into account in HRI research, including statistical and generalizability concerns.

This material has been published by Cambridge University Press as Human-Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781009424233 (https://www.cambridge.org/9781009424233).

Finally, 10.9 covers ethical considerations to keep in mind when designing a study. The overall aim of the chapter is to provide a basis from which to make initial study design choices and then delve more deeply into research methods to develop your own novel HRI studies.

10.1 Defining a research question and approach

Defining a good research question is one of the hardest tasks of a researcher. To form a strong research question, a researcher must consider previous relevant work and replicate or extend it to contribute new scientific insights. In HRI, such insights can come in the form of knowledge about human cognition and behavior, guidelines for robot design, technical aspects of the robot, or findings that can inform the application of robots in different use contexts.

Research questions in HRI might arise from theoretical considerations, such as the expectation that people will treat robots as social, or from the pragmatic need to test the usability of a certain robot feature or function. We recommend searching publications across disciplinary databases to incorporate research findings from multiple fields of relevant expertise. Ideally, you would look for a well-established phenomenon or theory and seek to replicate and extend it in your new research project, independently of whether it is about humans or robots. Research on interactions among humans can easily serve as a blueprint for human–robot research. Existing work in HRI, psychology, sociology, anthropology, design, and media communications can provide relevant insights into the underpinnings of smooth, successful, and acceptable HRI or into the optimal human-centered design of a novel robot platform.

To illustrate, in the 1990s, Reeves and Nass (1996) proposed the "computers as social actors" (CASA) approach and sought to replicate classic psychological findings in the context of human–computer interaction (HCI). In their seminal work, the authors conducted studies that provide evidence for the hypothesis that computers are treated just like human interaction partners. Moreover, they found that such behavior occurs quite automatically. For instance, they showed that humans give higher ratings if a computer asks about its own performance than when they have to rate the performance on a different computer, which indicates that people are polite to computers. Later on, the CASA approach was successfully extended to HRI through a wide array of studies, including some exploring the attribution of gender to robots (Eyssel and Hegel, 2012) and users' mental models of robots (Walden et al., 2015) and others studying the effects of perceptions of social presence and agency in caregiving (Kim et al., 2013) and educational scenarios (Edwards et al., 2016). This paradigm continues to inspire new research in HRI.

10.1.1 Is your research exploratory or confirmatory?

Broadly speaking, research can be classified as either exploratory or confirmatory. Exploratory research questions deal with phenomena that have not previously been examined in detail and aim at finding out the general "lay of the land" in a specific domain. For example, you might ask, "How do people adopt and use a robot vacuum cleaner in their home over one month?" or "Do large language models contain sufficient world knowledge to power a conversation with a robot?" Exploratory research assumes that there is not enough relevant prior information about the phenomenon to formulate testable expectations about the potential outcomes of the study, and it therefore seeks to explore what factors might be important and which outcomes are possible.

In an exploratory HRI study, Forlizzi and DiSalvo (2006) investigated how a vacuum-cleaning robot is integrated into the homes of real people. Their findings produced many surprises for the research community, including that people would treat autonomous robotic vacuums as social actors, that such vacuums could inspire teenagers to clean their rooms, and even that some pet–robot interaction occurred (see Figure 10.1).

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

When there is enough information to formulate hypotheses about the possible outcomes of an intervention, we enter the domain of confirmatory research. The goal of confirmatory research is to test hypotheses. In your hypothesis, you need to spell out the findings that you anticipate prior to starting your study and explain why you think those findings should be expected. A key point here is to formulate a question in such a way that it is verifiable. Take this example from everyday life: You might know that teenagers are often interested in new gadgets and technologies but tend to avoid doing chores. This may lead you to expect that introducing a robotic vacuum cleaner into their homes will increase their engagement with cleaning compared to introducing a normal top-of-the-line vacuum cleaner. You would then design your study in such a way that it answers the following research question: "Do teenagers engage in more cleaning with a robotic vacuum cleaner compared to a conventional vacuum cleaner?"

You might consider registering your hypothesis prior to conducting your experimental study at one of the many sites available for that purpose, such as the Center for Open Science (https://osf.io/prereg), AsPredicted (https://aspredicted.org), or the U.S. National Library of Medicine (https://clinicaltrials.gov). This will keep your work in line with the standards and rigor in empirical sciences and makes it clear that you have not adjusted your hypothesis to fit the data or have reported only carefully selected results (Nosek et al., 2017).

The teenagers and cleaning example shows how hypotheses can be inspired by commonsense knowledge, but you can also build on prior empirical research and social theory to develop hypotheses about HRI. One such example is the social conformity theory of Solomon Asch, who showed how people tend to conform to peer pressure. In an elegant experiment, he showed that when people complete a simple visual task in a group setting, they are more likely to give the same response as others in the group even if they know the response is wrong (Asch, 1951). This classic experiment can be run with a group made up of robots rather than people. Will people conform to robots? Studies have shown that adults do not, but children do (Brandstetter et al., 2014; Vollmer et al., 2018).

10.1.2 Are you establishing correlation or causation?

Along with deciding whether your research questions call for an exploratory or confirmatory approach, you need to decide whether you want to establish correlation or causation between the variables of interest in your research study.

In correlational studies, we can show a clear pattern by which the variables change value in relation to each other, but we cannot know what causes this relationship. A correlational survey study of teenagers using the Roomba could measure whether there is a statistical relationship between households owning a Roomba and the amount of time teenagers spend cleaning. We would, however, not necessarily know why this relationship happens. It might be that teenagers who own a Roomba are more tidy to start with or that their parents ask them to clean more often. To make the claim that a Roomba would increase the time spent cleaning, you would need to compare the behaviors of two similar groups of teenagers by giving one group a Roomba and the other group a regular vacuum cleaner, then measuring the outcomes.

This requires an experimental study design to investigate the causal relationship and show that a change in one variable actually leads to a change in the other. We do this by dividing a sample into two (or more) groups at random. This randomization should ensure that there are no preexisting differences between the groups. Then, the manipulation is introduced: the groups are treated exactly the same *except* for the variable that we think has an effect. In the Roomba example, this could mean that one group gets a Roomba, and the other gets a new regular vacuum cleaner. Finally, the variable of interest is measured in both groups. Due to randomization and the otherwise similar treatment, any major difference that is observed would be the result of our manipulation.

The difference between correlation and causation is important because it defines what conclusions can be drawn from the findings. Correlation says nothing other than “these things happen to occur simultaneously”—for example, there will be a strong correlation between the number of firefighters on the scene and the damage recorded after the fire. This does not, of course, mean that the damage was caused by firefighters and that we should stop sending firefighters whenever there’s a fire. Sometimes a correlation even pops up for no reason at all, a so-called *spurious correlation*. An example of a spurious correlation is the strong ($\rho = 0.97, r^2 = 0.896$) relationship between U.S. per capita cheese consumption and the number of people who died by becoming tangled in their bedsheets (see Figure 10.2[1]). As explained

[1] Data from: www.tylervigen.com/

This material has been published by Cambridge University Press as Human-Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781009424233 (https://www.cambridge.org/9781009424233).

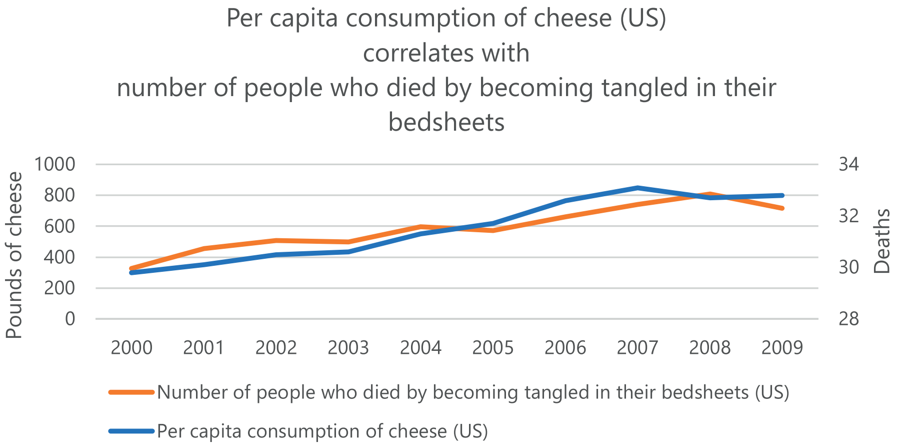


Figure 10.2 A strong correlation that has no causal relationship.

in Section 10.8.1, finding at least one spurious correlation becomes more likely as you run more statistical tests, and unless the relationship is obviously ridiculous (like with the cheese and the bedsheets), there is no way of telling apart correlations indicating a "real" relationship from the spurious ones. Thus, even if you are doing exploratory and correlational research, you should not just test for anything and everything.

10.2 Choosing among qualitative, quantitative, and mixed methods

How you define your research question will also affect what type of methods you should use to answer it. Qualitative methods allow researchers to understand the qualities of an interaction that are difficult to capture in numbers. This requires researchers to identify and interpret the underlying meaning or thematic patterns that they see in social interaction. The data that are derived from these studies typically cannot be expressed numerically, which disqualifies this approach from being used to establish correlations or causation. That is not to imply that qualitative research has little use to science; it tends to result in rich data, which can be used to generate new hypotheses or theories to test.

Quantitative methods, in contrast, often take the shape of surveys or controlled experiments and produce data that can be expressed numerically and analyzed statistically to check for correlations and causation. They will therefore allow you to make predictions or even establish cause and effect. Observational studies (see Section 10.2.4) can produce both qualitative and quantitative data, which can be used to investigate commonly seen patterns in interaction and correlations between the characteristics of people, robots, or context. For instance, you might find from observation and interviews that the number of times adolescents clean with the Roomba can be related to their personality characteristics, such as self-reported conscientiousness. The interviews might also tell you that people talk about the Roomba as a social actor, calling it a "he" or "she" rather than an "it" (i.e., a tool).

Finally, your research questions might call for a mixed-methods approach, which may include exploratory research using interviews, focus groups, or observation of naturalistic interaction to identify emergent factors significant to HRI, followed by experiments to confirm these relationships. For example, if your interviews lead you to think that the autonomous behavior of the Roomba is what makes it seem social to people, you could set up an experiment to test this. Such an experiment would have two groups of participants, whom you present with either an autonomous Roomba or a Roomba that they steer using a game controller. You can then measure the level of sociality they ascribe to each Roomba and test if these are significantly different from one another.

10.2.1 User studies

User studies are experiments in which you bring people in to interact with a robot. Not all HRI research requires a user study—for example, you might just want to test the navigation capability of your receptionist robot. However, most HRI research at some point will involve a study in which you measure how users respond to variations of the robot, the interaction itself, or the context of the interaction. These different variations are called *experimental conditions*. The critical feature of a user study is the random assignment of a large enough sample of research participants to your experimental conditions. Experimental conditions typically emerge from the factors that you consider of importance or interest and should be outlined in your research design. For instance, assume we want to test whether people apply human stereotypes to a gendered robot. To test this, we run an experiment using a male and a female robot prototype. The robot's gender is called the *independent variable*, which is the aspect of the experiment that is controlled or (in experiments) manipulated. Because we test two robot versions, male versus female, the independent variable has two levels. The resulting research design thus leaves us with two conditions to which we randomly assign our research participants.

If we think that gender stereotyping of a robot also depends on the gender of the human watching it, we want to test not just for the effect of robot prototype gender but also take into account participant gender as well. We thus add a second independent variable to our design: participant gender. Because we cannot manipulate this variable (we cannot randomly assign a gender to each participant who walks into our lab), participant gender would be called a *quasi-experimental factor*. Our study design now has a 2 × 2 format: robot gender (male vs. female) and participant gender (male vs. female). In our analysis, we will thus be comparing four groups, or "cells" in our design: males rating a male robot, males rating a female robot, females rating a male robot, and females rating a male robot.

Now the question is: How exactly do we measure what we want to know? The variables we measure are called *dependent variables*. We know from the psychological literature that women are commonly perceived as communal and warm, whereas men are perceived as more assertive (Bem, 1974; Cuddy

This material has been published by Cambridge University Press as Human-Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781009424233 (https://www.cambridge.org/9781009424233).

et al., 2008). We can use this information to measure to what extent our male and female robot prototypes are being stereotyped. Indeed, previous research studies have shown that manipulating robot gender leads to a stereotypical perception of traits in robots (Eyssel and Hegel, 2012). People seem to reproduce the stereotypes that are common among humans in the context of robots.

Not only does the dependent variable need to be well designed, but it is also important that the independent variable (i.e., the construct of interest) is validated. Can we be sure that our study participants actually recognized the robots as male or female? To establish the validity of our results, we need to know whether robot gender was operationalized successfully. We can do this by including a manipulation check in our study to see that our experimental treatment was indeed effective, that is, that our participants indeed perceived the robot with male gender cues as male and the robot with female cues as female. This could be done simply by adding a post-interaction question asking them to identify the gender of the robot and/or by seeing whether they refer to the robot by a specific gender when they talk about it after the interaction. Only once this is established can researchers be sure that the operationalization—that is, the translation of the theoretical construct of interest into a measurement or manipulation—was effective.

10.2.2 Survey studies

Sometimes HRI researchers choose to use a *survey*, which is a list of questions to be answered by participants. Answers are often given through multiple-choice options or some sort of rating scale. One commonly used type of scale is the Likert scale (pronounced as “lick-ert,” not “like-ert”). A Likert item asks respondents to rate statements about their attitudes and opinions on a topic based on how much they agree—for example: “Rate the statement ‘I found the robot friendly’ on a scale of 1 (‘Strongly agree’) to 5 (‘Strongly disagree’).” Another form of scale that is often used is the semantic differential scale, which asks respondents to evaluate the qualities of an artifact, or their attitudes, on a spectrum between two opposing terms (e.g., scary–friendly, competent–incompetent).

Multiple-choice or scale-based questions make the survey easier to analyze later on but require careful design while developing the survey to make sure that the questions are appropriately measuring the concepts the researchers are interested in. Along with making up their own questions and scales, researchers can use questions and scales developed and evaluated by other researchers to measure concepts of interest (e.g., evaluating participant personality with the Big Five Scale (John and Srivastava, 1999) or evaluating robot sociality with the Robot Social Attributes Scale (Carpinella et al., 2017)). Finally, researchers sometimes include open-ended questions in surveys as well, particularly when it is important to allow respondents to provide answers based on their own terms and categories or to understand their thought process or understanding of concepts while answering the survey (e.g., “Describe your ideal robot before you answer the following questions about it"). Because survey research is well established in the social sciences, there are many handbooks that describe how to go about constructing and performing surveys (for some examples, see Fowler (1995, 2013)).

Surveys allow researchers to investigate correlations between various factors relevant to HRI in a broad population. Such surveys often involve hundreds of participants and accommodate analyses with many different factors. Some surveys try to have a representative sample of participants, which can involve making sure the number of participants in certain categories (e.g., gender, age, ethnicity) corresponds to their percentage in the general population or weighting the collected data to achieve representative ratios.

10.2.3 System studies

Whereas user studies are used to report on people's attitudes toward and interaction with robots, system studies are those that evaluate the technical capabilities of the robot. A system study might involve users, but user involvement is not always needed. At the same time, system studies do require the same rigor expected from user studies. This means that verifiable research hypotheses and performance claims, a study protocol, and clear metrics are all key to system studies.

For example, when designing an interactive robot for children, you might want to know how well automated speech recognition works for your target user group (Kennedy et al., 2017). Speech recognition has been designed to work well for adults, but it might not be suitable for children due to their voices having a higher pitch and their speech often containing more disfluencies and ungrammatical utterances. To test whether speech recognition works for child speech, you could ask children to interact with your robot, but a better idea would be to use recordings of children's speech and pull these through the speech-recognition software. The benefit of this approach is that the experiment is repeatable: you can try different parameter settings in the software or even swap different speech-recognition engines and assess the performance using the same recordings.

Systems studies are often used to assess the perceptual capabilities of the robot. Capabilities such as face recognition, facial emotion classification, or sentiment detection from voice are best assessed using consistent test data sets with well-established metrics. For some capabilities, there are existing data sets that can be used to assess the performance of the robot. For face recognition, several data sets exist, for example, the IMDB-WIKI, which contains images of people extracted from the IMDb database and Wikipedia; in addition to labels, the images contain gender and age information (Rothe et al., 2016). The use of well-established metrics allows you to compare the performance of your robot to that of others. Classification problems often have agreed-on methods of reporting performance, such as reporting the accuracy of the classification (the number of correct classifications divided by the total number classifications, including the ones that are wrong) or the precision and recall. Speech-recognition performance is often expressed as a word error rate (WER), which is the total number of substitutions, deletions, and insertions in the text divided by the number of words in the actual spoken sentence. So if "Can you bring me a drink please" is recognized as "Can bring me a pink sneeze," that is a WER of $(2 + 1 + 0)/7 = 0.43$. It is worth exploring what the accepted metrics are in a particular discipline and rigorously sticking to the accepted method for evaluating and reporting system performance.

10.2.4 Observational studies

As robots have become more robust, more reliable, easier to use, and cheaper, it has become viable for HRI researchers to study how people and robots interact in various naturalistic contexts using observational methods. Observing how people interact with robots, for example, by studying where they place robots in their environments and how they respond to different kinds of verbal and nonverbal cues performed by robots, allows researchers to understand how HRI can unfold in a more natural way, without researchers directly intervening in the interaction.

Observational studies can be exploratory, involving putting a robot into a specific environment to see how interactions there unfold. An example of such an observational study is the work of Chang and Šabanović (2015), who put a seal companion robot in a public space in a nursing home and observed when and how different people interacted with the robot. The findings included frequency counts of interactions with the robot, as well as the identification of different social factors (e.g., participant gender, social mediation effects) that affected whether and for how long people interacted with the robot. The researchers did not manipulate anything about the robot or the environment. They just observed.

Observational studies can also be performed to evaluate, by means of a field experiment, how effective a robot is for a particular task or the effect of certain design variables on interactions. Researchers from the Advanced Telecommunications Research Institute (ATR) in Japan have performed several observational studies of interactions between the humanoid Robovie and mall customers. These studies represent a particularly fruitful iterative form of design and evaluation using observational techniques. In the initial stages of the study, researchers observed general human behaviors and analyzed these observations to identify particular behavioral patterns, which they then used to develop behavioral models for the robot. The robot was then placed in the mall, and people's reactions to it were evaluated to see if the behavioral models had the expected positive effects on people's responses.

Observational studies can rely on data collected in several different ways: observational notes and logs collected by a researcher in person, manual annotations of video recordings of interactions between people and robots, and robot logs from interactions with people.

In-person observation provides the possibility for researchers to have a better understanding of the broader context of interaction because they can see and hear things that might not initially be in the data-collection protocol. This can lead to amendments to the protocol or can be represented in notes that can help guide later analysis and interpretation of the data. In-person observation, however, is limited by the sensory capabilities of observers at the time of coding and does not allow for others to go back and review the coded observations. In terms of establishing interrater reliability (i.e., to what extent various people agree on an interpretation of an observation; e.g., was it a "social behavior" when a passerby moved out of the way to let the robot pass through?), more than one coder needs to be present in the context at the same time, which can be inconvenient and become a distraction to other people in the space because of the presence of multiple researchers.

Video coding, on the other hand, allows researchers to review observations as many times as needed, potentially revise their coding schemes, revise their codes of observations, and easily provide data to a second coder for establishing interrater reliability. Video, however, has a limited view defined by whatever is visible from the chosen camera angle. This may cause researchers to miss some relevant aspects of the interaction, so it is important to clearly define what the camera should be focused on before the video observation starts so that important things are not missed. Although video coding may seem more convenient and preferable overall, some contexts (e.g., nursing homes, hospitals, or schools) may not allow researchers to record video, so in-person coding may be necessary.

Finally, robot logs are limited by the robot's ability to sense and categorize different human actions but have the benefit of being able to provide data about both the robot's state and actions and the human actions it perceived at the same time. It is, of course, possible to combine these different data sources to improve the accuracy of the data.

Both in-person coding and video annotations require the development of a coding scheme that coders will follow systematically. This coding scheme can be developed based on theoretical or practical interests and expectations, or it can be developed in a bottom-up manner by identifying points of particular interest in a portion of the data and then going through the rest of the corpus to understand related patterns. It is very important to pilot test the coding scheme to identify missing components and overlapping or unclear codes so that coders can be in clear agreement about what the codes mean before they start (particularly for in-person coding, where you can't go back to view the interaction). Video analysis is also quite labor intensive, so properly defining how fine-grained you need the coding scheme to be can save time and effort. Aside from providing frequency counts of certain types of behaviors or identifying qualities and patterns of interaction, observational coding of interaction behaviors can also provide particularly interesting temporal patterns of behavior, which can show the effects of certain robot behaviors on people's actions (e.g., how a particular gaze cue by a robot is followed by a joint-attention behavior by a person).

10.2.5 Ethnographic studies

Along with behavioral observation, HRI researchers also engage in more in-depth and often long-term ethnographic observations, in which they not only seek to identify certain behavioral and interaction patterns among humans and robots but also to understand what those patterns mean to people and how they are connected with the broader environmental, organizational, social, and cultural contexts in which those interactions take place. Ethnographic observations can include all aspects of interactions between people and robots, including behaviors, speech, gestures, and posture. They also include information on the context in which those occur, including the daily practices, values, goals, beliefs, and discourse of different stakeholders, which include but are not limited to people who directly interact with the robot.

Whereas behavioral observation is inspired by ethology and the desire to explore and build explanatory models of animal and human behavior, ethnographic observation is based on the theory and practices of anthropology and the goals of understanding sociocultural experiences holistically. Ethnographic observation is often performed over longer periods of time, from a few months to a few years, which is necessary for the observer to get a more complete and emergent sense of the cultural logic of the research site. Ethnographic studies can be performed by participants as outside observers but also through participant observation, where the researcher takes part in the activity under study to better understand the experience. The former type of study is currently more widely represented in HRI, although social studies of robot design often take the latter approach. Ethnographic study is also often coupled with a "grounded theory" approach to data analysis, which assumes that the collection and interpretation of data are ongoing throughout the project, with the researcher regularly engaging in reflection on the questions that guide the research, methods of data collection and analysis, and potential interpretations of the data, thus iterating as the study goes along.

Ethnographic studies are still relatively rare in HRI, partly because of the labor involved in collecting data over longer periods of time but also because there have not been many robots that are technically capable of taking part in long-term interactions with people. Some successful examples of ethnographic studies include a one-year-long study of a service robot in a hospital that showed that the patient type in the context, oncology or postnatal, determined whether the robot was appreciated or hated (and sometimes kicked and sworn at) by nurses (Mutlu and Forlizzi, 2008). Forlizzi and DiSalvo (2006) did an ethnographic study in which they gave families either a robotic Roomba vacuum or the latest version of a conventional vacuum to use over several months. They learned that people treated the robot, but not the conventional vacuum, as a social agent and that having a robotic vacuum changed the way the family cleaned, particularly inspiring teenagers and men to participate. Leite et al. (2012) performed an ethnographic study with a social robot that could respond empathically to children in an elementary school. The study found that the task scenario and children's specific preferences influenced their experiences of the robot's empathy. Several ethnographic studies have also been performed with scientists using robots. Vertesi (2015) studied National Aeronautics and Space Administration (NASA) scientists' interactions with a remote Rover and showed how the organizational structure of the team affected the team members' use and experience of the robot. The study also showed that scientists performed aspects of the robot's behaviors with their own bodies, creating a team identity for themselves in the process.

Ethnographic studies are particularly valuable because HRI is a young field and thus is still developing a corpus of theoretical and empirical work that can identify the most relevant factors we need to pay attention to, not only in the design of robots but also in their implementation in different environments.

10.2.6 Conversational analysis

Conversational analysis (CA) is a method in which the verbal and nonverbal aspects of an interaction are reported in great detail (Sidnell, 2011). This is not limited to conversation only, as the name might imply, but can be applied to any form of interaction between people or between people and technology.

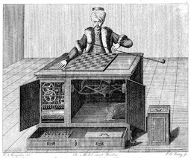
The process of CA starts by recording an interaction between two or more parties. Whereas this used to be audio recording, nowadays, video recording is more convenient, and several cameras can be used to capture the interaction from different angles. The participants being recorded might or might not be aware of the recording. From the recording, a very detailed transcription is produced, including turn-taking cues such as pauses in conversation, emotional cues such as laughter, behaviors performed while conversing, and other details of the interaction. Depending on the research question, the temporal resolution of the transcription can be brought down to the frame rate of the video recording. This can capture small actions, such as blinking and other eye movements, gestures, and changes in body posture. Fischer et al. (2013) used CA to investigate how the contingency of robot feedback affects the quality of verbal HRI. In their experiments, participants instructed the humanoid robot iCub how to stack some shapes in a contingent and noncontingent condition. Analysis of participants' linguistic behaviors, including verbosity, attention-getting tactics, and word diversity, showed that contingency had an impact on the participants' tutoring behaviors and therefore can be important for learning by demonstration.

CA will pay specific attention to elements in the verbal interaction, such as turn-taking, back-channeling, overlap of speaking, repair statements, echo utterances, and discourse markers. In HRI, CA can be used to analyze in great detail how people interact with social robots and whether they employ similar conversational strategies with robots as they do with people.

10.2.7 Crowdsourcing participants

HRI studies make extensive use of crowdsourcing to collect data and run studies. Crowdsourcing is the practice of obtaining responses from a large number of people, either paid or unpaid, via online methods. In recent years, the use of online crowdsourcing platforms has allowed researchers to run user studies and gather large amounts of data with relatively little effort and to gather data from subjects they would typically struggle to reach (Doan et al., 2011). The online platform can be entirely built by the researchers, but more often, existing online tools are used to recruit, run, and analyze user studies. The most widely used tools are Amazon Mechanical Turk (MTurk or AMT) (see Figure 10.3) and Prolific. These services allow you to post jobs: usually, short user studies in which participants are asked to watch a number of images or videos containing robots or interactions with robots and then answer questions about the material.

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



Crowdsourcing allows researchers to gather large amounts of data in a short time frame and for a modest cost. Taking part in a study will earn each participant a small financial reward, typically only a few U.S. dollars, with the price set depending on the complexity of the task, the time it is expected to take, and the quality rating of the respondent.

Increasingly, crowdsourcing is being used to evaluate the technical aspects of robots. Interactive robots often need to display behavior—such as gaze fixation, back-channeling or co-speech gestures—that is difficult or even impossible to objectively evaluate. There is no equation to capture how good co-speech gestures are, and there is no formula to say how empathetic a robot's voice sounds. Instead, subjective evaluations are used. In this technique, people are asked to rate the behavior of the robot, and crowdsourcing offers an effective and cheap method to collect responses from a large variety of human raters (Wolfert et al., 2022).

Running crowdsourcing studies comes with its own set of unique challenges, though, the most important being the relatively low level of control the experimenter has over the subjects taking part in the study and the environment in which the study is executed. Any account that meets the broad inclusion criteria set by the crowdsourcing platform is allowed to take the job. However, the account that is logged in might not be being used by the actual person registered as taking part in the study. Participants could take your study while pursuing an array of other activities, such as eating ice cream while petting a cat, or they could be full of caffeine or sitting on a crowded bus while listening to loud music on headphones. Crowdsourcing is also open to malicious user behavior: participants often provide low-quality or deliberately incorrect responses.

To avoid some of these problems, it is good practice to include verification questions in your user study (Oppenheimer et al., 2009). These questions check whether participants pay attention and are engaged with the task. When showing a video, a number could be displayed for a few seconds, after which the video participants are asked to enter the number. Questions can also be used to ensure the participant is responding to the questions rather than just picking random responses, such as "Please click the third option from below."

After data collection, it is necessary to separate the wheat from the chaff. A first filter will be the responses to the verification questions; another

This material has been published by Cambridge University Press as Human-Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781009424233 (https://www.cambridge.org/9781009424233).

method is to exclude all responses that took less than a reasonable amount of time. For example, if you believe the study should take a minimum of 15 minutes, then any responses that are far under that time should be disregarded. Some crowdsourcing platforms allow you not to reward participants if their responses are of insufficient quality, which not only leaves those participants without pay but also negatively affects their ratings. This has shown to be an excellent incentive to improve the quality of responses. Given that data collected using crowdsourcing are inherently more variable than data collected in the lab, one way of addressing this problem is to collect more of these data.

Although crowdsourcing has been successfully used to replicate results from lab studies in social psychology, linguistics, and behavioral economics (Bartneck et al., 2015a; Goodman et al., 2013; Schnoebelen and Kuperman, 2010; Suri and Watts, 2011), the value of crowdsourcing to HRI needs to be considered on a case-by-case basis. Sometimes the physical presence of a robot is key to the participant's performance, precluding the use of crowdsourcing. Sometimes the effect you are measuring is small and would not show up when sampling a large and diverse population. Sometimes the population you need is scarce on crowdsourcing platforms, such as elderly users or Swedish primary school teachers. Sometimes the task requires a certain level of language proficiency. Crowdsourcing has its place in HRI research, but it should be used with care and consideration.

Computer-based studies in general come with some problems: Participant age or technical affinity may play a role—for example, seniors and very young participants might not be highly familiar with computers that are commonly used to collect data. At the same time, depending on the age and cognitive abilities of participants, they might be more or less able to understand what we think we want to measure. For that reason, new variants of questionnaires may be required if you study participants with mild cognitive impairments or if you study children. However, using plain language (Stoll et al., 2022) not only benefits the aforementioned audience but also nearly everyone who might lack reading skills or might be largely unfamiliar with a topic.

10.2.8 Case Studies

Another type of study to consider in HRI is the case-study research design. In this type of qualitative study, researchers compare the effects of an intervention on a single participant rather than a group of people. This is done by initially collecting baseline measures of the individual's behavior, which are compared with the subject's behavior during and after the intervention.

Case studies are used when recruiting large numbers of participants is difficult because of their rarity in the population or when individual differences between subjects are large and relevant to the phenomenon of interest. Multiple participants can be recruited for case studies, but the number of subjects is often small, and for the sake of analysis, each participant is treated as his or her own control.

Case studies are commonly used in medical and education research fields, and in the case of HRI, they are used in research on the effects of robots on individuals with autism. For example, Pop et al. (2013) performed single-case studies with three children to investigate whether the social robot Probo can help children with autism spectrum disorders better identify situation-based emotions. Tapus et al. (2012) similarly worked with four children with autism to see whether they would show more social engagement with the Nao robot than with humans, and they found large variability among their responses. This shows the importance of performing single-participant studies in cases where individuals of interest, such as those diagnosed with autism, vary widely in their behaviors; in such cases, averaging the responses of a group could mask important intervention effects because different individual responses would cancel each other out when aggregated.

10.3 Selecting research participants and study designs

10.3.1 Representativeness of your sample

Because people are a necessary component of HRI studies, several important decisions in HRI studies must be made regarding the participants in a study. One is who the participants will be. The usual suspects for empirical HRI research are university students because they are the most convenient population to access for academic researchers, have time for and interest in participating in studies, and are usually in close physical proximity to the laboratories where much of the HRI research is performed.

It is, however, important to consider the limitations of using university students as a "convenience sample," particularly in relation to the research questions posed. In an ideal world, we would aim for a large, representative sample of potential end users of robots so that we can claim that our findings hold for a wide range of users and have *external validity*—that is, they can tell us something about people and robots in situations outside the study itself. Such samples are very difficult to bring in for experimental studies but might be more achievable in surveys. In studies of the general perceptions of robots, HRI, similarly to psychological research, assumes that university students are "close enough" to the general population in terms of characteristics when it comes to broad social traits (e.g., stereotyping), cognitive performance (e.g., memory), and attitudes (e.g., fear of robots). Even when using university students, it is important to be mindful of and balance certain characteristics of the sample, such as gender or educational background, depending on whether these factors might be expected to have an effect on your results. For example, students in a computer science department would likely be seen as having more positive attitudes toward robots and having greater ease in using computing technology than a broader student population or the general population of potential users.

If your research questions relate to studying the characteristics of a specific population, such as older adults, or to investigating the effects of robot applications in specific domains, such as the treatment of children with diabetes, your choice of participants will need to be more specialized. The specificity of your research question and the claims you want to make will guide the level of specificity of your sample. It is not possible, for example, to claim that a robot will have positive effects on older adults experiencing cognitive decline if you run your study with university students or even with older adults who are not experiencing cognitive decline. A university student sample will also not be sufficient for investigating the use of robots to support learning in young children. Thus, before running your study, you need to make a careful decision about what kinds of people should take part in it. You will also need to consider how to get access to this population and how to recruit and motivate individuals to be in your study. You should also consider whether you will be able to bring people from this population to your lab, whether you need to go to another place to have contact with them, or whether an online study might be appropriate.

10.3.2 Sample size

Another consideration regarding research participants is the number of participants you might need to answer your research questions (Bartlett et al., 2022). This will depend both on the type of study and analysis you are doing (quantitative vs. qualitative, survey, experiment, or interview) and on the population you are working with (e.g., university students, or older adults, or children with diabetes). It is difficult to reliably test for an effect with a small sample size because people will always differ a little bit from one another. In a study on gender stereotypes, for example, some participants will consider all robots a bit more “warm” than others; other participants will think all robots possess typically “male” qualities. Such differences, which naturally occur in people, will add noise to the data. Unless the manipulation has an extremely large effect, the data that we gather from a small sample will not be enough to reliably detect an effect. The differences among people might cancel each other out, or the variability of their responses might be too large. If you want to reach a valid conclusion about cause and effect, you need to determine the right sample size for your study design.

How many participants you need to reliably find a difference between conditions also depends on the type of design you use. When using a *between-subjects design*, participants are randomly assigned to a condition. In our example, one group of participants would be presented with the “male” robot, whereas the other group of participants would be shown the “female” version. After answering questions using a Likert scale, the mean scores of each group can be compared. Alternatively, in a *within-subjects design*, one group of participants is exposed to both versions of the robot prototype and asked to evaluate both. Because the same person provides two evaluations, you cut down on the “noise” in your data, and the number of participants required will be lower for this design. However, not all research questions are suitable to be answered with a within-subjects design. For example, if you want to test if people recover faster from a broken leg when they have a robotic assistant that does walking exercises with them every day, you can hardly have them first heal on their own and then break the other leg so that they can recover again with their robot helper. Also, researchers have to be mindful of the order effect that may occur; maybe people will always like the first robot better than the second (e.g., because of the novelty). Thus, it is a good idea to *counterbalance* any conditions when running a within-subjects design. This means that half the participants will first interact with the female robot and then with the male, and vice versa for the other half.

To approximate a sufficient sample size to establish a statistical effect of the desired size, the internet offers a variety of tools, such as G*Power (Faul et al., 2007). However, researchers may not always be able to meet such recommendations because they are also constrained by the availability of resources, such as time, money, robots, and potential participants.

Studies that involve special populations, such as older adults with depression, may have to make do with a smaller number of participants because of the acknowledged difficulty in recruiting specific populations. In some cases, such as studies of children diagnosed with autism, where the participants are also widely diverse in the way they express themselves and experience the world, it is possible to treat participants as individual cases and study changes within each participant's behaviors and responses.

For qualitative studies, rather than focusing on a particular number of participants needed, the rule of thumb is to try to achieve "saturation" of the analytic themes and findings. The idea here is that the researchers can stop collecting new data once they find that the data they are collecting are simply adding to and repeating existing themes and findings rather than creating new ones. Although it is relatively easy to understand, this concept can be more challenging to operationalize and measure, so scholars have developed various ways of defining and quantifying data saturation in various studies (e.g., Lowe et al., 2018; Guest et al., 2020).

10.4 Defining the context of interaction

10.4.1 Location of study

For HRI in particular, an important distinction is between studies performed in the lab versus those performed in the field. Especially in the early years of HRI, the majority of research was performed in the controlled environment of the lab. Although robotic technology has certainly advanced over the years, and there are now robotic platforms robust enough to use outside of the lab, so-called "in the wild" studies are still relatively rare compared with the number of studies performed in the lab.

Studying interactions outside of the laboratory is important for understanding how people might interact with robots in natural circumstances, determining what kinds of HRI might emerge in those circumstances, and investigating the potential broader social effects of new robotic technologies.

On the other hand, laboratory studies benefit from the researchers' ability to strictly control the context and nature of people's interactions with a robot—the introduction, task, environment, and length of the interaction can be clearly defined by the researchers. In the lab, participants are asked to interact with the robot only in the way researchers suggest. This allows for the strict manipulation of desired variables.

In contrast, field studies are more flexible in what can happen and are therefore closer to what might occur in day-to-day HRI. In the field, participants can choose how, when, whether, and why they want to interact with a robot; they can even ignore it. Field studies, therefore, provide a space in which to observe and discover new emergent phenomena, new variables of interest and significance to interaction, and the form and consequences of HRI when it is outside of the researchers' control. Field studies also effectively show how complex interactions between different contextual variables, such as institutional culture or interactions among people, might affect the interaction.

10.4.2 Temporal context of HRI

A related distinction that has grown in importance in HRI is whether researchers are studying short-term or long-term interactions between people and robots. The majority of lab studies, by necessity of their design, focus on "the first 10 minutes of HRI"—how people respond to and make sense of their first introduction to a robot. Researchers widely acknowledge, however, that people will change their attitude toward the robot as time passes, and consequently, the way they interact with the robot will change as well. The first interaction suffers from the *novelty effect*: people are generally not familiar with robots, so their initial reactions might be quite different from their reactions over a longer period of time. Short-term studies therefore have limited validity in informing us about how people and robots will interact over a longer period of time. They do, however, tell us about the kinds of characteristics of people and features of the robot that will affect the initial encounter. Such studies are important for setting up a positive feedback loop of interaction, which can then support more positive effects in long-term interaction. Studies of longer-term interactions, which can take place over several days, weeks, months, or in a few cases, even years, allow us to see how interactions between people and robots develop and change over time, how robots are integrated into human social contexts, and how social interactions between people themselves may change because of the presence of a robot.

10.4.3 Social units of interaction in HRI

Interactions between people and robots can be studied through several different social units of analysis, which the social sciences see as distinct in terms of the aspects of cognition and interaction they enable (see Figure 10.4).

This material has been published by Cambridge University Press as Human-Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781009424233 (https://www.cambridge.org/9781009424233).

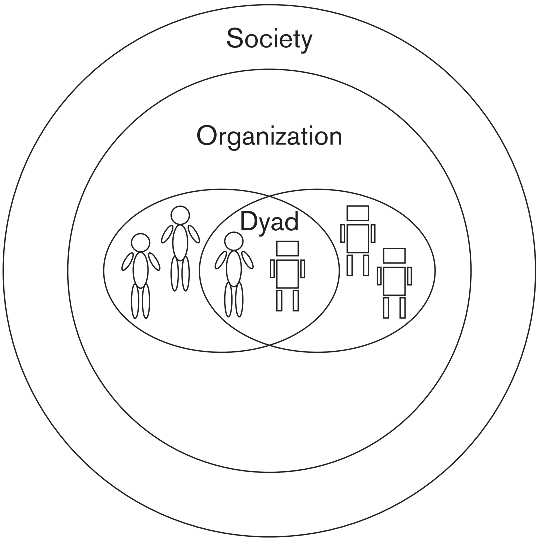


Figure 10.4 Units of analysis in HRI.

The most common unit, so far, has been the interaction dyad—one person and one robot interacting with each other. This is partly due to the early constraints of HRI—robots were difficult to procure and difficult to maintain and operate; hence, the most common form of HRI study was the lab experiment involving a single participant interacting with a single robot.

Figure 10.5 Robovie in school.

As early as 2006, the Robovie robot was one of the first robots capable of supporting group interactions at an elementary school (see Figure 10.5). It taught children English and tracked their social networks over time, keeping the children interested in interacting with the robot by unlocking secrets (Kanda et al., 2007b).

As robots have become more readily available and capable of interacting with more people and in more open-ended, naturalistic environments, the unit of analysis in HRI has expanded. Early studies of HRI “in the wild” showed that people actually often interact with robots not individually but in groups, a task for which most early robots were poorly equipped (Šabanović et al., 2006). Increasingly, HRI studies group interactions involving two or more people, both inside and outside of the lab. For example, Leite et al. (2015) found that children were better able to recall information from a story told by a group of robots when they interacted with them individually rather than in a group of three. Brscić et al. (2015) showed that children who come across a robot in a shopping mall abuse the robot only when they are in groups but not individually.

Social scientists distinguish between dyadic interactions and group interactions, and they consider the cognitive and behavioral aspects of each to be different. Groups bring in new perspectives on group effects, multi-party collaboration, team dynamics, and other such effects. Our vision of how we

This material has been published by Cambridge University Press as Human-Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781009424233 (https://www.cambridge.org/9781009424233).

will be interacting with robots in the future also presupposes that there will be many robots in our environment, so another aspect of group HRI studies has been exploring how multiple robots can interact with people, whether in teams, in swarms, or simply as co-present robotic actors.

When robots collaborate in teams, they are often perceived as having more social agency. For example, Carpenter (2016) found that robots used in military bomb-disposal teams were often seen by the soldiers as members of the group and that soldiers became attached to such robots, even expressing feelings of sadness when their team's robot was destroyed.

The increasing availability of robots for research in applied settings beyond the laboratory opens up another unit of analysis. That is, we can look at how HRI occurs within organizations, such as educational and nursing institutions or even the military. By studying HRI within organizations, it is not only possible to see the effect of individual factors on HRI but also the effect of the broader context, such as how existing labor distributions or roles affect the robot's function and its acceptance by workers, how the robot is adapted to existing practices, and how institutional values affect people's interpretations of the robot.

Mutlu and Forlizzi (2008) showed that introducing a robot into an organization, for example, reduced work for some while increasing it for others. At the same time, it is plausible that people in different roles (e.g., manager, nurse, janitor) can have different perceptions of a robot based on how it affects their work. In another ethnographic study on the use of the seal-like Paro robot in a nursing home, Chang and Šabanović (2015) showed that having even one person who acts as an advocate for the robot in an organization can lead to more people committing to try it out and make it work for them, by modeling positive experiences of using the robot and creating a "positive feedback loop" that supports the long-term adoption of the robot. An organization can also be set up in a particular way to support the functions of a robot. Vertesi's (2015) ethnographic study of the NASA Rover team showed that the need to balance the robot as a scarce resource shared by many different scientists and engineers worked well with an egalitarian setup of the team, where all team members needed to agree and say they were "happy" about the robot's next move. Now that it is possible, studying interactions between people and robots from an organizational standpoint seems necessary for the further development of the field and for our ability to design appropriate robots and social structures for the successful application of HRI in the real world (Jung and Hinds, 2018).

10.5 Choosing a robot for your study

Along with deciding how many and what types of participants you need to answer your research question, you will also need to decide on the characteristics of the robot(s) you need to use in your study. Factors you will need to decide on include the robot's appearance, functionality, and ease of use, among others. Whereas some of these decisions might be based on practical constraints, such as what types of robots are available to you or how much it would cost to purchase a new one, others will be guided by your research interests.

Robots can be seen as research tools, with which you can manipulate factors of interest and observe the effects of such manipulation on the outcome variables you want to measure. This approach is at the heart of experimental HRI research but can also be useful for more exploratory studies in which you may want to see if certain design factors produce differential effects on HRI. In order to use robots as a stimulus in HRI studies, we can manipulate their appearance, behavior, and communication mode and style, as well as their role in the interaction, among other characteristics. HRI researchers often use off-the-shelf robots for their studies, but they also sometimes design and test their own prototypes. When deciding what kind of robot to use, determining which hardware and software capabilities would be best for the study and the appropriate level of autonomy of the robot are important considerations.

There are some commercial robots that lend themselves well to HRI studies, such as the Nao (Aldebaran Robotics), Furhat (Furhat Robotics), QTrobot (LuxAI), or Paro (Intelligent System). Even when using a commercial robot, getting your robot up and running will require some basic programming skills. The Nao and the QTrobot can be programmed using a visual programming environment, which allows you to quickly go from the drawing board to a working robot. However, knowledge of more advanced control software and programming languages, such as the Robot Operating System (ROS), will allow you to greatly extend the repertoire of the robot's behavior and enrich the interaction. ROS contains a number of packages that implement sensory perception and visualization for different types of robots.

10.6 Setting up the mode of interaction

There are dozens of ways in which people and robots can be brought together for a study. People can meet an actual robot, or they can be shown pictures or videos of a robot. The robot can be fully autonomous or can be tele-operated by the experimenter. People can come to the lab, or the scientists can get out of the lab and bring their robots to the people. Sometimes, a single data point is all that is needed; on other occasions, only thousands of data points will do.

10.6.1 Wizard of Oz

In some HRI studies where the development of autonomous capabilities for the robot is not the focus of the research at hand, researchers commonly rely on the Wizard-of-Oz (WoZ) technique. WoZ involves deceiving study participants into thinking the robot is behaving autonomously, when it is actually being operated by a member of the research team. Research participants should then be informed about this deception in a post-experimental debriefing (see also Section 10.9).

Using WoZ, researchers can "pretend" that their robot has interactional skills that it does not have, either because they require further technical development or because additional time or skill must be expended on programming the robot. The WoZ approach is particularly suitable in situations in which technology has developed to a degree at which it is almost usable for HRI, such as speech recognition. Using a wizard to recognize the users' utterances makes an experiment more robust and the robot's behavior more realistic and believable, enabling an actual interaction flow. It could, however, be considered problematic to completely fake an AI system that can uphold a serious and prolonged conversation because that would be considered a very unrealistic level of capability for the robot.

WoZ can also be used to test people's perceptions of more advanced capabilities, such as a robot that can understand and respond to the social context in very nuanced ways (e.g., see Kahn et al. (2012)). For experimental studies, it is important to constrain the wizard's behavior so that the robot's behavior is kept consistent across conditions and does not introduce additional variation that can confound the analysis. WoZ can also be used as a way to collect data from participants to help develop a robot's design or autonomous capabilities (e.g., Martelaro and Ju, 2017; Hu et al., 2023; Sequeira et al., 2016).

> The WoZ method is named after a character in the movie of the same name. Dorothy and her companions set out to find the all-mighty Wizard of Oz who can return Dorothy to Kansas. They encounter the wizard in his castle and are afraid of his gigantic appearance, his authoritative voice, and the smoke and fire he emits. Only when Dorothy's dog, Toto, pulls away a curtain do they notice Professor Marvel, who is operating the machinery that controls the wizard. In HRI research, wizards often hide in the background and control the robot, giving the robot the semblance of having more advanced autonomous capabilities than it actually has. We all hope not to encounter Toto and be found out.

10.6.2 Real versus simulated interaction

Although the ideal way to gauge people's perceptions of and response to robots is in real-time, face-to-face interaction, it is still common for HRI researchers to present their participants only with video or photos of robots. In the field of HRI, there has been considerable discussion on whether video recordings of robots can be used as a replacement for live human–robot interactions. Whereas Dautenhahn et al. (2006) argue that the two interaction styles are broadly equivalent, Bainbridge et al. (2011) conclude that participants had a more positive experience interacting with physically present robots than with a video representation. Powers et al. (2007) also found large attitude differences between participants interacting with a co-located robot in comparison to a remote robot. Therefore, the use of visual stimuli alone limits the generalizability of study findings but can be appropriate for exploratory studies of the effects of certain factors (e.g., perceptions of different robot forms; see DiSalvo et al., 2002) or for studies in which accessing the appropriate population can be difficult, such as cross-cultural samples. Using videos to present robots to participants can also enable researchers to avoid problems associated with a less controlled experiment that involves actual interaction. Finally, videos and photos are particularly amenable for use in studies that take advantage of online participant pools, whether through universities, word-of-mouth referrals, or services like Amazon's Mechanical Turk.

10.7 Selecting appropriate HRI measures

In HRI, as in psychology and other social sciences, researchers commonly distinguish between direct versus indirect measures to assess attitudes toward people or objects. In the example of the "gendered" robot study described earlier, the study design relied on *direct measurements* of the dependent variables—asking participants to rate the robot's warmth and authoritativeness, for example.

Within both correlational and experimental studies, self-reports are often used to assess the constructs of interest, such as concepts or variables. Self-report measures commonly bear high face validity, meaning that people usually directly know what the researchers want to measure when they read the items of the given questionnaire. On the other hand, this makes it easy for participants to amend their actual opinion with the aim of pleasing researchers, to represent themselves in a positive light or "be a good participant." This aspect also holds true for interview techniques, which are a way to gather an even more holistic picture of participants' thoughts and feelings toward both humans and robots. Interviews can be structured or semistructured in nature. In structured interviews, the interviewer asks a set of predetermined questions, often in a specific order, whereas in a semistructured interview, the interviewer has more leeway in deviating from the script; for example, some questions may be planned, but others may arise spontaneously during the interview. Both types often use questions to which interviewees can respond in their own words. Such open-ended responses, however, require labor-intensive coding after transcription of the interview's content. Such interviews might be a useful complement to questionnaires, though, as illustrated by de Graaf et al.'s (2017) use of data from a long-term survey and an interview to explore the reasons why people choose not to use a communication robot in their homes. As their work has shown, a research participant might feel highly uncomfortable in the presence of an unfamiliar robot.

In some cases, however, participants might be reluctant to report their true feelings and attitudes on a questionnaire or when talking directly to an interviewer. They may also not be aware of and able to report some unconsciously held beliefs. In that situation, it might be useful to complement your set of direct measures with indirect ones. Reaction times are often used as a proxy for factors that are harder to measure, such as attention or engagement. Indirect measures can include the use of eye tracking as an indicator of attentional focus and cognitive processing or the use of physiological measures such as heart rate or skin conductance to give researchers an idea of participants' level of stress experienced during HRI. Whereas computerized measures of attitude (e.g., a variant of the so-called Implicit Association Test[2] to measure anthropomorphization) have become increasingly popular, physiological correlates of attitudes toward robots or other technologies are less frequently used in contemporary research. Computerized and physiological measures are often more difficult to administer and require specific equipment, and ultimately, the findings are not always interpretable in an unambiguous manner. For example, skin conductance can indicate that someone is excited, but it cannot reveal whether the excitement is due to fear or enjoyment. In addition, a study in which the skin conductance of participants was measured as they interacted with a Nao robot showed that skin conductance readings are, unfortunately, not very conclusive (Kuchenbrandt et al., 2014).

To circumvent difficulties in interpreting results, it is helpful to use a combination of direct and indirect measures or several indirect measures at once in one study to ensure that you are indeed measuring the construct, or variable, that you intend to measure. As a researcher, you should aim to establish that all measurements used in your research reliably and validly assess what they are supposed to capture. This can be done by carefully pilot testing your study designs and measures used, developing and even formally validating new measures, or using widely accepted and validated measures that you find in the literature.

10.8 Standards for statistical analysis

> *"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."*
>
> As the famous quote by statistician Sir Ronald Aylmer Fisher (February 17, 1890–July 29, 1962) points out, the earlier on you ask for advice on your experimental design and analysis, the more useful it will be. Most universities offer statistical consultation of some sort, but even informal discussions with peers and professors may prove of tremendous value.

Although statistics has a reputation for being confusing and incomprehensible, in reality, most statistical tests are built on three main measurements: tendency, variability, and number of observations. To understand how these three things influence statistical testing, imagine that you're trying to decide

[2] See https://implicit.harvard.edu/implicit/

which of two restaurants is better. You have never actually been to either, but you can pull up reviews easily enough to compare. What would help you decide whether restaurant A is actually better than restaurant B? Well, obviously, you would first look at the average reviews. If restaurant A has an average of 4.8 stars out of 5 and restaurant B averages 3.2 stars out of 5, you will be fairly certain that A is better than B. The closer those averages are to one another, the less certain you will be that one restaurant is actually better than the other. This would be an indication of the difference in *tendency* between both groups.

But that is not everything that you will want to take into account. If you see that restaurant A has an average of 5 stars but only 3 people left a review, whereas restaurant B has an average of 4.7 stars from over 1,000 reviewers, you may still opt for restaurant B. This would be because you—quite reasonably—assume that with so many reviewers, you get a better estimate of the “true” quality of the restaurant. This is the influence of *sample size*: the more responses we have, the more certain we can be that the tendency is an accurate representation of the truth. Another example you can think of here is trying to figure out whether or not a coin is fair. Getting heads 75% of the time will not let you say for sure if you flipped it only four times (and got three heads and one tail), whereas the same percentage of heads would be pretty convincing if you had flipped the coin 1,000 times.

And finally, there is the matter of variability. Say that restaurant A and B both have an average of 4.2 stars and both have the same number of reviews, but for restaurant A, these reviews range from 1 star to 5 stars, whereas restaurant B has mostly 4-star ratings with a few 5 stars. For which restaurant would you be more certain that the 4.2 star is an accurate indication of the quality? This is the importance of *variability*: the more variable the results, the less certain we are that our sample mean is an accurate indication of the “true” effect.

These three measurements—tendency, sample size, and variability—are often called *descriptive statistics*. They give a summary overview of data without yet comparing conditions or calculating correlations, and they should be used as the first stage of data analysis. Always provide means (which indicate the tendency), standard deviations (which are an indicator of variability when the data have a normal distribution—if this is not the case, you can provide a range), and the number of participants (sample size). In addition, demographics (e.g., age and gender) will give your reader an idea of whether your sample resembles the general population, and excluded data points, together with the reason for exclusion, need to be reported for integrity and transparency.

Next, your study will probably require inferential statistics. Most classical statistical tests combine the tendency (often the mean), sample size, and variability into a test statistic, which in turn is used to calculate the p-value: the probability of getting the data at hand if there had been no true effect. Going back to the restaurant analogy, the p-value indicates how likely it would have been to get the reviews we got *if both restaurants had been equally good*.

The smaller that chance is, the more confident we may be in the hypothesis that one restaurant is in fact better than the other. This is the logic behind null hypothesis significance testing (NHST). Different study designs warrant different kinds of statistical tests to get to the p-value. Although going into the details of the extensive number of statistical tests and procedures is beyond the scope of this chapter, the interested reader might consult the readily available literature, such as the work of Andy Field (2018).

Until recently, science relied on NHST to report on the importance of results. If the probability of the data under the null hypothesis is small enough (i.e., p-value is less than or equal to some threshold, typically 0.05), the result may be considered "significant," and the null hypothesis would be rejected in favor of the alternative hypothesis. On the face of it, this provides a useful means of characterizing the success (or failure) of a method or intervention.

The definition of the p-value may sound formal and confusing, but you have probably applied an intuitive version of it before. For example, take a look at the following headline, published in the *Moscow Times* (2020) at the start of the coronavirus pandemic: "Third Russian Doctor Falls from Hospital Window after Coronavirus Complaint."

Reading this headline may have made you wonder whether this unfortunate accident was indeed only that, an accident. Your suspicion would stem from your inference that under the null hypothesis (i.e., if there had been no conspiracy against critical doctors), the probability of three of these incidents in a row would have been quite low. Although you did not calculate a concrete value, this is, in essence, what the p-value boils down to.

10.8.1 Making sense of statistics

There are a few common misunderstandings and often-overlooked implications in NHST, which have given rise to a recent questioning of the overreliance on NHST and p-values (Nuzzo, 2014).

Assuming a threshold of $p \leq .05$, this still means that 5% of the time where the null hypothesis is true (i.e., there is nothing going on), the obtained data will look as if there is an effect. This would constitute a *Type I error*, or false positive. Because false positives look exactly like true positives, even a significant result cannot be taken as conclusive proof that there is an effect.

Moreover, the p-value is often wrongfully taken to indicate "the chance of a Type I error." This complete misinterpretation of the p-value is pervasive and widespread among both students and academics (Badenes-Ribera et al., 2015; Lyu et al., 2020). In reality, the p-value only indicates the chance of a Type I error *if nothing had been going on* (i.e., conditional on the null hypothesis), and the overall chance of the results being due to a Type I error cannot be computed.

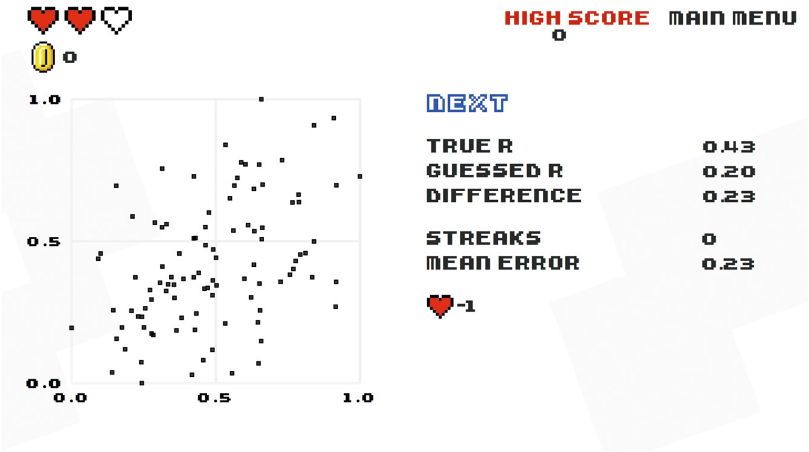
A fundamental issue with NHST concerns the inferences that one can and cannot draw from it. What is tested in NHST (the chance of finding the current data, provided that there is no true effect, or $p[A|B]$) is not what the researcher actually wants to know (the chance of a true effect, provided the current data, or $p[B|A]$). Although these *may* seem similar, their fundamental difference becomes clear when we consider sharks and death tolls. The chance of dying, provided that you are eaten by a shark, $p(dead|sharkbite)$, is pretty close to 1. However, the chance that you are eaten by a shark, provided that you are dying, $p(sharkbite|dead)$, is close to 0—for better or worse, most of us die from other causes than shark attacks. In his entertaining and remarkably accessible paper "The Earth Is Round ($p < 0.05$)," Jacob Cohen explains some of the problems with NHST in further depth (Cohen, 1994).

Related to the misunderstanding of the p-value is the misconception that p-values are stable; that is, if you conduct a study twice, you should get a similar p-value each time (Badenes-Ribera et al., 2015). Empirical results have suggested, and simulation studies have shown, that p-values are highly volatile in experiment replications. Repeating a study that has a significant p-value can result in the p-values of the replication study being in the range [0.00008,0.44] for 80% of the replication studies (Cumming, 2008). p-values are thus unreliable as a measure of how solid a result is.

Another common mistake is the conflation of a p-value with how big or important an effect is. The significance, size, and importance of an effect are three different things: a very small ("highly significant") p-value does not say anything about the size of the observed experimental effect. An effect size captures how large a change is between two conditions. It is calculated from the tendency and the variability of the data. The p-value, in addition, takes the sample size into account. Thus, the p-value can be considered an indication of how consistent the effect in the collected sample is, whereas the effect size indicates how large it is. These two measurements should both be seen as different from importance.

To illustrate the distinction between the three, consider the following situation: A new treatment has been developed for a medical condition. You compare this new treatment against the conventional treatment and find a significant yet small effect: recovery rates improve from 4% to 6%. What to make of this? Well, that depends. If the medical condition is foot fungus, you probably won't care much for a 2% higher chance of getting rid of your fungus. You would need a larger effect size to really care, especially if this new treatment is more expensive or has more side effects than the standard treatment. However, if the 2% increase refers to the chance of survival from a very aggressive kind of cancer, the very same effect size would probably be considered rather important.

Different statistical tests come with their own calculations of effect size; common effect sizes include Cohen's d (for a t-test), η_p^2 or ω^2 (for analysis



TRUE R 0.43
GUESSED R 0.20
DIFFERENCE 0.23

STREAKS 0
MEAN ERROR 0.23

Figure 10.6 If you had to make a guess, how strongly would you say the two variables in the plot are correlated? It has been shown that people find it very hard to infer the strength of a relationship from plots. On the website www.guessthecorrelation.com, you can try for yourself. (By the way, the correlation in the picture is *r* = .43, which is considered a medium effect.) (Source: Omar Wagih)

of variance (ANOVA)), and R^2_{adj} (for regression). For most effect sizes, there are guidelines available to help with the interpretation, which will provide a rule of thumb of what constitutes a "small," "medium," or "large" effect. For example, in Figure 10.6, a medium effect for a correlation is shown.

A final important implication from NHST involves capitalizing on Type I errors, also known as "p-hacking." You already encountered this in the discussion of spurious correlations in Section 10.1.2. The logic behind p-hacking is as follows: if a cutoff value of $p \leq .05$ is used, then if there is no true effect (i.e., under the null hypothesis), logically, one would expect a false positive 1 in 20 times. Thus, if you run enough tests, you will eventually find a significant result even if, in reality, there is none. If you then only report the significant results and leave out all the times where you found no effect, you can easily present your results as a valid new finding. p-hacking is especially problematic for observational studies because it is very easy to measure many variables and keep testing the relationship between different combinations of measurements until you find one that is significant.

10.8.2 *Good practices to overcome issues with classic statistical testing*

We can partially remedy these issues by reporting not only the p-values but also the confidence intervals (CIs) of our data. CIs do not compare data and therefore cannot be used to say if results are significant or not. Instead, they report on how confident we are that the population mean (which we estimate through the mean of our sample) lies between a minimum and maximum value of the CI. When reporting the 95% CI of data, this means that in a replication study, the mean of the replication data will have an 83% chance of being within the CI of the original experiment. Reporting CIs and effect sizes conveys additional information on the magnitude of an effect and the precision of the estimates provided. This information complements the significance and will help both you and your reader to make sense of the findings (Coe, 2002).

The p-value indicates the chance of a Type I error (false positive) under the null hypothesis. However, as mentioned in Section 10.3, the opposite is possible as well: a researcher can conduct an experiment, gather data, and then wrongfully conclude that there is no effect. This has been, not very creatively, named a *Type II error*, or a false negative. Type I and II errors can be avoided by making sure your experiment has sufficient statistical power to detect any true effects. Power depends on the same three measurements mentioned before—tendency, variability, and sample size—but of those three, only the last one is under your control. Thus, you have to make sure you collect either enough participants or enough data points per participant. This can be tricky, and the number of participants needed can increase dramatically depending on how complicated your study design is or how small the effect you're hoping to detect is. Software such as G*Power (Faul et al., 2007) allows you to calculate the power both before and after a study.

Another way to ensure that results are trustworthy and not the consequence of Type I or II errors is through replication. Psychology has recently seen a replication crisis (Maxwell et al., 2015), where a number of "established" effects failed to replicate. Although this is to be expected under NHST, p-hacking may have been partially responsible. In HRI, the reproducibility of research has been less prominent on the research agenda, but the recent concerns in the social science community have brought these topics into the purview of HRI researchers as well (Irfan et al., 2018). Replication of HRI results is also now more possible than before because of the wide availability of certain robot platforms (e.g., Nao or Baxter), in contrast to the earlier reliance by researchers on bespoke platforms.

Registration can facilitate replication and prevent p-hacking by forcing researchers to specify exactly what tests they are planning to run before collecting data (see p. 164). There has been a drive for sharing code for commonly available robots and, if possible, making the experimental procedures available to other HRI researchers in order to enable them to run the same experiment in their own labs, testing the generalizability of a certain research question across contexts (Baxter et al., 2016). Overall, the notion of generalizability is highly important, even though representative samples are hard to obtain in HRI research.

The choice of methodology also affects the degree to which we can generalize from our HRI studies in the laboratory to those findings obtained from field studies. Developing new robots, applying robots in different contexts, and understanding the potential consequences of robots for people in daily life may require a combination of the methods mentioned in this chapter. This does not need to be done in one research project or by a single researcher but could be accomplished by the HRI research community over time.

A final, radical way to overcome issues associated with NHST is to abandon NHST altogether. This can be done through the adoption of Bayesian inference, a method of statistical analysis that has been increasing in popularity (Van de Schoot et al., 2017). As noted in Section 10.8, NHST draws inferences conditional on the null hypothesis: How likely are these data if nothing had been going on? The outcome is dichotomous: a result is either significant, or it is not. In contrast, Bayesian statistics uses prior information to draw a hypothesis and updates this with the newly gathered data. The result is not a single estimate but rather a range of possible values and an indication of how much confidence can be placed in each estimate (Etz and Vandekerckhove, 2018). As a result, it is rare that a "hard conclusion" is drawn from Bayesian inference. Rather, previous beliefs are strengthened or weakened, depending on how the newly found data align with the prior data.

10.9 Ethical considerations in HRI studies

Last but not least, one important aspect to consider when dealing with human participants in HRI studies is the need to take into account the ethics of human-subjects research. Any research that involves human participants, whether correlational or experimental, qualitative or quantitative, online or in person, requires participants' informed consent before the research is started. That is, participants are informed about the nature of the study and what to expect, with an emphasis on the voluntary nature of their participation and information regarding the risks and benefits of taking part in a given research study. Before starting a study, either online or in the real world, participants have to declare that they understand what they will be asked to do and what will be done with the collected data and that they consent to participating. Many universities and institutions have specific guidelines on how participants can be recruited and informed about their participation in research studies. Researchers need to be aware of this and follow all policies to be able to present their results for publication after the study.

Sometimes, however, it is impossible to fully disclose the actual goals of the given research project. In that case, a cover story or deception is used. For instance, in WoZ studies, participants are led to believe that a robot can behave autonomously. In that case, it is key to provide post-experimental information, a so-called *debriefing*, to participants so that they do not go home from the study thinking that robots are currently able to function fully autonomously.

This is even more critical if a robot might provide the human interaction partner with fictitious feedback about the human's personality or performance. Of course, the participants then must be debriefed about the reason for providing made-up feedback, and they must be informed that this feedback was actually bogus. Again, this serves to ensure participants' psychological well-being beyond the duration of the study.

In the case of qualitative research, initial information about the study goals given to participants may be more cursory, but the common practice is to later inform study participants of the findings if they are interested. In some cases, researchers might even discuss their interpretations of the data with participants or collaboratively develop interpretations and future robot design and implementation guidelines based on the results.

In HRI research, we also have to consider the ethical aspects of having humans involved with robots—both in terms of physical and psychological safety and in terms of the implications an interaction could have for a given individual. Think, for example, of an elderly person who has had a robot in his or her home for a certain amount of time and might have gotten attached to the robot companion. Consequently, the day the robot is taken away, this will cause distress. Users' emotional reactions toward robots, the attachments they might build, and the void that results when the robot is taken away must be considered.

To make sure that you are complying with ethics regulations, you may consult with the various codes of ethical conduct, such as those provided by the American Psychological Association,[3] the American Anthropological Association,[4] or the Association for Computing Machinery.[5] Your university's ethics committee may provide more detailed feedback regarding your specific research study. Note that ethics approval is a requirement for publication in many scientific journals, so consider getting it before you start your data collection.

Along with ethical behavior toward research participants, researchers should also reflect on the ethical implications of their research aims, questions, and findings and make choices about what types of research to pursue, and how to go about it, with these implications in mind. Such ethical considerations can include questions about where to seek out and whether to accept funding, whether to participate in research that may inform particular corporations or governments, and even how to structure one's relationship with participants and their ability to provide input on the methods and presentation of research results.

More generally, the ethical and social consequences of the implementation of robots in society have to be taken into account. In most contemporary research projects that deal with smart homes or the deployment of robots in homes, care facilities, or public spaces, these aspects have to be investigated and addressed. Considering the ethical implications of digitalization and a potential hybrid human–robot society is a key societal issue that is now discussed at large, not solely by robot ethicists and philosophers.

10.10 Conclusion

HRI studies have a lot in common with work in several social science disciplines, including experimental psychology, anthropology, and sociology. It is good practice to be aware of scholarly norms and practices in the field or fields relevant to your work. HRI researchers are expected to be aware of and adopt the same rigor when collecting and reporting data as other scholars using the methods they have chosen.

[3] See www.apa.org/ethics/code/

[4] See https://s3.amazonaws.com/rdcms-aaa/files/production/public/FileDownloads/pdfs/issues/policy-advocacy/upload/ethicscode.pdf

[5] See www.acm.org/about-acm/code-of-ethics

HRI is also sensitive to the same problems that have plagued the social sciences for over a century. For example, in the drive to come up with original work, HRI experiments are almost never repeated. There is also a considerable publication bias, with positive results more likely to make it to publication, whereas negative results, less exciting results, or less conclusive findings tend not to get published or to go unnoticed. However, HRI has opportunities that were not on offer until recently. Experimental data, including large video logs, can now be fully stored and shared with others, ready for scrutiny or additional analyses. Methods, protocols, and results are now more available than ever before, largely due to the drive toward open-access publishing and preregistration of experiments.

Although there are new and exciting publishing options available, the HRI community is also exposed to the financial and social constraints of academic publishing. Although conferences offer a fast and predictable publication process, they do require considerable financial resources to travel to the event. Publishing in reputable open-access journals also comes with a considerable price tag. Flaky journals and conferences (Bartneck, 2021) offer much more affordable options without offering any advantage over just posting your article online yourself. Researchers often have no choice but to fall back to the traditional publishing channels, such as commercial journals, that do not charge the individual authors, but the libraries of their institutions.

The scientific publishing environment has changed and will continue to change. A big step forward is when large funding agencies require their funded projects to publish in open-access formats. In the meantime, researchers can choose to deposit their work in the institutional repositories of their institutions or on their private websites. This approach is often referred to as *green open access*, also known as *self-archiving*. It has been shown that making articles openly available increases their citations (Gargouri et al., 2010).

HRI researchers can also find relevant methodological approaches and discussions in the related field of HCI, which has a longer history of performing user studies, system evaluations, and theory building around the use of computing technologies in society and can provide guidelines and critical perspectives pertinent to HRI research. HRI researchers can learn from discussions about how to incorporate contextual variables into their work, how to think critically about design and study methods, and how to work more closely with the potential users of new robotic technologies through prior work in HCI. It is also, however, important to remember that HRI deals with robots, which are not only a different, embodied technology compared to computers but also pose different technical and social challenges for research.

Questions for you to think about:

- In some instances, it is not ethical or possible to answer a research question with an experiment. Can you think of such an instance? How would you address ethical issues related to the setup of your study?

How might you address concerns about the inclusion of vulnerable populations (e.g., children, older adults with cognitive impairments) in your study?

- "Significance" has been considered a misleading term because it says nothing about the relevance of a finding. Can you think of a situation where finding a significant small effect is relevant? What about a situation where it is irrelevant?
- Say you want to set up an experiment in which you assess how well a robot tutor teaches children. How would you set up your study? How would you measure the robot's ability as a tutor? What confounding factors do you expect?
- HRI studies often seek to address people's subjective experiences of robots—their enjoyment of the interaction, for example. How would you measure enjoyment, incorporating both direct and indirect and subjective and behavioral measures? How would you make sure that your enjoyment measure has construct validity—that it is actually measuring enjoyment with the robot, not just general happiness, or reflecting the participant trying to please the experimenter?
- How would you approach a user evaluation of your prototype differently from a systems evaluation? What types of questions would you want to answer in each type of evaluation? What kinds of measures would you use in each type of evaluation?

10.11 Exercises

The answers to these questions are available in the Appendix.

*** Exercise 10.1 Convenience sample** What is a convenience sample? Select one option from the following list:

1. A group of participants that you recruited in a convenience store
2. A group of participants recruited online through a crowdsourcing service
3. A group selected based on participants' easy accessibility or proximity to the researcher, such as university students
4. A statistical technique used to minimize bias in research studies.

*** Exercise 10.2 Types of studies** What type of study offers the possibility of establishing correlation or even causation? Select one option from the following list:

1. A qualitative study
2. A descriptive study
3. A cross-sectional study
4. A quantitative study

** **Exercise 10.3 What do participants see?** In which experimental design do participants see all experimental conditions? Select one option from the following list:

1. Within-subject design
2. Between-subject design
3. Longitudinal design

** **Exercise 10.4 Correlations and causation** Correlation and causation are important concepts in scientific study. Which statement is correct? Select one option from the following list:

1. Correlation causes causation.
2. They are the same thing—if variable A is correlated to variable B, then it also causes B.
3. Causation is a prerequisite for correlation.
4. Correlation is a necessary but insufficient criterion for causation.
5. *Causation* is a synonym for *correlation*.

** **Exercise 10.5 Variables** There are two types of variables in scientific studies. Select one or more options from the following list:

1. Independent variables are aspects that the experimenter manipulates.
2. Measurements are independent variables.
3. Dependent variables are aspects that the experimenter manipulates.
4. The experimenter manipulates measurements.
5. Dependent variables are aspects that the experimenter measures.

** **Exercise 10.6 Causal relationships** Only certain study types allow you to establish a causal relationship. Which studies allow you to establish a causal relationship? Select one or more options from the following list:

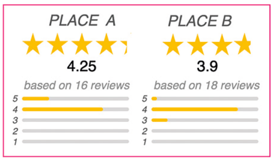
1. Observational studies
2. Ethnographic studies
3. Conversational analysis
4. Controlled studies
5. Case studies
6. System studies

*** **Exercise 10.7 Statistical inference** A researcher uses the significance level of $p \leq .05$ to test the relationship between robot likability and 40 other measured items. In reality, not one of these 40 items is related to robot likability. On average, how many significant results would you expect?

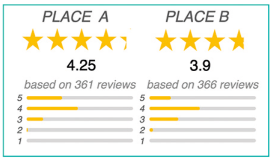
1. Zero
2. Five
3. Two
4. The question cannot be answered with the information given.

This material has been published by Cambridge University Press as Human-Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781009424233 (https://www.cambridge.org/9781009424233).

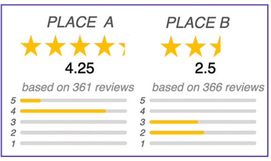
**** Exercise 10.8 Building blocks** Each of the following reviews indicates the importance of a different aspect of your data collection. Pair the images (a–c) with the names of the concepts.



a.



b.



c.

1. Tendency_____
2. Variability_____
3. Sample size_____

Future reading:

- Bethel, Cindy L., and Murphy, Robin R. Review of human studies methods in HRI and recommendations. *International Journal of Social Robotics*, 2(4):347–359, 2010. doi: 10.1007/s12369-010-0064-9. URL https://doi.org/10.1007/s12369-010-0064-9
- Field, Andy, and Hole, Graham. *How to Design and Report Experiments*. SAGE Publications, Thousand Oaks, CA, 2002. ISBN 978085702829. URL http://worldcat.org/title/how-to-design-and-report-experiments/oclc/961100072
- Hoffman, Guy, and Zhao, Xuan. A primer for conducting experiments in human–robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(1):1–31, 2020. doi: 10.1145/3412374. URL https://doi.org/10.1145/3412374
- Riek, Laurel D. Wizard of Oz studies in HRI: A systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, 1(1): 119–136, 2012. doi: 10.5898/JHRI.1.1.Riek. URL https://doi.org/10.5898/JHRI.1.1.Riek
- Baxter, Paul, Kennedy, James, Senft, Emmanuel, Lemaignan, Severin, and Belpaeme, Tony. From characterising three years of HRI to methodology and reporting recommendations. In *11th ACM/IEEE International Conference on Human-Robot Interaction*, pages 391–398. Institute of Electrical and Electronics Engineers, Piscataway, NJ, 2016. ISBN 978-1-4673-8370-7. doi: 10.1109/HRI.2016.7451777. URL https://doi.org/10.1109/HRI.2016.7451777
- Šabanović, Selma, Michalowski, Marek P., and Simmons, Reid. Robots in the wild: Observing human-robot social interaction outside the lab. In *9th IEEE International Workshop on Advanced Motion Control*, pages 596–601. Institute of Electrical and Electronics Engineers, Piscataway, NJ, 2006. ISBN 0-7803-9511-1. doi: 10.1109/AMC.2006.1631758. URL https://doi.org/10.1109/AMC.2006.1631758

- Young, James E., Sung, JaYoung, Voida, Amy, Sharlin, Ehud, Igarashi, Takeo, Christensen, Henrik I., and Grinter, Rebecca E. Evaluating human-robot interaction. *International Journal of Social Robotics*, 3 (1):53–67, 2011. doi: 10.1007/s12369-010-0081-8. URL https://doi.org/10.1007/s12369-010-0081-8