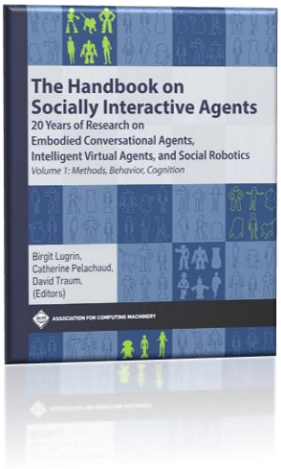
Empirical Methods in the Social Science for Researching Socially Interactive Agents

Astrid Rosenthal-von der Pütten and Anna M. H. Abrams



Author note:

This is a preprint. The final article is published in "The Handbook on Socially Interactive Agents" by ACM books.

Citation information:

Rosenthal-von der Pütten, A., and Abrams, A. (2021). Empirical Methods in the Social Science for Researching Socially Interactive Agents. In B. Lugrin, C. Pelachaud, D. Traum (Eds.), *Handbook on Socially Interactive Agents – 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics*, Volume 1: Methods, Behavior, Cognition (pp. 21-76). ACM.

DOI of the final chapter: 10.1145/3477322.3477325

DOI of volume 1 of the handbook: 10.1145/3477322

Correspondence concerning this chapter should be addressed to Astrid Rosenthal-von der Pütten (arvdp@humtec.rwth-aachen.de) and Anna Abrams (anna.abrams@humtec.rwth-aachen.de)

2 Empirical Methods in the Social Science for Researching Socially Interactive Agents

Astrid Rosenthal von der Pütten and Anna M. H. Abrams

2.1 Motivation

This introductory methods chapter is meant to be an informative overview for all non-social scientists who work with socially interactive agents (SIAs) and who would like to familiarize themselves with empirical methodologies in psychology and the social sciences. It is primarily written for young scholars, that is, undergraduate or graduate students who are new to this field of research and new to empirical methods in the social sciences. We will clarify the research process and explain methods for studying research questions surrounding human-centered development, testing and distribution of SIAs. In particular, we will provide answers to the following questions:

- What do we mean by methods in empirical social sciences? (Section 2.1.2)
- Why do I need methodological knowledge in empirical social sciences? (Section 2.1.1)
- Which research questions are addressed in empirical social sciences? (Sections 2.1.2 and 2.2)
- Which empirical methods should I use to address my research question? (Section 2.2)
- How does the chosen method work, in principle, and what aspects are important to consider when constructing, conducting, and analyzing my study and its results? (Section 2.2)
- Where can I find additional resources about methods in the empirical social sciences? (Section 2.3)
- What are the hot topics discussed in the community concerning methods? What are the current challenges and future directions? (Sections 2.4 and 2.5)

This chapter will also be useful for established scholars in the field as we provide an overview of different methods that can serve as inspiration. Furthermore, we included helpful

material such as lists of online-tools, questionnaires and specialized methods books and will point you into the right direction for further reading.

2.1.1 Why Do I Need Methodological Knowledge in the Empirical Social Sciences?

Depending on your discipline, you have a specific understanding of the term ”methods.” An engineer might understand methods as different systematic approaches that can be followed in order to reach the desired (technical) solution to a problem. The engineering method consists of stages such as idea, concept, planning, design, and then development of the former into a working prototype that demonstrates the solution to the problem [Ertas and Jones 1996]. The solution may be a tangible working prototype or an intangible working simulation. This prototype is being tested and debugged before launch. In computer science, depending on the problem statement, you might use theoretical, experimental or simulation computer science methods. For instance, the experimental computer science approach [Zelkowitz and Wallace 1998] serves to identify concepts that facilitate solutions to a problem and then evaluate these solutions. One example for this evaluation process would be simulation studies with which researchers can evaluate a technology by executing the product using a model of the real environment testing whether their hypothesis of the environment's reaction to the technology is supported. These are examples of methods that need no human in the loop (except for the engineer or computer scientist). In contrast to engineering and computer science, in psychology and the social sciences, the human being and its relation to other human beings is the central focus of the research endeavor. Psychology is a scientific attempt to understand and explain human mental processes and behavior. Psychological science includes fields such as perception, cognition, attention, emotion, intelligence, subjective experiences, motivation, brain functioning, and personality. In social psychology this extends to interaction between people, such as interpersonal relationships. The social sciences are concerned with the scientific study of human society and social relationships.

The term SIA already implies why you will need to gain at least some knowledge about social science concepts and methods. SIAs are meant to be “socially” interactive, drawing on social psychological principles of interaction. Moreover, SIAs are developed to be deployed in social settings (rather than caged robot arms in production lines). Thus, their development and deployment involves an additional problem space than the technical questions that we have discussed above. For this additional problem it will be useful to know about empirical methods in psychology and the social sciences.

Consider that you followed a systematic approach to develop a social robot that helps to gather supplies in a hospital and assists nurses. You have run simulations to test whether it moves correctly and whether speech input is processed as intended. You have bench-marked two different navigation systems and two different natural language processing units and identified which one performs better on your training data. Now, you are ready to give the social robot the go to interact with humans. Will the human, let's say his name is Ben, find

the robot useful? Is the interaction smooth? Does Ben understand the functionality of the robot? Does he like working with it? Does Ben consider the robot a team member? Does the social robot change the way how the human team members work with each other, and if yes, in what way? When you want to answer these questions, you need to know about the process of studying human perception, human behavior, and human attitude building. Ideally, engineers, computer scientists, and researchers in the field of psychology and social sciences work together in an interdisciplinary team from the start 'til the end of a development process following a human-centered design approach.

2.1.2 What Are Methods in the Empirical Social Sciences?

There are different methods for the acquisition of knowledge. We consider ourselves as social *scientists* and will therefore apply an empirical approach to acquiring knowledge instead of knowing because we have a "gut feeling," because it has always been like that, or because an authority said so. We will apply the *empirical method* that uses observation or direct sensory experience to obtain knowledge and uses evidence for verification of information [Gravetter and Forzano 2012, pp.13-15]. Within the empirical method, we follow either the *hypothetico-deductive model* of the scientific method and engage in an "approach to acquiring knowledge that involves formulating specific questions and then systematically finding answers." [Gravetter and Forzano 2012, p.16]. In contrast, there are also systematic methodologies based on empirical data but use *inductive reasoning*, for example, focusing on the construction of (new) theories through methodical gathering and analysis of data, such as grounded theory. This approach will only be briefly covered in this chapter (see Sections 2.2.3.1 and 2.4.2), but you will find recommendations for further reading in Section 2.3.

Once you have specified your research question or hypothesis, you have to think about your research strategy. In section 2.2 you will learn more about different research strategies. Most commonly, in the field of SIAs, researchers conduct evaluation studies. Evaluation is the process of developing and implementing a plan to assess something (e.g., your SIA) against the background of a specific research question or hypothesis using a systematic approach to assessment through previously defined measures (see Section 2.2.1). These measures can be quantitative and qualitative 2.4.2. Evaluations serve to determine the merit, worth, or value of something to inform judgements about the relative strengths and weaknesses, and the impact of variables. Since they are so prevalent in SIA research, we will put a focus on evaluation studies that can be realized differently, see Section 2.2.3.

2.2 Models and Approaches

How do you proceed once you have made up your mind that you want to do a study? In the following, we will guide you through the research process step by step. This section includes the research process in eight steps (see Section 2.2.1). Please note: the elaborations regarding the steps and important concepts and factors for each step are limited. In this book chapter,

we can only provide a glimpse into the broad topic of empirical social science methods. In addition, you will find recommendations for further reading throughout this section and in Section 2.3. We provide two scenarios to exemplify how researchers derive to a study design and which methodological choices they make considering the appropriateness of different methodological options. The following two scenarios are meant to give you concrete examples for methodological options when explaining the research process steps in Section 2.2.1; however, we also go through the full procedure of how to plan, conduct, analyze, and report a study using the two examples in Sections 2.2.2.1 and 2.2.2.2 to provide a more "hands-on" guide.

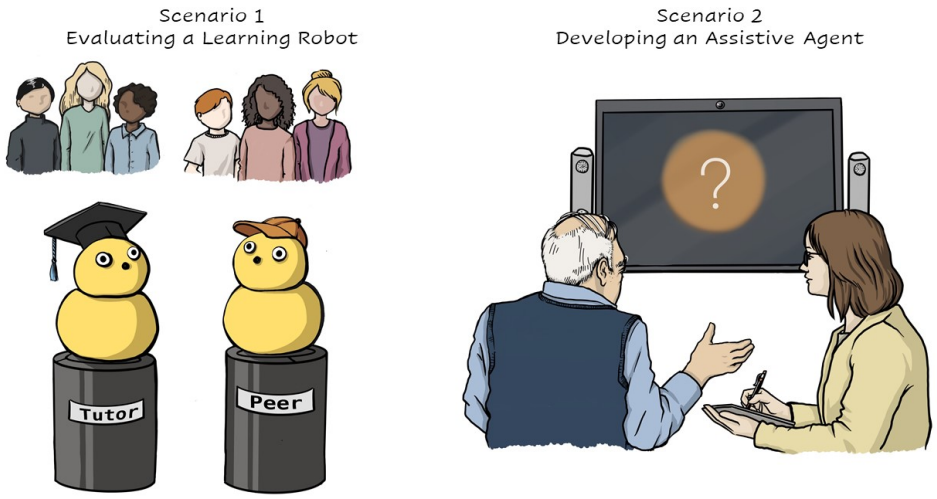


Figure 2.1 Example scenarios for study design.

Example 1—Evaluating a Learning Robot: Imagine there is a competition within your class on social robotics. Using the Keepon robot platform, the students in your class form two teams, each building a social robot to assist children with vocabulary learning for Spanish. The robots differ with regard to the social roles they take on in interaction (see Figure 2.1). The robot of the T-Team acts like a tutor, while the robot of the P-Team acts like a peer. You want to know which robot is better in helping children and which team has won the competition.

Example 2—Developing an Agent for Assisted Living: You are working at a research lab in a third party-funded project with the aim of developing a virtual assistant for older adults to be installed in their homes (see Figure 2.1). You are at the beginning of the project and want to know who exactly the target group is for this technology, what the virtual assistant should be

capable of, and how it should look like. At the end of the project there should be a prototype and an estimate of whether this might be a successful product on the market.

2.2.1 The Research Process

The textbooks on empirical methods agree on the nature of the research process as involving at least eight steps [e.g., Gravetter and Forzano 2012]:

- Step 1—Find a research topic
- Step 2—Form a research question or hypothesis
- Step 3—Define the research strategy and experiment design
- Step 4—Operationalization of variables
- Step 5—Define and select sample
- Step 6—Conduct the study / data collection
- Step 7—Data processing and data analysis
- Step 8—Report results

For each step, we will provide an overview about (i) what has to be considered in the step, (ii) which methodological decisions a researcher has to take, and (iii) what the methodological alternatives are for the respective decision.

2.2.1.1 Step 1—Find a Research Topic

The research process begins with identifying the research topic. In our examples, the research topic is given because the lecturer decided to run a competition, or the funding agency provided money to develop an assistant to bring to market. However, when you are about to do a Bachelor's, Master's, or PhD thesis, you will have to define your own research topic. You might want to identify a human need and develop a SIA addressing human needs. You might be inspired by one of the topics covered in this book and want to contribute to this area, or you might have observed a social phenomenon in interactions with SIAs that in your opinion deserves further investigation. All of these exemplary approaches are valid methods to identify and define a research topic.

2.2.1.2 Step 2—Form a Research Question or Hypothesis

Once you identify a research topic, you will have to review the literature in that field and find the specific research question(s) you want to address. If applicable (and in most cases it is applicable) you should consult theories that are relevant to your research topic. The literature review will help you to *define your central concepts* and get an overview of which research questions have already been addressed, what *empirical evidence* is available, and where the *research gaps* are. This allows you to formulate research questions or derive hypotheses that

are based on prior findings regarding that research question (see Section 2.2.2.1 on how to do this based on our examples).

2.2.1.3 Step 3—Define the Research Strategy and Experiment Design

There are many different ways to design a study. Your research strategy and study design depend on the type of research question or hypothesis you have proposed. Remember that a "research strategy is a general approach to research determined by the kind of question that the research study hopes to answer." [Gravetter and Forzano 2012, p.159].

> We will review different *types of research strategies* and explain when they are applicable: (i) the descriptive research strategy, (ii) the correlational research strategy, and (iii) experimental or quasi-experimental research strategies.
>
> Moreover, regarding the latter research strategy, experiments, we provide additional information on how to design experiments and explain three *types of experiment designs*: (i) the within-subjects design, (ii) the between-subjects design, and (iii) the factorial design.

Research strategies. The *descriptive research strategy* is "intended to answer questions about the current state of individual variables for a specific group of individuals" [Gravetter and Forzano 2012, p.160] and is not concerned with relationships between variables. For instance, you could assess how much money people would be willing to spend on a virtual assistant or which functions they would want to have incorporated into the system. If you're going to examine the relationship between variables, there are two different ways to do so.

One approach includes simple observation of variables of interest, as they exist naturally for a set of individuals. This is called *correlational research strategy*. If you are interested in the amount of money older adults are willing to spend depending on their income, you would choose a correlational strategy. You would assess people's willingness for investment and their income and run a statistical test on this data to discover a correlation. You would continue to examine whether there is any pattern of relationship between the variables and how strong this relationship is. This strategy can only *describe* a relationship but cannot *explain* the relationship because *correlation is not causation*.

Another approach to examining relationships follows an *experimental or quasi-experimental research strategy*. The *experimental research strategy* is "intended to answer cause-and-effect questions about the relationship between two variables" [Gravetter and Forzano 2012, p.163]. You can answer questions such as "does interacting with a robot peer lead to longer attention in a learning task compared to interacting with a robot tutor?" To answer cause-and-effect questions, you manipulate one variable (the *independent variable*) to create so-called treatment conditions (robot peer vs. robot tutor). In addition, you prepare for measurement of a second variable (the *dependent variable*) to obtain a set of scores within each treatment condition (attention span while learning). It is of great importance that all other potentially

influencing variables are controlled, as far possible. By *controlling all other variables*, you can conclude that differences in the scores of your dependent variable between treatment conditions is due to your manipulation of the independent variable. With regard to our example, you would design an experiment in which one group of children is interacting with the peer robot and one group of children is interacting with the tutor robot (two treatment conditions, independent variable: social role of robot) and you measure how long they focus their attention on the task (dependent variable: attention) and compare the scores between the two groups. It is crucial that you *randomly assign participants* to one of the groups. In the context of our example, this can easily be done by inviting children into your lab and randomly assigning them to treatment conditions.

Sometimes, however, it is not so easy to assign participants randomly to the experimental groups. Imagine, you are conducting the experiment in a school. For a whole week, you put the peer robot into one classroom and the tutor robot into another classroom. You cannot resolve the class structure for a week and randomly assign children, thus, you use two naturally existing groups: the classrooms. At the end of the experiment you are comparing learning gains in each class by administering a vocabulary test. This is called a *quasi-experimental research strategy*. Although, quasi-experimental settings use some of the rigor and control of true experiments, they are always flawed to a certain extend and cannot obtain an absolute cause-and-effect answer, because there might exist other group factors systematically influencing the outcome. For instance, the human teacher in one classroom might motivate children more to use the social robot for learning vocabulary, thereby generating more learning time and greater learning gain (thus, in this example, the possible *confounding variable* is: motivation by teacher).

Experiment design. There are, however, many more methodological decisions required when planning an experiment. For instance, you have to decide whether to use a within-subjects design, a between-subjects design, or a factorial design. As for the *within-subjects design*, you would use a single group of participants who receive or experience all of the treatment conditions. Thus, a within-subjects design looks for differences between treatment conditions within the same group of participants. In this case you would have one group of children who interact with both robots, the peer and the tutor. In contrast, the *between-subjects design* requires separate independent groups of participants for each condition. In this case you would use two groups of children. Each group interacts with only one version of the robot. Sometimes, researchers want to investigate more than one independent variable. This would require designing a *factorial study design*. In the context of the current example, you might be interested in the question of whether girls and boys react differently to peer or tutor robots, thereby introducing a second independent variable (in this case a so-called *quasi-independent variable* because you cannot actively manipulate the gender of your participants, but there are naturally existing groups). When two or more independent variables are combined in a single study, they are called factors. Our example would be a two-factor design in which both factors

have two values, resulting in a 2 x 2 factorial design with the factors gender (values: boy or girl) and the robot's social role (values: peer or tutor). You can design this study as a complete between-subjects design or as a so-called *mixed design* in which one factor is a between factor (gender) and one is a within factor (robot's social role).

When planning an experiment, please note that you might want to include a *control condition* (or a control group). A control condition refers to a non-treatment condition in an experiment where participants do not receive the treatment being evaluated. Here, a reference classroom group that does not interact with a robot but has normal class and is also measured in the dependent variable.

2.2.1.4 Step 4—Operationalization of Variables

The next important step in planning your study is the operationalization of your variables.

> In this step, we explain (i) what *operational definitions* are, (ii) why it is important to consider different *modalities of measurement*, and (iii) what *scales of measurement* exist.

In step 2 of the research process, the task was to identify theories relevant to your research question and to define appropriate constructs. The "problem" with *constructs* is that they are *hypothetical attributes* or mechanisms that help explain and predict behavior in a theory. Examples of constructs are motivation, knowledge, intelligence, or cognitive load. These constructs cannot be observed or measured directly, but it is possible to *observe and measure the external factors and the external behaviors* associated with the construct. Constructs can be influenced by external stimuli and in turn can influence external behavior. For instance, the theory of similarity attraction suggests that people are like others who they perceive as being similar to themselves, rather than dissimilar. Attraction is the relevant construct here. Attraction is hard to measure directly because it is a mental process. However, we can manipulate external factors such as similarity of the other person (e.g., similar = same gender/attitude/similar appearance; dissimilar = opposite gender/diverging attitude/diverging appearance). Moreover, we can observe and measure external behavior that might be affected by attraction such as a rating for how much we like that other person. What is needed is an *operational definition* that "specifies a measurement procedure (a set of operations) for measuring an external, observable behavior, and uses the resulting measurements as a definition and a measurement of the hypothetical construct" [Gravetter and Forzano 2012, p.105]. This process is also referred to as *operationalization*. In our example, the construct similarity can be operationally defined in a variety of ways. For instance, for our group of participants evaluating the assistive agent, we created an agent more similar (matching gender) or dissimilar to them (opposite gender). Hence, we are comparing two different levels of similarity that in this case is defined by whether or not the agent has the same gender.

A simple way to come to the operational definition for the variables of interest is to consult previous research that made use of the same variable, because this research should report in detail how the variables have been defined and measured. By adopting these definitions and measurements in your study your results will be directly comparable to the results obtained in previous research.

Usually, there are different options for measuring any particular construct and variable. For example, when you want to assess acquired knowledge in a specific language you could use self-report and ask people how much they think they have learned, you could administer language tests (e.g., vocabulary) or observe whether the verbal behavior in that language has changed and is more fluent, more verbose, and contains less grammatical errors than before a treatment. In this example we would use different *modalities of measurement*: *self-report measures* such as interviews and questionnaires, and *behavioral measures* such as performance tests or behavior in interactions. A third modality are *physiological measures* (e.g., galvanic skin response, heart rate, or brain imaging techniques). All three modalities have certain advantages and disadvantages that can influence the quality of the measurement. There are two criteria for the evaluation of quality of operationalizations of variables and these are *validity* and *reliability*. A valid measurement has been demonstrated to actually measure what it claims to be measuring and a reliable measurement is able to produce identical results when it is used repeatedly to measure the same individual under the same conditions (see Gravetter and Forzano [2012, pp.107-119]). If participants deliberately lie in a self-report this poses a threat to the validity of your measurement. In case you decide to use increased heart rate as a measure for similarity attraction you also might face a validity problem. Heart rate can increase due to a number of causes such as fear, anxiety, arousal, or embarrassment. The question is how can you be sure that measurement of heart rate is in fact a measurement for fear? To determine the validity and reliability of measures you should learn and read more about different types of validity and reliability in a methods book (see Section 2.3 for suggestions, e.g., some types of reliability can be tested for with statistical tests) and consult more closely the discussions in previous work using the variables you are using.

Once you have chosen the measures that you want to use in your study, you should be aware of the scale of measurement. Traditionally, there are four types of measurement scales: nominal scales, ordinal scales, and interval and ratio scales. *Nominal scales* represent qualitative (not quantitative) differences in the variable measured (some are female or male; being female is not superior nor inferior to being male). Categories on an *ordinal scale* are organized sequentially and consists of a series of ranks (e.g., first, second, third; small, medium, large). With an ordinal scale, you can determine not only differences but also the direction of differences (not the magnitude of differences). Interval and ratio scales are organized sequentially, and all categories have the same size [e.g., degrees in Celsius, each interval (degree) has the same size]. Hence, interval and ratio scales allow the determination of difference as well as its direction and magnitude. Interval scales have an arbitrary zero

point (e.g., Celsius or Fahrenheit have an arbitrary zero point in addition to positive and negative values) while ratio scales have a meaningful zero point. For ratio-scaled variables, zero is the complete absence of something. The scale of measurement of your variables also determines which statistical test you can use when describing your data and when trying to discover relationships between variables. In this regard, please note that so-called Likert scales (explanation can be found below in the examples) that are most frequently used in self-assessments are ordinal scaled but given the robustness of many parametric tests can be used as interval scales in statistical testing (see Norman [2010]).

2.2.1.5 Step 5—Define and Select Sample

Once you have established your study design and measures, you should invest some thought into defining and selecting your sample.

> In this step we explain (i) what is a *population*, a *target population*, and a *sample*; (ii) different *sampling procedures* and when to use them; and (iii) how to determine the adequate *sample size* for your study by using *power analysis*. We therefore briefly explain *statistical hypothesis testing*.

First, we have to distinguish between the *population*, being the large group of interest to a researcher, and the *sample*, the small set of individuals who participate in the study. Very often, you will have a so-called *target population* that is defined by the researcher's specific interests. By target population, researchers address a group of individuals in the target population that shares one specific characteristic. For instance, a target population could constitute all German children in fourth grade or all individuals over 70 years living alone in an independent home. Usually, researchers do not have the means to draw a sample from the whole target population (all children in second grade), but from an *accessible population* (e.g., all children in second grade in one city). However, the goal is always to *generalize study results* of the sample to the population. Therefore, researchers seek to find a *representative sample* that closely mirrors or resembles the population and its defined characteristics. When the sample does not closely resemble the population but has different characteristics from those of the population, this is called a *biased sample*. Researchers have to be careful which sampling procedures they use in order to avoid sampling bias.

The likelihood of the sample being representative or biased depends on the procedure that is used to select participants for your study. There are two types of *sampling procedures*: probability sampling methods and non-probability sampling methods. *Probability sampling* methods require that the odds of selecting a particular individual are known and can be calculated. In order to do so, you must (i) know the exact size of the population and all its members, (ii) each individual in the population must have a specified probability of selection, and (iii) selection of individuals must be a random process. For *non-probability sampling* methods, the odds of selection are not known, the researcher does not know the population

size and cannot list all members of the population. In this case, you do not use an unbiased method of selection. Thus, non-probability sampling methods have a greater risk of producing a biased sample. For the research field of SIA, not all population parameters are understood and can be identified. It is, thus, unlikely that you will be able to perform probability sampling methods. You will more likely perform non-probability sampling methods, such as convenience sampling. *Convenience sampling* means that you will be using those individuals who you have easy access to. Availability and individuals' willingness to participate are the decisive factors here. These are, for instance, students who are enrolled in one of your classes, or the children of the elementary school where you know teachers who are willing to help you in doing a study, or those people in the mall that happen to be there when you are conducting a field trial with your new social robot. Although convenience samples are obviously convenient, that is, less expensive and easier to get, they are also more prone to be biased. There are, however, ways to handle potential bias. You can ensure that your sample is reasonably representative and not strongly biased; for instance, you can work with schools from different districts of the city and be careful to select a broad cross-section of children (males and females, with siblings and only child, with and without immigration background). Moreover, you should describe your sample in detail in your research report and thus allow other researchers to evaluate how representative or biased your sample might have been and take this into consideration when evaluating the results of your study.

Once you know how you want to select your sample you have to determine the required *sample size* for your study—how large should the sample be in order to be representative? A general principle from statistics is the law of large numbers: the larger the sample size, the more representative the sample. There are, however, also practical limits to the sample size (e.g., time and expenses). Thus, most often you will have to compromise between the benefits and advantages of a large sample size and the costs of running a study with many participants. A *rule of thumb* is that you need about 25–30 individuals in every group you are testing [cf. Gravetter and Forzano 2012, p.142] because accuracy of the sample mean in relation to population mean increases with sample size, but the improvement of accuracy slows dramatically once the sample size is around 30 (per experimental condition!). Because of this limited added accuracy, researchers often opt for a sample size of 25–30 per condition.

The sample size is also determined by other statistical factors that can be taken into account in a so-called *power analysis*, which is a statistical procedure to determine the required sample size for detecting an effect of a given size with a given degree of confidence.

Power stands for the probability of finding an existing effect and is influenced by the significance level, the sample size, and the effect size (high power diminishes the risk of false negatives). Given any three of these four components, we can estimate the fourth. Hence, when we know the significance level (e.g., $p < 0.05$), the assumed effect size of the effect we are looking for (e.g., d = 0.5, which would constitute a medium sized effect in a *t*-test), and the power we want to use in our study (e.g., 80%), we can calculate the required sample size

for a *t*-test (e.g., 102 participants, 51 in each group). In a *t*-test, you determine the differences in means of two groups (children interacting with tutor or with peer robot). On the other hand, if you have a given maximal sample size (e.g., you only have access to 40 people with a very specific characteristic and no chance to get access to more individuals of that target population), the power analysis can determine the probability of detecting an effect of a given size with a given level of confidence. If you plan an experiment with two groups, trying to find an effect of medium size with 40 participants, the probability of determining this effect will be extremely low (power = 46%). This means that your study would have a 46% chance of finding a statistically significant effect of treatment condition given there really is an important difference between the treatment conditions. This might lead you to overthink and revise this experiment design. Statistics books often feature lists with examples of power analyses. There are also freely available software tools that help with performing power analyses (e.g., G*Power3; http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/).

2.2.1.6 Step 6—Conduct the Study/Data Collection

Before conducting a study, the last step should be to critically review everything that you have prepared and decided so far from an ethical viewpoint (see Chapter 3).

> In this step we explain (i) *research ethics*, (ii) *informed consent* and *debriefing* of participants, and (iii) provide *useful tips* for conducting a study.

Considering *research ethics* is very important and in many countries it is institutionalized with so-called *internal review boards* (IRBs) or *ethics committees*. A common process involves obligatory notifications to the IRB about every study involving human subjects. These reports should include detailed descriptions of the entire study, discussion of potential ethical concerns, and specification of measures to reduce potential harm to human subjects. The IRB commonly reviews study proposals and judges them upon ethical innocuousness. A positive evaluation of the ethics board is the official permission for conducting the study. IRBs often perform a risk–benefit analysis and assess the individual risks a participant is subjected to in a study and the benefits your research provides for society. One might tend to think that most research in SIA does not involve high risk for participants because people will not be physically harmed. Unfortunately, this is a fallacy because psychological harm can result from some studies. You might think that administering an IQ test in your study is a low risk endeavor. However, when a person participates in this test and receives a low IQ score, this can seriously threaten the person's self-concept. IRBs usually provide guidelines on how to conduct studies that follow ethical standards. IRBs are governed by Title 45 Code of Federal Regulations Part 46 of the United States. Many other countries have similar rules for the establishment and working processes of ethical review boards. If there is no official regulation of the state, then, very often, universities and academic associations have committed themselves to establish an ethical review board. Even when there is no institution requiring you to do an ethical review,

your research integrity should tell you to follow ethical guidelines and seek for guidance in this matter. Most academic journals and conferences will ask you to state whether your research has been IRB reviewed and might reject research that has not. Sometimes this can be avoided when you can explain in detail what measures you have undertaken in order to ensure ethical standards.

The American Psychological Association (APA) provides their *"Ethical Principles of Psychologists and Code of Conduct"* online for your reference (https://www.apa.org). The absolute basics for research are *informed consent* and *debriefing* (see section 8 in APA Ethics Code).

Informed consent means that you inform the participant that he or she is about to take part in a study and get permission to collect data [.20 2010]. This is especially important when you are collecting data that cannot be anonymized such as audio or video data. In this case, IRBs often require a process for data handling and data protection. The APA Ethics Code describes informed consent as follows: "psychologists inform participants about (1) the purpose of the research, expected duration and procedures; (2) their right to decline to participate and to withdraw from the research once participation has begun; (3) the foreseeable consequences of declining or withdrawing; (4) reasonably foreseeable factors that may be expected to influence their willingness to participate such as potential risks, discomfort or adverse effects; (5) any prospective research benefits; (6) limits of confidentiality; (7) incentives for participation; and (8) whom to contact for questions about the research and research participants' rights. They provide opportunity for the prospective participants to ask questions and receive answers." (APA Ethics Code, Section 8.02). Some universities or their IRBs provide examples or guidelines on how to construct an appropriate informed consent form.

Moreover, you should *debrief participants* properly, which means that "psychologists provide a prompt opportunity for participants to obtain appropriate information about the nature, results, and conclusions of the research, and they take reasonable steps to correct any misconceptions that participants may have of which the psychologists are aware. If scientific or humane values justify delaying or withholding this information, psychologists take reasonable measures to reduce the risk of harm. When psychologists become aware that research procedures have harmed a participant, they take reasonable steps to minimize the harm." (APA Ethics Code, Section 8.08). One specialty that frequently occurs in studies with SIAs is that researchers use a so-called Wizard-of-Oz (WoZ) scenario [Dahlbäck et al. 1993]. This means that participants ostensibly interact with an autonomous system, but actually the social robot or virtual agent is controlled by a so-called "wizard," a hidden confederate of the experimenter controlling the actions of the robot or virtual agent. In this setup, participants are deceived about the true nature of the SIA (for a review on WoZ in HRI, see Riek [2012]). Deception is to be avoided unless the researcher has determined that "the use of deceptive techniques is justified by the study's significant prospective scientific, educational, or applied value and that effective non-deceptive alternative procedures are not feasible" (APA Ethics

Code, Section 8.08). If your study setup includes any kind of deception, you are obliged to debrief participants as early as possible about the deception, preferably at the conclusion of their participation but no later than at the conclusion of the data collection, and permit participants to withdraw their data. For further discussion on deception in research, see, for instance, Christensen [1988] and Tai [2012].

When you have received the IRB approval, you can start recruiting participants, conducting the experiment, and collecting your data. Here are some *useful tips* that you usually do not find in a textbook but are based on experience. When recruiting participants, *always recruit more participants than you need*. There is always someone who does not show up or your technology is on strike on one day. You will experience that not all test trials produce suitable data to be included in your dataset. Thus, plan to recruit more participants than you need in order to cope for any dropouts. Clearly specify the *inclusion and exclusion criteria* for study participation. For instance, if it is crucial for your study that participants can properly hear the robot, an exclusion criterion will be impaired (and not corrected) hearing. If the study investigates how girls react to a robot, boys cannot participate. Clearly stating the inclusion and exclusion criteria in study advertising helps to avoid frustrating situations for you and your participants.

If you are conducting the experiment with more than one experimenter, try to be consistent in how the experimenters are conducting the experiment. Ideally, *counterbalance experimenters to experimental conditions* so that you do not have unexpected experimenter effects (see Section 2.2.3.1 for experimenter/interviewer effects). It helps when you prepare an *experimenter script* that lays out what should be said (and how) and what should not be said during the experiment when talking to the participant. For interaction studies in the lab, prepare an *experimenter checklist* that lists all the steps of your study. This prevents you from forgetting something and risking data loss, for example,

- which equipment has to be made ready, switched on, started before data collection?
- which documents have to be put out ready? (e.g., informed consent forms, written debriefing)
- how and where to store data after the interaction? (e.g., which server, folder)

An additional useful tip is to conduct *testing runs* with other uninformed lab members in order to see how long the study takes (this also helps in planning time slots for participants), whether everything runs smoothly, and whether participants do understand every task within the experiment. When you debrief participants, *ask them first whether they noticed something or found something strange*. This is especially important for WoZ settings in order to check whether the deception has been successful or whether it has been detected by the participant (in this case, these data cannot be included in the analysis). However, answers to this question can reveal other flaws in your study setup that can be changed when detected early. Last

but not least, keep records of all participants (in anonymized form) where you note relevant information such as technical errors, failed manipulations or deceptions (e.g., in WoZ), or other peculiarities. For more useful tips, consult the paper by Bethel and Murphy [2010].

2.2.1.7 Step 7—Data Processing and Data Analysis

After data collection, the most exciting step follows—data analysis.

> In this step, we explain (i) when you need to consider *processing of data* and (ii) how to make a plan for *statistical analysis* of your data.

First, you might have to process some of the data. For instance, you have to extract from a continuous video how often and how long children looked to the peer or tutor robot. In case of a study involving a questionnaire, you usually collapse data of a well-established questionnaire into a sum score or a mean value that is then used in further data analysis. Some processing is not very challenging or hard work. Other procedures require more effort and need quality checks. In the case of behavioral coding (e.g., for signs of attention by children in the interaction with a robot), coding is best done by two people. This procedure serves as a quality check because it enables you to detect the degree of agreement between the two coders to ensure that the coding results are valid (interrater/intercoder reliability, see Sections 2.2.3.3 and 2.2.3.1 for more detailed information).

After data processing, you can start with data analysis. In case you want to do a cause-and-effect analysis, you will have to consider your study design (correlational, between-subjects, within-subjects, or mixed-design), the measures for the independent and dependent variables, and their characteristics (e.g., nominal, ordinal, interval scaled), and then chose an appropriate statistical test. For detailed explanations on how to run statistical tests (e.g., the *t*-test that has been briefly discussed above), please refer to further reading: some books provide decision trees for choosing the appropriate test [Field 2018]. Some statistics book publishers also have companion websites with useful tools such as Andy Field's *Discovering Statistics* book series, which also features a "which stats test" online (http://methods.sagepub.com/which-stats-test). In any case, you will need to familiarize yourself with the most common statistical tests, how to run them, and how to interpret and report their results. We present some examples based on our two scenarios in Section 2.2.2.1.

2.2.1.8 Step 8—Report Results

The last step in your research endeavor is to report your results.

> In this step, we explain (i) the *structure of a research report* and (ii) general recommendations regarding *writing style*.

The form of the report depends on the addressee. Thus, the form of your report might depend on the funding agency of your research project, your lecturer, or the research community

via a scientific journal or conference publication. However, there are some general guidelines that should always be followed. A good research report should describe in detail the research process and the theoretical and methodological decisions that were made during this process. You should provide an objective description of the outcome of your research project, which typically includes the measurements that were taken and the statistical summary and interpretation of those measurements. And very importantly, your research study "grows out of an existing body of knowledge and adds to that body of knowledge. The research report should show the connection between the present study and past knowledge." [Gravetter and Forzano 2012, p.488]. The basic structure of a paper follows the IMRAD acronym:

- **I**ntroduction (Which question was asked?)
- **M**ethods (How was it studied?)
- **R**esults (What was found?), **A**nd
- **D**iscussion (What do the findings mean?)

Writing a good paper is a science in its own right. There are articles and even books [Hall 2012; Field and Hole 2013] that summarize guidelines and best practices and help with starting to write. They explain which information should be presented in which way in what chapter. When writing up psychological research, the APA publication manual is a very good reference [American Psychological Association 2020]. The APA also provides specialized guides for reporting quantitative [Cooper 2020] and qualitative research [Levitt 2020] in psychology. Most importantly, the *APA Publication Manual* tells you how to cite properly for psychology journals. However, writing a really excellent paper is a skill, and learning this skill will require experience and practice.

The general recommendation for writing style is to write in an impersonal and objective style and avoiding ambiguity, colloquialisms, and jargon. You should also try to avoid biased language (e.g., "older adults" is less biased than "the elderly"). Typically, research reports are written in past tense or past perfect when describing prior work (introduction and theoretical background sections of report), how you decided to set up and conduct the study (method section of report), and when presenting performed analyses and their results (results section of report). When interpreting and discussion the results you should switch to present tense. The work of other researchers must be properly cited in your research report to avoid plagiarism. Journals and academic conferences with proceedings usually have one predefined citation style that has to be followed when submitting your work. For psychology journals and conferences, this is the APA citation style. Proceedings of conferences in computer science are often realized by specialized publishers in that area such as IEEE, ACM, or Springer Nature, which all have their own citation styles. Find out which citation style is to be used before preparing the manuscript. Software for reference management and knowledge organization can be very useful to collect prior work and properly cite this work. Examples are EndNote,

Mendeley, Citavi, and Zotero (a comparison of reference management software can be found here: https://en.wikipedia.org/wiki/Comparison-of$_{r}eference_{m}anagement_{s}oftware$).

2.2.2 Two Exemplary Research Projects

2.2.2.1 Example 1—Evaluating a Learning Robot

For the scenario description, please refer to the introduction (see 2.2. In this scenario, your research topic is given (see Section 2.2.1.1): you shall assess the impact of different social roles that are implemented in a social robot that is supposed to help children with learning Spanish vocabulary. In order to formulate a research question or hypothesis (see Section 2.2.1.2), it is advisable to consult research from the field of education research, pedagogy, and social psychology when studying the impact of different roles in learning situations. For instance, how exactly do you conceptualize the role of a tutor and the role of a peer? How do peers and tutors behave differently in learning situations and what are the impacts of these different roles? You will find conceptualizations and prior evidence in existing literature (e.g., Belpaeme et al. [2018]). Belpaeme and colleagues [2018] review concepts and existing studies in this area and state that a peer has "the potential of being less intimidating than a tutor or teacher, peer-to-peer interactions can have significant advantages over tutor-to-student interactions" (page 6). For instance, in interactions with a peer robot, longer periods of attention on learning tasks, faster responses, and more accurate responses were observable compared to interactions with a tutor robot [Zaga et al. 2015]. There is certainly more evidence to find, but for our example, we will use this one prior finding in order to pose the hypothesis that the peer robot will elicit longer attention in the task of learning vocabulary and when asked for vocabulary the children's responses will be faster and more accurate. However, the (very limited) literature we reviewed in our example does not allow us to state a hypothesis on long-term learning gains. Thus, here we can only pose the following research question: what is the influence of the robot's social role on children's learning gain of Spanish vocabulary?

Next, you define the research strategy and experiment design (see Section 2.2.1.3). Your hypothesis suggests a relationship between the robot's social role and children's attention and recall promptness and correctness. Moreover, you assume that the social role might have an influence on long-term learning gain. Hence, your independent variable is the social role of the robot. The dependent variables are children's attention, recall promptness and recall correctness as well as learning gain. You plan to invite children into your lab and let them interact with one of your robots. Thus, you can randomly assign participants to your groups, making this a true experiment. Suppose you decide for independent groups of participants, hence you have a between-subjects design.

After you have identified your independent (social role of the robot) and dependent variables (children's attention, vocabulary recall promptness and correctness, and learning gain), it is time to operationalize these variables (see Section 2.2.1.4). For your independent variable, this means that you need to provide definitions of the two social roles as well as

definitions about the behavior that is connected to the two roles that can be implemented into the social robot. For instance, in Zaga et al. [2015] the differences in social role (peer vs. tutor) were established via the style of interaction through the design of gestures, speech, and postures with peer or tutor characteristics based on literature on teachers' multi-modal expressions and on peer collaboration. With regard to your dependent variables, this means likewise that you need a definition and a specified way to measure variables. In the work by Zaga and colleagues, focus of attention was measured by gaze behaviors of the participants directed at the robot and to the task (counts and duration of gaze behaviors: behavioral measure, ratio scale). You might also decide to include a self-report measure and ask children whether it was easy for them to be attentive during the task on a scale of 1 to 5 with 1 being "very hard" and 5 being "very easy" (self-report measure, ordinal/interval scale). This type of scale is known as *Likert scale* [cf. **?**, for an analysis and discussion on the usage of Likert scales in the field of human–robot interaction]. Moreover, you want to know how promptly and correctly children can recall the vocabulary during the interaction. Hence, you are measuring the time they take to recall an item (behavioral measure, ratio scale) and the number of correct items (behavioral measure, ratio scale). Lastly, you administer a vocabulary test 1 week after the treatment in order to assess learning gains in Spanish vocabulary.

In the next step, you need to define and select an adequate sample (see Section 2.2.1.5). Your target population are elementary school children in the fourth grade. You will probably identify an accessible population of elementary school children in a local elementary school where you have contact to—making this a convenience sample. In order to avoid sampling bias as much as possible, you could contact a second school in a different district. Your power analysis for a *t*-test given an alpha level of 0.05, power of 80%, and a medium effect size (d = 0.5) shows that 102 participants are required for your study.

In order to conduct the study (see Section 2.2.1.6), you need to prepare some documents. Especially, you will need the informed consent of the parents of the participating children because children cannot give informed consent themselves. Since you decided that you will run a laboratory study, children and their parents will probably only be available during the (late) afternoon hours or on weekends. You prepare an experimenter checklist that tells you to have study materials ready and to not forget to switch on the cameras because you need videos to determine the attention allocation of the children. You decide to debrief parents about the manipulation of the study immediately after their children's participation. However, you decide to postpone debriefing for the children until after data collection has been concluded because you fear that the children will talk about the manipulations to others while the study is still ongoing. After data collection and analysis, children can be debriefed and informed about the results together in their classroom. During the study, imagine that two children were so shy that they did not engage in the interaction at all. One interaction was interrupted by a phone call and a second interaction failed because the robot did not produce speech output

anymore. You note these four cases in your record list and plan to discuss with the team which cases have to be excluded from the dataset.

After data collection, processing of data and data analysis follows (see Section 2.2.1.7). In order to extract behavioral measures, you will need to process the video data. This means that you will have to do coding on the videos, that is, when and how long participants showed gaze behaviors toward the robot, how prompt their reactions were and how correct. You should prepare a coding guideline that describes what behavior can be interpreted as attention, and hand this to two coders who code the data. Afterward you calculate to what extent the two coders agree in their coding by assessing the inter-coder reliability. Depending on the measure, you need to calculate kappa statistics, correlation coefficients, or intraclass correlation coefficients [cf. Cohen 1960; Shrout and Fleiss 1979]. You determine the number of gazes and their duration; the latter is summed up. You also count the number of correct recalls during interaction and determine the mean reaction time to recall. Finally, you have a look at the free recall vocabulary test and get a test score for each child. Then you can perform independent *t*-tests on attention allocation, recall promptness and accuracy, and learning gain. You find that your hypothesis is supported—children pay greater attention to the peer robot, and direct recall is more often correct in this group than in the tutor group. However, recall promptness did not differ between groups nor did learning gain. Finally, you have to write a research report for the social robotics course (see Section 2.2.1.8).

2.2.2.2 Example 2—Developing an Agent for Assisted Living

For the scenario description, please refer to the introduction (see Section 2.2). In this scenario. you have the mission of building an assistive agent for older adults. You want to find out more about your target group in order to build a useful and pleasant intelligent virtual agent (IVA) that people will buy to use at home (see Section 2.2.1.1).

This example is a bit more complex than the Learning Robot Scenario. When you start to develop an IVA from scratch, your development process will actually involve a number of different research questions and thereby different studies (see Section 2.2.1.2). At first you want to find out more about your target group and their needs in order to determine the functionalities of the virtual assistant. Here the research questions might be (amongst others): What needs do older adults express to have and which, in their view, could be addressed by an IVA? What functionalities do they want to have in an IVA? How do they envision an IVA to behave and look like? And also important for the project would be the question: How much are customers willing to spend on a virtual assistant? Later on, when you start to prototype, you will probably perform a perception-only non-interactive study (see Section 2.2.3.2) showing participants pictures or short videos of different versions of the virtual agent with regard to looks and behavior. By this, you want to determine: Which agent is preferred by the majority of the participants and why? The next stage would then be conducting interaction studies (see Section 2.2.3.3), probably in a laboratory situation and later on in field experiments in future

users' homes. Let us assume that you are one of the first projects to ever try to develop a virtual assistant. In this case, there would be little prior evidence to build upon and a lot of your work will be exploratory. Thus, you would rather pose research questions than specific hypotheses. In order to keep it simple, we will outline two different studies in the following.

Study 1—Survey. Your first study will be a mix of descriptive and correlation research strategy (see Section 2.2.1.2). You would conduct a survey assessing the demographics of your assumed target group, including their income in order to examine the assumed relationship between people's willingness to invest in a robot and their income. Moreover, your survey could be asking people about their attitudes toward IVAs and which tasks they want the IVA to perform. In case you already envision some functionalities, you can ask people whether they would use the IVA for these functions. Using the survey method, you can query a large number of people. However, such a survey also has its limitations because people cannot explain why they have specific attitudes or why they would use a virtual assistant for one task but not for another task. A solution to this problem might be to back up your survey with another research approach. You could invite a smaller group of participants to interview them more deeply about this topic. During interviews, participants have the chance to elaborate on the "why." You could also invite a group of people to participate in a so-called focus group, which is a semi-structured interview held with a group of people. Please refer to Section 2.2.3.1 for more detailed information on different types of interviews.

As for operationalization of variables (see Section 2.2.1.4) in the survey, which only uses self-report measures, you could, for example, assess demographics such as gender (male, female, or divers; nominal scale), age (years; ratio scale), and income (euros or USD; ratio scale) as well as people's willingness to invest in a robot (euros or USD; ratio scale), their attitude toward virtual assistants, and which tasks they envision the agent for. Let us assume that you did not find a well-established questionnaire to establish attitudes toward virtual assistants. Hence, you decided to *ad-hocly* create some questionnaire items that you think are valid measures such as "I think virtual assistants are useful" or "I believe that using a virtual assistant improves my everyday life," and let people rate these questions on a scale of 1 to 5 with 1 being "I disagree" and 5 being "I agree" (self-report; ordinal/interval scale). You also provide participants a list of functionalities and ask whether or not the virtual assistant should be capable of those functionalities (e.g., "virtual agent has access to calendar": yes/no; nominal scale; self-report).

In the next step, you need to define and select an adequate sample (see Section 2.2.1.5). Your target population for the survey study are older adults aged over 70 who live in their own homes or in assisted living environments (in contrast to nursing homes). Let us assume that you work together with four cities and have access to the population register of these cities. Thereby, you know the exact size of the target population (in these four cities) and all its members. By this you can assign each individual in the population a specified probability of selection, and randomly select the number of individuals you need. Based on a power

analysis, you know you will need at least 134 individuals for performing the correlation analysis between income and willingness to invest. However, it might be advisable to recruit more individuals following the law of large numbers.

When planning to conduct the survey (see Section 2.2.1.6), you will need to include the informed consent into the online survey on the first page (see Perrault and Keating [2018] for recommendations on how to construct informed consent in online studies). You have sent 1000 potential participants a written invitation with the link to the online study because you expect that only a fraction of invited persons will actively make the effort to participate. The return rate for surveys are sometimes as low as 10% to 20%. Another recruitment tool many researchers are using nowadays are platforms such as Amazon Mechanical Turk (MTurk) or CrowdFlower, which have the advantage of fast completion of the study, access to otherwise hard to reach target populations, and more diversity in samples, such as specific occupational groups or people with a specific health condition [cf. Casler et al. 2013; Smith et al. 2015; Hauser and Schwarz 2016], but also come with disadvantages [cf. Fleischer et al. 2015; Smith et al. 2016]. The usage of crowdsourcing websites has been discussed in different disciplines [cf. Necka et al. 2016; Shank 2016; Follmer et al. 2017].

The survey study will predominantly be analyzed (see Section 2.2.1.7) on a descriptive level stating the percentage of male and female respondents, their mean age, and their mean income (which can also be presented per groups using cross-tables). You can run a correlation analysis on the relation between income and willingness to invest in an IVA, which shows that there is no relationship. Rather, the descriptive data suggests that there is low variability in what people are willing to invest regardless of their income. Moreover, the descriptive data gives you an impression on which tasks the participants judge as suitable for an IVA and which are not.

The results of the survey are part of a research report to be delivered to the funding agency. (see Section 2.2.1.8)

Study 2—Perception-only non-interactive study. Your second study will probably be a study evaluating the participants' perception of different IVAs (see Section 2.2.3.2), in which you compare different looks of the IVAs. Let us assume that you designed two versions of a virtual assistant: a female version and a male version. Based on work in social psychology on similarity attraction [Montoya et al. 2008], you develop the hypothesis that an agent matching the participants' gender might be preferred over non-matching agent (see Section 2.2.1.2). In summary, you are using a 2 x 2 mixed factorial design with the quasi-independent between-subjects factor participant gender (male, female) and the within-subjects factor agent gender (male, female).

As for operationalization of variables (see Section 2.2.1.3), you use participant gender (male, female; nominal scale) and the virtual assistant gender (male, female; nominal scale) as independent variables. As dependent variables, you want to assess the perceived likability of the agent, its perceived similarity to the participant, and participants' usage intentions. It is

advisable to include a so-called manipulation check for similarity. A *manipulation check* "is an additional measure to assess how the participants perceived and interpreted the manipulation and/or assess the direct effect of the manipulation" [Gravetter and Forzano 2012, p.268]. In your case, this could be an item asking "How similar is this agent to you?" You could again use a Likert scale ranging from 1 "not at all similar" to 5 "very similar" (self-report, ordinal scale). Your manipulation of similarity would be successful when participants rate an agent matching their gender significantly more similar than participants evaluating an agent not matching their gender. As dependent variables you use the perceived likability of the agent and ask participants "How likable is the agent?" on a scale ranging from 1 "not at all likable" to 5 "very likable" (self-report, ordinal/interval scale) as well as their usage intention and ask "How likely are you to use this agent" on a scale ranging from 1 "will definitely not use" to 5 "will definitely use" (self-report, ordinal/interval scale).

In the next step, you need to define and select an adequate sample (see Section 2.2.1.5). The power analysis for your 2 x 2 mixed factorial design tells you that you will need at least 34 individuals, that is, 17 female and 17 male participants since this is your between-subjects factor. This is probably going to be a convenience sample. For instance, you could launch an advertisement in a local newspaper stating that you are looking for study participants.

You prepare your perception-only non-interactive study in the laboratory (see Section 2.2.1.6). When advertising the study (which includes older adults), you explicitly state that participants should have normal or corrected to normal vision and hearing. Participants give informed consent before starting the experiment and are debriefed after completion of the study.

When analyzing the data (see Section 2.2.1.7), you first do your manipulation check by calculating a mixed design repeated measures analysis of variance (ANOVA) with the between-subjects factor participant gender and the within-subjects factor agent gender on the dependent variable "perceived similarity." The analysis reveals an interaction effect showing that indeed a gender-matched agent is perceived as being significantly more similar to the participants than a non-matched agent, that is, women rate the female agent as more similar than the male agent and men vice versa rate the male agent more similar to themselves. However, when performing the same analysis for "likability of the agent" you see that the similarity attraction effect is only observable in men. Men rate the gender-matched male agent as significantly more likable than the female agent, while women did not show a preference for one or the other agent. Moreover, there was no difference in usage intentions. Given your results, it would be advisable to either continue developing both agents or, if this is too costly, to continue developing the male agent only since women did not show a preference for agent gender and men preferred a male agent.

You plan to submit your research result to a human–technology interaction journal since those journals accept interdisciplinary works (see Section 2.2.1.7). You inform yourself about the journals guidelines for authors, the required template, and the citation format.

2.2.3 Types of Studies Most Commonly Used in SIA Research

As we learned above, the type of study you are conducting depends on your research question and the research strategy you choose in order to answer it. In the example of developing an assistive agent for older adults, we saw that in the early stages of the development it might be advisable to conduct *interviews* in order to receive more in-depth information about people's needs and wishes for what an agent should be capable of. Later on, the team developed prototypes of the virtual agent such as different graphical renderings of the virtual agents in still pictures or maybe animations of behavior in short videos. This can be presented to participants in *perception-only non-interactive studies* to receive feedback that can be used in further iterations of the development process. In the later stages of the development process you will have an autonomous or semi-autonomous agent that can be used in *interaction studies*. These three types of studies, interview, perception-only non-interactive study, and interaction study, are most frequently used in SIA research. We will discuss their advantages and disadvantages.

2.2.3.1 Interview

An interview usually follows an exploratory research agenda and is commonly used in ethnographic or field studies [Qu and Dumay 2011]. Mostly, researchers have broader research questions rather than narrow hypotheses that are addressed in interview studies. Even though interviews are sometimes used to enrich data from quantitative studies, they are more commonly used to get a first but deep understanding of how individuals experience, perceive, think or feel about, and evaluate a certain topic. This leads us to a very fundamental fact about interviews: they always produce data that rely on subjects' *introspection*. Introspection entails a lot of subjectivity, which researchers usually try to avoid when doing social research as subjectivity introduces biases. With interviews, this is not the case. In interviews, gathering *subjective* and introspective data is wanted. In interviews, you do not aim to find quantifiable and generalizable phenomena. Interviews deliver *qualitative data*, which stands in contrast to most other methods introduced in this chapter. *Interviews are described as centered around the interviewee, qualitative, descriptive, presuppositionless, focused, open for ambiguity and changes, taking place in an interpersonal interaction* [Kvale 1983].

Whether interviews are a valid method for your research depend on the questions that you are asking. The kind of questions that you can address by applying a qualitative research methodology such as an interview technique are questions concerning the "why." You will never be able to identify every possible answer that people might give to the why-question (e.g., why is this conversational agent appealing to a specific user?), so you will not be able to hand people a questionnaire with multiple-choice items covering all possible answer options. In order to really understand why people feel how they feel or think how they think, you cannot give them predefined answers to choose from but have to let them talk freely. Interviews give you a very deep insight into real needs and demands (careful: you are not producing

statistically significant results generalizable to all other older adults). Thus, interviews are often used in the analysis of needs and requirements for technical systems. Some researchers use interviews as part of a *participatory design* method along with other qualitative methods such as focus groups, co-creation workshops, and paper prototyping (e.g. Frauenberger et al. [2011], Šabanović et al. [2015], and Lee et al. [2017]).

Types of Interviews. There are many different types of interview techniques and for a complete overview, please refer to one of the recommended textbooks in Section 2.3. Here, we will give you an introduction into three different types of interviews concerning their degree of structure [Qu and Dumay 2011]:

- *Structured* The interviewer asks the interviewee predefined questions according to a rigid interview guide that is followed strictly throughout the interview. The guide does not only give the exact wording for questions, but also for the introduction into the topic and any other information that is given to the interviewee. The underlying assumption is that correctly and unambiguously asked questions will all deliver relevant information. On the one hand, compared with the other structured types, the structured interview produces the most comparable and generalizable data. On the other hand, there is no room for flexibility and spontaneity. Thus, we argue that the structured interview is a bad compromise from both worlds, qualitative and quantitative research, and would like to advise you to use a less structured approach.
- *Unstructured* The unstructured interview is the least formal and predefined type of interview. It is rooted in ethnographic research where an interviewer tries to understand someone's perspective entirely and most data is gathered through conversation rather than a pre-prepared line of questions. In an unstructured interview, the interviewer adapts and reacts dynamically to the topics brought up by the interviewee. The underlying assumption is that researchers cannot know the relevant questions in advance. The advantage of this approach is its openness to any given topic that an interviewee addresses. However, by being an active part of the conversation, the interviewer risks getting involved and shaping and steering the conversation too much. Additionally, data are not easily comparable between interviewees.
- *Semi-structured* The semi-structured interview technique lies within both extreme ends. Methodologically, it is not pinpointed to the exact middle but is interpreted differently (sometimes more structured, sometimes more unstructured) by different researchers and research fields. Semi-structured interviews make use of systematic, standardized interview guides but allow for interim questions and unplanned exploration of certain topics. Thus, the interviewer aims to ask certain pre-planned questions in the same standardized way to all interviewees to create comparable answers but gathers individual responses from specific interim questions. Successful execution of a semi-structured

interview requires a well-trained interviewer to make sure the interview situation stays under the interviewer's control even though there are free talking passages.

These types of interviews can be applied to many different scenarios. Some interviews are conducted as part of *contextual inquiries*, which includes placing the interviewee in a relevant context. Instead of inviting older adults into the lab to do an interview on assistive devices, you might as well visit them in their homes or care facilities and ask them questions while they are in their usual habitat. Verbal data are usually *combined and analyzed in the context of behavioral observations*.

Advantages and Disadvantages.

- *Time:* Qualitative research is not about measuring and numbers. Having a small sample is likely to decrease the time for data gathering and its analysis. (However, sometimes, conducting 10 interviews might result in more work than a quantitative poll with 100 participants.)
- *Information density and quality:* Well-conducted interviews will deliver more in-depth information about the participants. Participants will be able to verbally communicate their answers that will deliver more extensive answers than anonymous participants typing answers on their keyboards in front of a computer. There is an added value and potential to combine verbal information with behavioral data analysis in face-to-face interviews.
- *Trust:* During an interview situation, a good interviewer will build a personal bond with the interviewee and cause the interviewee to trust and share many information (refer to the biases below to read about the drawbacks of this situation). The interviewer will also never be able to validate whether an anonymous participant in an online survey was actually part of the target group or only in it for the incentive.

Even though, potentially, there are many more biases in interview studies, a selection of four biases is presented and described in Figure 2.2.

Practical Tips: How to Conduct Interviews. We would like to give you some tips from our personal experience in conducting interviews that you might find practical.

1. *Educate yourself:* Try to learn as much as you can about the topic in question in order to formulate relevant questions and prepare your interview guide. Learn as much as you can about the target group in order to pose questions that are sensible and appropriate, and study possible biases and pitfalls in interviews in order to prevent them. Enroll in an interviewer training course.
2. *Use this simple principle:* You can only get answers to questions that you have asked. Thus, careful preparation is as important as in any other research study. What do you want to know? What do you have to ask? If you have very specific questions, you should

Biases in Interviews	Description
Experimenter Bias	There are many biases that derive from the participant or the experimental situation, but there are also biases caused by the researchers themselves. The experimenter bias entails many effects on the participant and the experiment that the researcher introduces to an experiment that are not wanted, e. g. effects caused by the experimenter's gender, age, outer appearance or use of words (read more on these biases and how to overcome them in interviews, e. g. in Chenail, 2011).
Suggestion	An interviewer has to be very well-trained to not use any kind of suggestion, verbally or non-verbally. Questions such as "Did you like that?" (ask *how did you like that* or even better *what is your opinion*) infer positivity and are among the more obvious forms of suggestion. Any behavioural gestures such as nodding, smiling and hand gestures can signal a wrong impression to the interviewee.
The Good Participant (or Good Subject Effect)	Interviews build on trust, atmosphere and an interpersonal connection between interviewee and interviewer. This can cause biased results. Participants might want to deliver a good story or be helpful. They might want to give answers that they expect the interviewer to be wanting to hear. This effect is also present in quantitative laboratory studies (Nichols & Maner, 2010).
Introspection	People cannot always know about the reasons for how they feel, think and behave. Much research is concerned with "the introspection illusion" that doubts the validity of introspective information. Individuals overvalue their capability to introspect constantly also in daily life (e. g. Pronin, 2009).

Figure 2.2 Selection of biases that potentially influence results in an interview

choose a more structured approach. If you do not have specific questions but want to have a first glimpse into a topic, chose a more unstructured approach.

3. *Embrace pauses:* In normal conversations we tend to fill pauses because silence is sometimes considered awkward. As an interviewer, you should not fill these pauses with paraphrasing and repetitive questioning. Be patient. Let interviewees fill pauses and be surprised how many extra information you get (also, spontaneous paraphrasing should be avoided because it influences participants in an unsystematic way and produces biased answers).

Data Analysis. Analysis of data depends heavily on the type of interview. Unstructured interviews do not produce very comparable datasets and should not be interpreted as such. Data from such interviews is usually described per participant and summarized without counting or measuring data. The more structured the interview is, the more comparison between participants can be integrated in the analysis of data, and the more descriptive statistical analyses can be used to analyze data. *Audio and video recordings* are usually used for analysis. Audio recordings are transcribed. Coding schemes can be used to cluster answers into categories, for example, positive, negative, and neutral for general valence of an answer. For coding, two coders can categorize, and inter-coder reliability can be calculated (e.g., Cohen's kappa, see Section 2.2.1.7).

2.2.3.2 Perception-only Non-interactive Studies

In perception-only non-interactive studies (or short perception studies), participants are presented with stimulus material such as pictures, videos, audio files, or written descriptions of socials robots or virtual agents that shall be evaluated, for example with regard to design, appearance, or behavior (i.e., participants view stimulus material *without directly interacting* with an SIA). The goal is to develop a detailed understanding about how people interpret and reason about a robot's appearance and behavior. Such an understanding is not only crucial for advancing our basic knowledge about human–robot interaction but also for our ability to design a robot's appearance and behavior that is easy to recognize and to interpret. Sometimes perception-only studies are combined with in-depth interviews (see Section 2.2.3.1). In this case, participants are presented with stimuli of SIAs and are asked to give short ratings and subsequently elaborate on why they have given such ratings [Rosenthal-von der Pütten and Krämer 2015].

Types of Perception Studies. Many perception studies in development processes are used for evaluation of different designs to identify the best design option for the SIA. Other perception studies rather explore psychological phenomena by using controlled stimuli such as videos or pictures. This has the advantage of control as stimuli can be controlled whereas in interaction studies the interaction unfolds between the interactants (the SIA and human), giving the researcher less control about what exactly happens.

Advantages and Disadvantages. In contrast to interaction studies, the advantage of perception studies is that they are less error-prone since the stimulus material consists of pictures, descriptions, or videos of SIAs. In interaction studies dropouts happen regularly due to malfunctions of the social robot or virtual agent. This risk is diminished in perception studies. Moreover, perception studies require less personnel than interaction studies where often more than one experimenter is needed to run the study. In fact, perception-only studies can often be performed using online survey platforms that can be completed by a large number of participants simultaneously. Crowdsourcing platforms such as MTurk or CrowdFlower facilitate the recruitment of participants (see Section 2.2.1.6). Lastly, in perception studies researchers can exert a higher control on the experimental setting because the stimuli are exactly the same for all participants; whereas in interaction studies the interaction unfolds between the two interactants, thus always generating variability in the flow of interaction. On the downside, perception studies lack external validity because the stimuli are not presented in context, that is, the context of a real interaction situation. For some research questions this is more problematic than for others. For instance, in a perception study the nonverbal behavior reviewed in small videos might be clearly recognized as dominant or submissive. However, this effect might be diminished when the dominant behavior sequences are presented in a longer interaction phase together with other nonverbal behaviors.

Practical tips: How to Conduct Perception Studies. We would like to give you some tips on what to consider when designing perception-only non-interaction studies.

1. *Experiment design of recognition studies:* You have to make important methodological decisions when designing your recognition study, for example, whether each participant is shown only one type of behavior (between-subject design) or whether each participant is shown multiple types of behavior (within-subject design). A within-subjects design study may cause bias in that participants are prone to engage in more direct comparison between the various stimuli. In case this establishes a confound with regard to your research question a between-subjects design would be more advisable. Furthermore, you have to decide which response format to use in assessing people's ability to recognize a certain behavior, for instance, a forced-choice or Likert-scale response format. Previous research has shown that such methodological decisions about the study design and response format have large implications for the conclusions we draw, for example, people's ability to reliably distinguish between emotion expressions is highly contingent on the particular response format a study employed [Russell 1994]. These considerations are valid for all types of studies; recognition studies are especially prone to generating distortions of recognition rates based on methodological choices.
2. *Framing of the study:* Make sure that participants know what their task is and which perspective they should take when making evaluations, especially when you are using an online study. For some research questions you want the participant to act as an observer of a situation; for other research questions you want the participants to put themselves in the shoes of a person in the portrayed situation.

Data and Data Analysis. Unless you are combining stimulus presentation and rating with an interview (see Section 2.2.3.1 for data analysis tips) that would require a mixed methods approach to data analysis (see Section 2.4.2), most perception studies will result in self-report data that usually need less preprocessing compared to interaction studies with video coding or transcription of interactions. The descriptive and inferential statistical analysis follows according to the previously defined hypotheses (see Section 2.2.1.7).

2.2.3.3 Interaction Studies

In addition to the data from perception and evaluation of agents from observation of stimulus material, meaningful data can be gathered from interaction studies. In interaction studies, participants are invited to directly interact with an SIA. Thus, an interaction study always includes at least one participant and at least one SIA joined up in some form. Information is drawn from verbal and behavioral observations made during the interaction and usually added by pre- and/or post-interaction questionnaires where people report, for instance, prior experiences with SIAs, attitudes toward SIAs, and their perception and evaluation of the interaction itself. Studies involving interaction can either be of qualitative nature where, for

example, an interaction is followed by an interview or where an interaction is part of a single case study. An interaction study can also be part of an experiment where the kind of interaction itself is varied and serves as the independent variable or different kinds of participants are confronted with the same kind of interaction (see Section 2.2.1).

The major advantage of a real interaction is the *external validity* of the results. Only in a real interaction scenario can data be gathered either in experiments or qualitative studies that can be transferred and interpreted for real-life interactions. In addition to data from the interaction itself, participants can be asked to rate and evaluate the SIA and their real interaction experience. These data paired with observational data will deliver a dataset that can give a very holistic understanding of human–SIA interactions. Mixed methods are powerful tools in understanding phenomena.

Type of Interaction Studies: Interaction Settings and Methods. Interaction studies can be conducted in different settings. Two of the most prominent settings to conduct a study are lab vs. field. Lab studies take place in research institutes and usually involve highly controlled conditions under which the study is conducted. Field experiments are conducted "in the wild," in natural settings. Commonly it is assumed that the lab provides higher internal validity (more control of the independent variable) and field studies produce higher external validity (higher generalizability due to natural surrounding) [Reis and Judd 2013]. Even though, in empirical science, it is never that easy (e.g., you can have unexpected influences in the lab that endanger the internal validity of the experiment), we can state that *different degrees of internal and external validity* have to be considered when choosing the setting of an experiment and that the setting enables you to exert more or less control over the situation. Many interaction studies with socially interactive robots take place in semi-public spaces such as hospitals, airports, and shopping malls. These places have one advantage over public spaces—people are usually prepared and warned about camera surveillance upon entrance in these areas. Even as a researcher with good intentions, you may not film and record individuals without asking for consent. Thus, careful planning and choosing of a data collection sight does not only include concerns about the validity of the experimental result but requires *considerations about ethics, privacy rights, and data security* (see Section 2.2.1.6).

Types of Interaction Studies: Types of Interactions. An interaction can either be virtual or with a physically embodied agent. On many occasions, testing and studies are conducted when the agent is still being developed or not at all developed. In these cases, studies in virtual reality can be conducted to gain a first impression of how an interaction might take place. Another option to study not yet developed autonomous systems are studies with a *WoZ design*. A WoZ design involves a person who remotely operates the agent. There are different types of SIA capabilities that are often simulated using a WoZ design, among others: natural language processing, navigation and mobility, and nonverbal behavior [Riek 2012]. A structured guide and training for the wizard controlling the agent are necessary to ensure

reliability and consistency. Only reliable testing conditions will deliver data that can be used for analysis.

Advantages and Disadvantages. In contrast to other types of studies, interaction studies' main advantage is the possibility of observing a real interaction and combining different types of data together. However, the amount of time and preparation that has to be put in to set up an interaction study sometimes exceeds other types of studies. In addition to the usual preparation for each step of an experiment (see Section 2.2.1), research teams have to develop and test the functioning of the interactive agent and make sure the system runs reliably throughout the experiment. The extent of preparation time and resources that are needed are dependent on the agent and the type of interaction.

Practical Tips: How to Conduct Interaction Studies. Some practical tips for running an experiment are discussed here. Also consult Section 2.2.1.6 for more tips as well as the paper by Bethel and Murphy [2010].

1. *Experimenter script:* prepare a script for the experimenter and confederate (and wizard) that defines what has to be done and said in which way in order to avoid possible confounds to your experiment.
2. *Experimenter checklist:* the checklist should include all the steps of your study so that you do not forget to switch on a device and risk data loss.
3. *Testing runs*: invite other lab members to be pretend-participants in your study to search for possible misunderstandings and misconceptions before you conduct the study with real participants.

Data and Data Analysis. Analysis of data is preceded by preparing your data. Depending on the type of study, preparation includes transcribing video and audio recordings of the interaction, coding and categorizing the transcripts, and transferring the processed data together with data from any additional material (e.g., questionnaires) into your statistics software. Descriptive and inference statistical analyses follow according to the previously defined hypotheses (see Section 2.2.1.7).

2.3 Research Tools

There are many (free) resources available that will help you with constructing and conducting your study, analyzing your data, and reporting your results. In this section we provide you with lists of useful online resources, recommendations for further reading on quantitative and qualitative research methods, statistics, and reporting your scientific work as well as survey and experiment tools. Moreover, we provide you with an overview of ready-to-use questionnaires that could be helpful for your study of SIAs.

Useful Online Resources and Online Research Tools:

- Power analysis: G*Power3; http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/

- Statistical test decision tree: http://methods.sagepub.com/which-stats-test
- The Ethics Code of the American Psychological Association: https://www.apa.org/ethics/code/
- www.surveymonkey.com (commercially available online survey tool in 17 languages)
- www.qualtrics.com (commercially available online survey tool in 62 languages)
- www.soscisurvey.de (free online survey tool, user interface only in German)
- https://www.psytoolkit.org/

Books on Quantitative and Qualitative Research Methodology, Statistics, and Reporting:

- F. J. Gravetter, L. A. Forzano. 2012. *Research Methods for the Behavioral Sciences*. (4. Aufl.). Wadsworth, Cengage Learning, Belmont, CA.
- A. Field and G. Hole. 2013. *How to Design and Report Experiments*. Repr. SAGE, Los Angeles.
- A. Field. 2018. *Discovering Statistics Using IBM SPSS Statistics*. (5th. ed.). SAGE. (also available for R and SAS), Los Angeles, London, New Delhi, Singapore, Washington, DC, Melbourne.
- C. Jost, B. Le Pévédic, T. Belpaeme, C. Bethel, D. Chrysostomou, N. Crook, M. Grandgeorge, and N. Mirnig. 2020. *Human–Robot Interaction. Evaluation Methods and Their Standardization*. Springer International Publishing (12), Cham, IL.
- E. Lyons, A. Coyle. 2016. *Analysing Qualitative Data in Psychology*. (2nd. ed.). SAGE, Los Angeles.
- P. Leavy (Ed.). 2015. *The Oxford Handbook of Qualitative Research*. Oxford Library of Psychology. Oxford University Press, Oxford.
- George M. Hall. 2012. *How to Write a Paper*. John Wiley Sons, Ltd, Chichester, UK.
- American Psychological Association. 2020. Publication Manual of the American Psychological Association. *The Official Guide to APA Style*. 2020. (7th. ed.).
- Harris M. Cooper. 2020. *Reporting Quantitative Research in Psychology. How to Meet APA Style Journal Article Reporting Standards*. (2nd. ed., revised).
- H. M. Levitt. 2020. *Reporting Qualitative Research in Psychology. How to Meet APA Style Journal Article Reporting Standards*. (Revised ed.).

Questionnaires Commonly Used in SIA Research. Although we look back to two decades of research on SIAs, there has long been a lack of standardized measures with regard to evaluation of interactions with SIAs, especially concerning the "newer" field of social robots. However, in the last five years a significant effort within the research community has been put into addressing this gap, resulting in a constantly growing body of work around questionnaires or other forms of standardized assessments and tests around interactions with SIAs. To

facilitate your research endeavor we collected and systematized questionnaires according to whether they are *dependent variables* that evaluate the outcome of the interaction in some form or whether they are potential *moderating* or *mediating variables*.

A *moderator variable* has the potential to change the strength or direction of an effect between two variables of interest, meaning it affects the relationship between the independent variable and the dependent variable. Gender is often included as a moderator, but also different psychological profiles (e.g., low or high loneliness, low or high self-efficacy in HRI) can have a moderating effect on the relationship. For example, you could find that the usage of an emotionally expressive SIA (in contrast to a non-expressive agent) has a greater impact on acceptance of that agent in female than in male participants. Or you might find that a highly self-disclosing agent (compared to a non-disclosing agent) has a greater impact on perceived likability of the agent for people scoring high in loneliness.

In contrast, a *mediator variable* is a variable that causes mediation in the dependent and the independent variables. In other words, it explains the how or why of an observed relationship between the dependent variable and the independent variable (IV), assuming that the independent variable does not influence the dependent variable (DV) directly but instead does so by means of a third variable. This can either be a complete mediation, meaning the full effect from IV on DV is caused by the mediator variable, or it can be a partial mediation in which only a part of the effect of IV on DV is cause by the mediator (see Jose [2013] for further reading on statistical moderation and mediation analyses).

Moreover, questionnaires assessing *dependent variables*, that is, the outcome of, for example, an interaction, are systematized according to the aspects of the interaction they are measuring (e.g., system performance, social evaluation of SIAs, acceptance, evaluation of overall interaction). Please find an extensive list of available questionnaires in Appendix A.

2.4 Current Challenges

Since SIAs are still a quite young research field, research in this field face many challenges. We decided to address three of these challenges since they are directly related to methods: the replication crisis, the unnecessary conflict between quantitative and qualitative methods, and the lack of long-term and field studies.

2.4.1 Replication Crisis

In a tremendous effort, the Open Science Collaboration made the attempt to replicate 100 psychology studies. In their paper published in 2015 [Open Science Collaboration 2015], the authors reported that only 39% of these studies have replicated the original result. This is rooted on the already ongoing debate about the so-called replication crisis that seems to be especially pronounced in psychology and the social sciences. However, researchers from other disciplines likewise report that their studies were not able to reproduce findings by other scholars and even their own prior work [Baker 2016]. Recently, Irfan et al. [2018] discussed how

the replication crisis in psychology impacts research in the field of human–robot interaction; their arguments also hold for research on virtual agents. They argue that the consequences of the crisis in psychology also affects SIA research because "we do either use research methods similar to those used in other disciplines (and psychology in particular), or rely directly on insights and results handed down from other disciplines" (p.14). Many scholars developing SIAs do so in interdisciplinary teams working closely with psychologists drawing on prior research on social interaction and social relationships in order to design the "socialness" of the SIA. When, however, classical effects from social psychology cannot be replicated with humans, we cannot expect the effect to emerge in interactions with SIAs. Irfan et al.'s good advice is to attempt to replicate the social psychology effect with humans first before running a study with social robots. If the effect can be replicated with humans but not with social robots (or virtual agents), we can classify our null results better; for instance, this would suggest that the social phenomenon is likely not the same or not at all occurring in interactions with SIAs. The takeaway message of Irfan et al. is to be critical and approach the classic effects from social psychology in textbooks with the necessary skepticism that is advisable not only for social psychology but for any discipline. However, we should not be too pessimistic because in fact many classics from social psychology have been shown to be existent in interactions with computers, virtual agents, and social robots as the group of Clifford Nass first and many researchers later have demonstrated (e.g., Reeves and Nass [1996], Hoffmann et al. [2009], von der Pütten et al. [2010], and Eyssel and Hegel [2012]). Nevertheless, the SIA research community should avoid replicating the replication crisis, that is, we should avoid letting the same mechanisms, such as the file drawer problem or publication bias, skew the output of our scientific work by reporting only "significant" results. The community has been picking up on this recently. For instance, the International Conference on Human–Robot Interaction launched a new "reproducability in Human–Robot-Interaction" track that welcomes contributions that "reproduce, replicates, or re-creates prior HRI work (or fails to)" and "provide new HRI artifacts (e.g., datasets, software)" that facilitate reproducability. Selected journals welcome pre-registered reports for replication studies, agreeing to publish the paper regardless of the outcome. Moreover, we should all value studies regardless of whether they produced significant results. Null-results studies are as informative. Defalcating these results can inflate the importance of single significant studies, for instance when a dozen research teams try to replicate a study, fail, and do not report the replication failure, the significant original study gains unwarranted attention given the majority of null effects. Thus the advice is, even when it is harder to publish non-significant results, try. Or at least publish your work accessibly for other scholars in preprint servers such as www.psycharchives.com.

2.4.2 Quantitative and Qualitative Research Methods

The difference between quantitative and qualitative studies can be easily summarized: quantitative studies produce data that can be counted and measured, qualitative studies produce data

that cannot. Both approaches have their drawbacks and advantages, but both should be equally valued as scientific approaches for data collection in order to answer important research questions. As Hammarberg et al. [2016] point out "qualitative and quantitative research methods are often juxtaposed as representing two different world views. In quantitative circles, qualitative research is commonly viewed with suspicion and considered lightweight because it involves small samples which may not be representative of the broader population, it is seen as not objective, and the results are assessed as biased by the researchers' own experiences or opinions. In qualitative circles, quantitative research can be dismissed as over-simplifying individual experience in the cause of generalization, failing to acknowledge researcher biases and expectations in research design, and requiring guesswork to understand the human meaning of aggregate data." (p. 498). "Choosing sides" in research is misleading. We would like to point out that careful and thorough reflection of the research topic and, especially, of the research question should guide the choice for an appropriate method. "The crucial part is to know when to use what method." [Hammarberg et al. 2016, p.498].

- *Quantitative research method:* When you have specific hypotheses, you can identify, isolate, and operationalize variables; when you want to unravel relationships and differences and you want to generalize results and make statements for the population, then choose a quantitative research method (see Section 2.2.1 for a detailed description on how to conduct experiments).
- *Qualitative research method:* When you have a research question that asks about subjective experiences and perspective, you have a specific, small target group from which you want to get background information from or you want to have an in-depth understanding of a specific case, then choose a qualitative research method (see Section 2.2.3.1 for an example of a qualitative study method).

Both methodological approaches can be used in combination, either to research different aspects of the same research question or to research the same aspect and complement and enrich the dataset. In consideration of the exemplary study on robots for learning, results from a quantitative study on the learning outcome could be enriched by interviews assessing students' personal experience of learning together with an agent.

Theoretically and pragmatically, it is not always clear cut which between qualitative and quantitative methods should be used and, sometimes, qualitative and quantitative study approaches are merged together. Strictly speaking, once data is counted and measured, a study is not seen as purely qualitative anymore. However, sometimes qualitative methods are used and data is categorized into clusters that can be analyzed with descriptive statistics. A good example for successfully merging qualitative approaches with quantitative analytical approaches in a virtual agent study is the study by Opfermann and Pitsch [2017]. Special user groups (older adults and individuals with mild cognitive impairments) were confronted

with an embodied conversational agent in a WoZ study. Authors studied the influence of continuous reprompts by the agent that indicate non-understanding of users' interaction attempts and reactions. Analysis of data was done by a sequential protocol of qualitative single-case conversational analyses for each participant followed by quantitative coding including categorizing and frequency counting to compare user behavior and find patterns. Authors conclude with a specific advice for reprompts as an error handling strategy: a reprompt should be given once and it should be unambiguous ("Do you mean yes? Say yes or no." Read Opfermann and Pitsch 2017 for details and for more practical implications).

Valuing both approaches (quantitative and qualitative) but understanding their advantages and disadvantages and, especially, knowing when to use them (and when not to use them) is very important for researchers in psychology and the social sciences. Scholars in SIA should be open-minded and use either approach when appropriate.

2.4.3 Field Studies and Long-term Studies

The third grand challenge in the field of SIA is that we still face a significant lack of field studies and long-term studies. SIAs are envisioned to provide assistance or service, to work together with humans in mixed teams in different working environments, or to offer some form of companionship. As a matter of fact, SIA will have to deal with more than one human in complex social environments. In utter contrast to this envisioned future scenario, research on SIAs has primarily focused on laboratory experiments, examining the interaction between a single human and a single social robot or virtual agent, while research on multi-agent systems is still young and research on HRI groups has only recently begun. As Jung and Hinds [2018] pointed out, this dyad-based research of HRI in laboratory settings "has helped establish a fundamental understanding about people as they interact with robots," but "our theories reflect an oversimplified view of HRI" (p. 1). Although the need for a paradigm shift from studying dyadic interactions in laboratory settings to studying (group) interactions in complex environments has been identified and advocated for [Jung and Hinds 2018], research in this regard is still scarce. This is mostly due to the fact that field studies require a robustly running system that can deal with environmental changes and challenges. Running these studies is expensive, time-consuming, and pose many ethical challenges with regard to informed consent, data protection, and many more. Depending on the type of target group, it is challenging to find participants who agree, for example, to try a robot in their homes for a longer period of time and agree to be under "constant evaluation." As a result, sample sizes of field and long-term studies are often small, leading to negative reviewer comments about the power of the studies. As a community, we should value the tremendous effort that goes into a field or long-term study. Even when the results are not statistically generalizable, these studies provide us with badly needed insights into how our SIAs perform and are perceived in the complex social environments that we design them for. Field and long-term studies that might

result in smaller sample sizes can benefit greatly from combining quantitative and qualitative approaches to assess how successfully an SIA is integrated in the social environment.

2.5 Future Directions

When reviewing the grand challenges that we face regarding methods, we can directly infer future directions for our research. Since the research field of SIA is still quite young most empirical work has been pioneering. In order to establish results and effects we need replication studies that test the robustness of these effects as well as their generalizability to different cultural contexts (see Strait et al. [2020] for an example of cross-cultural replication). Moreover, we should welcome the diversity of our research community and embrace the potential that it offers for interlacing qualitative and quantitative methods. We hope we were able to illustrate how the two methodological traditions can be of mutual benefit instead of hindrance, and especially when it comes to field- and long-term studies that are placed in social context qualitative methods are well suited to take this social context into account, granting a more holistic understanding of the interaction situation and its social meaning than using quantitative methods alone.

2.6 Summary

In this chapter on methods from the social sciences and psychology that can be used for research on SIAs, we provided you with a broad overview about all relevant concepts you should have heard of and taken into account when planning to conduct studies involving human participants. However, keep in mind that methods are an integral part of all disciplines, which usually makes up a significant part of your expert knowledge in a given discipline. This means that although we provided you with the fundamentals, you might need to study more about methods to get to know the sufficiency, or in other words: You have to do the duty first before you can show off. For most of the methodological considerations that we covered in a section, there are specialized books or many research papers that deal with aspects of methodology. It is advisable that once you have chosen a rough direction, you consult more specialized literature on the specific method of your choice. We hope that our suggestions for further reading and recommendations for support tools will facilitate this process. In addition, we hope you will learn that social science methods are not only a duty to fulfil but can be a pleasure as well.

Bibliography

2010. Informed Consent. In N. J. Salkind, ed., *Encyclopedia of research design*. SAGE, Thousand Oaks, Calif. ISBN 9781412961271. DOI: 10.4135/9781412961288.n188.

A. P. Association. 2020. *Publication manual of the American Psychological Association: The official guide to APA style*, seventh edition. ISBN 9781433832154.

M. Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604): 452–454. DOI: 10.1038/533452a.

T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka. 2018. Social robots for education: A review. *Science Robotics*, 3(21): eaat5954. DOI: 10.1126/scirobotics.aat5954.

C. L. Bethel and R. R. Murphy. 2010. Review of Human Studies Methods in HRI and Recommendations. *International Journal of Social Robotics*, 2(4): 347–359. ISSN 1875-4791. DOI: 10.1007/s12369-010-0064-9.

K. Casler, L. Bickel, and E. Hackett. 2013. Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, 29(6): 2156–2160. ISSN 0747-5632. DOI: 10.1016/j.chb.2013.05.009.

L. Christensen. 1988. Deception in Psychological Research. *Personality and Social Psychology Bulletin*, 14(4): 664–675. ISSN 0146-1672. DOI: 10.1177/0146167288144002.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1): 37–46. ISSN 0013-1644. DOI: 10.1177/001316446002000104.

O. S. Collaboration et al. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251).

H. M. Cooper. 2020. *Reporting quantitative research in psychology: How to meet APA style journal article reporting standards*, second edition, revised. APA style products. ISBN 9781433832833.

N. Dahlbäck, A. Jönsson, and L. Ahrenberg. 1993. Wizard of Oz studies: why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces*, pp. 193–200.

A. Ertas and J. C. Jones. 1996. *The engineering design process*, 2. ed. Wiley, New York. ISBN 0471136999.

F. Eyssel and F. Hegel. 2012. (s) he's got the look: Gender stereotyping of robots 1. *Journal of Applied Social Psychology*, 42(9): 2213–2230.

A. Field. 2018. *Discovering statistics using IBM SPSS statistics*, 5th edition. SAGE, Los Angeles and London and New Delhi and Singapore and Washington DC and Melbourne. ISBN 9781526419521.

A. Field and G. Hole. 2013. *How to design and report experiments*, repr. SAGE, Los Angeles. ISBN 9780761973829.

A. Fleischer, A. D. Mead, and J. Huang. 2015. Inattentive Responding in MTurk and Other Online Samples. *Industrial and Organizational Psychology*, 8(2): 196–202. ISSN 1754-9426. DOI: 10.1017/iop.2015.25.

D. J. Follmer, R. A. Sperling, and H. K. Suen. 2017. The Role of MTurk in Education Research: Advantages, Issues, and Future Directions. *Educational Researcher*, 46(6): 329–334. ISSN 0013-189X. DOI: 10.3102/0013189X17725519.

C. Frauenberger, J. Good, and W. Keay-Bright. 2011. Designing technology for children with special needs: bridging perspectives through participatory design. *CoDesign*, 7(1): 1–28. ISSN 1571-0882. DOI: 10.1080/15710882.2011.587013.

F. J. Gravetter and L. A. Forzano. 2012. *Research Methods for the Behavioral Sciences*, 4. Wadsworth, Cengage Learning, Belmont, CA.

G. M. Hall. 2012. *How to Write a Paper*. John Wiley & Sons, Ltd, Chichester, UK. ISBN 9781118488713. DOI: 10.1002/9781118488713.

K. Hammarberg, M. Kirkman, and S. de Lacey. 2016. Qualitative research methods: when to use them and how to judge them. *Human reproduction (Oxford, England)*, 31(3): 498–501. DOI: 10.1093/humrep/dev334.

D. J. Hauser and N. Schwarz. 2016. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*, 48(1): 400–407. DOI: 10.3758/s13428-015-0578-z.

L. Hoffmann, N. C. Krämer, A. Lam-Chi, and S. Kopp. 2009. Media equation revisited: do users show polite reactions towards an embodied agent? In *International Workshop on Intelligent Virtual Agents*, pp. 159–165. Springer.

B. Irfan, J. Kennedy, S. Lemaignan, F. Papadopoulos, E. Senft, and T. Belpaeme. 2018. Social Psychology and Human-Robot Interaction. In T. Kanda, S. Ŝabanović, G. Hoffman, and A. Tapus, eds., *HRI'18 companion, March 5-8, 2018, Chicago, IL, USA*, pp. 13–20. ACM, New York, NY, USA. ISBN 9781450356152. DOI: 10.1145/3173386.3173389.

P. E. Jose. 2013. *Doing statistical mediation and moderation*. Methodology in the social sciences. The Guilford Press, New York. ISBN 9781462508211.

M. Jung and P. Hinds. 2018. Robots in the Wild. *ACM Transactions on Human-Robot Interaction*, 7(1): 1–5. ISSN 2573-9522. DOI: 10.1145/3208975.

S. Kvale. 1983. The Qualitative Research Interview. *Journal of Phenomenological Psychology*, 14(1-2): 171–196. DOI: 10.1163/156916283X00090.

H. R. Lee, S. Šabanović, W.-L. Chang, S. Nagata, J. Piatt, C. Bennett, and D. Hakken. 2017. Steps Toward Participatory Design of Social Robots. In B. Mutlu, M. Tscheligi, A. Weiss, and J. E. Young, eds., *HRI'17*, pp. 244–253. IEEE, Piscataway, NJ. ISBN 9781450343367. DOI: 10.1145/2909824.3020237.

H. M. Levitt. 2020. *Reporting qualitative research in psychology: How to meet APA style journal article reporting standards*, revised edition. ISBN 9781433833434.

R. M. Montoya, R. S. Horton, and J. Kirchner. 2008. Is actual similarity necessary for attraction? A meta-analysis of actual and perceived similarity. *Journal of Social and Personal Relationships*, 25(6): 889–922. ISSN 0265-4075. DOI: 10.1177/0265407508096700.

E. A. Necka, S. Cacioppo, G. J. Norman, and J. T. Cacioppo. 2016. Measuring the Prevalence of Problematic Respondent Behaviors among MTurk, Campus, and Community Participants. *PloS one*, 11(6): e0157732. DOI: 10.1371/journal.pone.0157732.

G. Norman. 2010. Likert scales, levels of measurement and the laws of statistics. *Advances in health sciences education : theory and practice*, 15(5): 625–632. DOI: 10.1007/s10459-010-9222-y.

C. Opfermann and K. Pitsch. 2017. Reprompts as error handling strategy in human-agent-dialog? User responses to a system's display of non-understanding. In *Human-robot collaboration and human assistance for an improved quality of life*, pp. 310–316. IEEE, Piscataway, NJ. ISBN 978-1-5386-3518-6. DOI: 10.1109/ROMAN.2017.8172319.

E. K. Perrault and D. M. Keating. 2018. Seeking Ways to Inform the Uninformed: Improving the Informed Consent Process in Online Social Science Research. *Journal of empirical research on human research ethics : JERHRE*, 13(1): 50–60. DOI: 10.1177/1556264617738846.

S. Q. Qu and J. Dumay. 2011. The qualitative research interview. *Qualitative Research in Accounting & Management*, 8(3): 238–264. ISSN 1176-6093. DOI: 10.1108/11766091111162070.

B. Reeves and C. I. Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge university press.

H. T. Reis and C. M. Judd. 2013. *Handbook of Research Methods in Social and Personality Psychology*. Cambridge University Press, New York. ISBN 9780511996481. DOI: 10.1017/CBO9780511996481.

L. Riek. 2012. Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *Journal of Human-Robot Interaction*, pp. 119–136. DOI: 10.5898/JHRI.1.1.Riek.

A. M. Rosenthal-von der Pütten and N. C. Krämer. 2015. Individuals' Evaluations of and Attitudes Towards Potentially Uncanny Robots. *International Journal of Social Robotics*, 7(5): 799–824. ISSN 1875-4791. DOI: 10.1007/s12369-015-0321-z.

J. A. Russell. 1994. Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological bulletin*, 115(1): 102.

S. Šabanović, W.-L. Chang, C. C. Bennett, J. A. Piatt, and D. Hakken. 2015. A Robot of My Own: Participatory Design of Socially Assistive Robots for Independently Living Older Adults Diagnosed with Depression. In J. Zhou and G. Salvendy, eds., *Design for aging*, volume 9193 of *Lecture notes in computer science Information systems and applications, incl. Internet/web, and HCI*, pp. 104–114. Springer, Cham. ISBN 978-3-319-20891-6. DOI: 10.1007/978-3-319-20892-3_11.

D. B. Shank. 2016. Using Crowdsourcing Websites for Sociological Research: The Case of Amazon Mechanical Turk. *The American Sociologist*, 47(1): 47–55. ISSN 0003-1232. DOI: 10.1007/s12108-015-9266-9.

P. E. Shrout and J. L. Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2): 420–428. ISSN 0033-2909. DOI: 10.1037/0033-2909.86.2.420.

N. A. Smith, I. E. Sabat, L. R. Martinez, K. Weaver, and S. Xu. 2015. A Convenient Solution: Using MTurk To Sample From Hard-To-Reach Populations. *Industrial and Organizational Psychology*, 8(2): 220–228. ISSN 1754-9426. DOI: 10.1017/iop.2015.29.

S. M. Smith, C. A. Roster, L. L. Golden, and G. S. Albaum. 2016. A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples. *Journal of Business Research*, 69(8): 3139–3148. ISSN 01482963. DOI: 10.1016/j.jbusres.2015.12.002.

M. Strait, F. Lier, J. Bernotat, S. Wachsmuth, F. Eyssel, R. Goldstone, and S. Šabanović. 2020. A Three-Site Reproduction of the Joint Simon Effect with the NAO Robot. In T. Belpaeme, J. Young, H. Gunes, and L. Riek, eds., *HRI '20*, pp. 103–111. Association for Computing Machinery, New York, NY. ISBN 9781450367462. DOI: 10.1145/3319502.3374783.

M. C.-T. Tai. 2012. Deception and informed consent in social, behavioral, and educational research (SBER). *Tzu Chi Medical Journal*, 24(4): 218–222. ISSN 10163190. DOI: 10.1016/j.tcmj.2012.05.003.

A. M. von der Pütten, N. C. Krämer, J. Gratch, and S.-H. Kang. 2010. "It doesn't matter what you are!" Explaining social effects of agents and avatars. *Computers in Human Behavior*, 26(6): 1641–1650. ISSN 0747-5632. DOI: 10.1016/j.chb.2010.06.012.

C. Zaga, M. Lohse, K. P. Truong, and V. Evers. 2015. The Effect of a Robot's Social Character on Children's Task Engagement: Peer Versus Tutor. In A. Tapus, E. André, J.-C. Martin, F. Ferland, and M. Ammi, eds., *Social robotics*, volume 9388 of *Lecture notes in computer science Lecture notes in artificial intelligence*, pp. 704–713. Springer, Cham and Heidelberg and New York and Dordrecht and London. ISBN 978-3-319-25553-8. DOI: 10.1007/978-3-319-25554-5_70.

M. V. Zelkowitz and D. R. Wallace. 1998. Experimental models for validating technology. *Computer*, 31(5): 23–31. ISSN 00189162. DOI: 10.1109/2.675630.

Table 1: Appendix - Questionnaires in HRI.

Name of Questionnaire	Measured Construct(s)	IV, DV, MV	Reference
Evaluation of Agents / Interactions with Agents			
General Impressions of Humanoids	Familiarity	DV	Kamide et al. (2013)
General Impressions of Humanoids	Repulsion	DV	Kamide et al. (2013)
General Impressions of Humanoids	Utility	DV	Kamide et al. (2013)
General Impressions of Humanoids	Performance	DV	Kamide et al. (2013)
General Impressions of Humanoids	Motion	DV	Kamide et al. (2013)
General Impressions of Humanoids	Voice	DV	Kamide et al. (2013)
General Impressions of Humanoids	Sound	DV	Kamide et al. (2013)
General Impressions of Humanoids	Humanness	DV	Kamide et al. (2013)
General Impressions of Humanoids	Entitativity	DV	Kamide et al. (2013)
The Human-Robot Interaction Evaluation Scale (HRIES)	Sociability	DV	Spatola et al. (2020)
The Human-Robot Interaction Evaluation Scale (HRIES)	Animacy	DV	Spatola et al. (2020)
The Human-Robot Interaction Evaluation Scale (HRIES)	Agency	DV	Spatola et al. (2020)
The Human-Robot Interaction Evaluation Scale (HRIES)	Disturbance	DV	Spatola et al. (2020)
Godspeed Questionnaire	Animacy	DV	Bartneck et al. (2009)
Godspeed Questionnaire	Anthropomorphism	DV	Bartneck et al. (2009)
Godspeed Questionnaire	Likeability	DV	Bartneck et al. (2009)
Godspeed Questionnaire	Perceived Intelligence	DV	Bartneck et al. (2009)
Godspeed Questionnaire	Perceived Safety	DV	Bartneck et al. (2009)
Animated Character and Interface Evaluation	Anxiety	DV	Rickenberg & Reeves (2000)
Animated Character and Interface Evaluation	Task Performance	DV	Rickenberg & Reeves (2000)
Continued on next page			

Table 1 – continued from previous page

Name of Questionnaire	Measured Construct(s)	IV, DV, MV	Reference
Animated Character and Interface Evaluation	Liking	DV	Rickenberg & Reeves (2000)
The Robotic Social Attributes Scale (RoSAS)	Warmth	DV	Carpinella et al. (2017)
The Robotic Social Attributes Scale (RoSAS)	Competence	DV	Carpinella et al. (2017)
The Robotic Social Attributes Scale (RoSAS)	Discomfort	DV	Carpinella et al. (2017)
Attitudes, Emotions and Expectations in Interaction			
Rapport–Expectation Robot Scale (RERS)	Expectation as a conversation partner	IV, DV, MV	Nomura & Kanda (2016)
Rapport–Expectation Robot Scale (RERS)	Expectation for togetherness	IV, DV, MV	Nomura & Kanda (2016)
Robot Anxiety Scale (RAS)	Anxiety toward Communication Capability of Robots	IV, DV, MV	Nomura et al. (2006)
Robot Anxiety Scale (RAS)	Anxiety toward Behavioral Characteristics of Robots	IV, DV, MV	Nomura et al. (2006)
Robot Anxiety Scale (RAS)	Anxiety toward Discourse with Robots	IV, DV, MV	Nomura et al. (2006)
Assessment of Attitudes Towards Social Robots (ASOR)	Mental capacities	IV, DV, MV	Damholdt et al. (2020)
Assessment of Attitudes Towards Social Robots (ASOR)	Socio-practical capacities	IV, DV, MV	Damholdt et al. (2020)
Assessment of Attitudes Towards Social Robots (ASOR)	Socio-moral status	IV, DV, MV	Damholdt et al. (2020)
Negative Attitudes towards Robots Scale (NARS)	Negative Attitude toward Situationsof Interaction with Robots	IV, DV, MV	Nomura et al. (2006)
Negative Attitudes towards Robots Scale (NARS)	Negative Attitude toward Social Influence of Robots	IV, DV, MV	Nomura et al. (2006)
Negative Attitudes towards Robots Scale (NARS)	Negative Attitude toward Emotions in Interaction with Robots	IV, DV, MV	Nomura et al. (2006)

Continued on next page

Table 1 – continued from previous page

Name of Questionnaire	Measured Construct(s)	IV, DV, MV	Reference
Frankenstein Syndrom Questionnaire (FSQ)	General anxiety toward humanoid robots	DV	Nomura et al. (2012)
Frankenstein Syndrom Questionnaire (FSQ)	Apprehension toward social risks of humanoid robots	DV	Nomura et al. (2012)
Frankenstein Syndrom Questionnaire (FSQ)	Trustworthiness for developers of humanoid robots	DV	Nomura et al. (2012)
Frankenstein Syndrom Questionnaire (FSQ)	Expectation for humanoid robots in daily life	DV	Nomura et al. (2012)
Measurement of Moral Concern for Robots	Basic moral concern		Nomura et al. (2019)
Measurement of Moral Concern for Robots	Concern for psychological harm		Nomura et al. (2019)
Self-Efficacy in HRI	Self-efficacy expectations	IV, DV, MV	Rosenthal-von der Pütten & Bock (2018)
Embodiment, Physical Presence, Social Presence and Co-Presence			
Social Presence Survey	Social Presence	DV or MV	Bailenson et al. (2003)
Networked Minds Questionnaire of Social Presence	Social Presence	DV or MV	Biocca et al. (2003)
Networked Minds Questionnaire of Social Presence	Co-presence	DV or MV	Biocca et al. (2003)
Networked Minds Questionnaire of Social Presence	Subjective symmetry	DV or MV	Biocca et al. (2003)
Networked Minds Questionnaire of Social Presence	Intersubjective symmetry	DV or MV	Biocca et al. (2003)
Kidd and Breazeal Questionnaire	Perceived Presence	DV or MV	Kidd & Breazeal (2004)
Lombart & Ditton Presence Questionnaire	Presence	DV or MV	Lombard et al. (2000)
Embodiment and Corporeality Questionnaire	Corporeality	DV or MV	Hoffmann et al. (2018)
Embodiment and Corporeality Questionnaire	Mobility and Tactile Interaction	DV or MV	Hoffmann et al. (2018)
Embodiment and Corporeality Questionnaire	Shared Perception	DV or MV	Hoffmann et al. (2018)
Embodiment and Corporeality Questionnaire	Nonverbal Expressiveness	DV or MV	Hoffmann et al. (2018)

Continued on next page

Table 1 – continued from previous page

Name of Questionnaire	Measured Construct(s)	IV, DV, MV	Reference
Usability and User Experience			
Hoonhout Enjoyability Scale	Product Enjoyability	DV	Hoonhout (2002)
User Experience Questionnaire	User Experience	DV	Laugwitz et al. (2008)
System Usability Scale	System Usibility	DV	Brooke (1996)
Questionnaires for Children and Adolescents			
Children's Social Behavior Questionnaire	Social Behavior, Children	IV	Hartman et al. (2006)
Emotion Awareness Questionnaire for children	Emotion, Children	IV	Rieffe et al. (2008)
Technology-Specific Satisfaction Scale (TSSS) (child)	Satisfaction, Children	DV	Alves-Oliveira et al. (2015)
Technology-Specific Expectations Scale (TSES) (child)	Expectations (Capabilites & Fiction), Children	DV	Alves-Oliveira et al. (2015)
Trust in Technology			
Human-Robot Trust Scale	Trust (Robot)	IV, DV, MV	Schaefer (2013)
Scale of Trust in Automated Systems	Trust (System)	IV, DV, MV	Jian et al. (2000)
Human-Computer Trust Scale	Trust (Computer; Reliability, Technical Competence, Perceived Understandability, Faith, Personal Attachment)	IV, DV, MV	Madsen & Gregor (2000)
Psychological States, Emotion, Motivation, Sastisfaction and Stress			
Self Assessment Manikin and Semantic Differential	Emotional state	DV or MV	Bradley & Lang (1994)
Positive and Negative Affect Schedule (PANAS)	Affective state	DV or MV	Watson et al. (1988)
Satisfaction With Life Scale	Satisfaction with life	IV	Diener et a. (1985)
Situational Motivation Scale (SIMS)	Intrinsic motivation, identified regulation, external regulation, and amotivation	DV or MV	Guay et al. (2000)
Academic motivation scale	Intrinsic, Extrinsic, and Amotivation in Education	IV, DV, MV	Vallerand et al. (1992)

Continued on next page

Table 1 – continued from previous page

Name of Questionnaire	Measured Construct(s)	IV, DV, MV	Reference
Students' motivation toward science learning (SMTSL)	Motivation to learn science	IV, DV, MV	Tuan et al. (2005)
English Language Learner Motivation Scale (ELLMS)	Motivation to learn the english language	IV, DV, MV	Ardasheva et al. (2012)
UCLA Loneliness Scale	Loneliness	IV	Russel (1996)
Percieved Stress Scale (PSS)	Stress	IV, DV, MV	Cohen et al. (1983)
New General Self-Efficacy Scale	General self efficacy	IV, DV, MV	Chen et al. (2001)
Standardized Mini-Mental State Examination Development	Mental State, Cognitive Abilities	IV	Crum et al. (1993)
CES-D Scale: A Self-Report Depression Scale	Depressive symptomatology in general population	IV	Radloff (1977)
Psychological Traits and Diagnostic Measurements			
Eysenck Personality Questionnaire	Personality	IV	Francis et al. (1992)
Big Five Questionnaire	Personality	IV	Caprara et al (1993)
Barratt-Impulsiveness-Scale (BIS 11)	Impulsiveness	IV	Patton et al. (1995)
The Aggression Questionnaire	Physical Aggression, Verbal Aggression, Anger, Hostility	IV	Buss & Perry (1992)
Buss-Perry Aggression Questionnaire short form	Physical Aggression, Verbal Aggression, Anger, Hostility	IV	Bryant & Smith (2003)
Emotion Regulation Questionnaire	Emotion regulation (Suppression and reappraisal)	IV	Gross & John (2003)
Task Related Cognitive Load			
Questionnaire for Placement Committees	Cognitive Development	DV	Fridin & Belokopytov (2014)
NASA Task Load Index Questionnaire	Cognitive Load	DV	Hart (2006)
Cognitive Load Questionnaire	Cognitive Load	DV	Sweller (1988)