Designing and Conducting Human-Robot Interaction Studies: Guidelines and an interactive specification approach

LEIMIN TIAN, NICOLE ROBINSON, PAMELA CARRENO-MEDRANO, WESLEY P. CHAN, MARAM SAKR, TINA WU, ELAHE ABDI, ELIZABETH CROFT, DANA KULIĆ, Faculty of Engineering, Monash University, Australia

Human-Robot Interaction (HRI) is a rapidly growing field. As personal and service robotics advance into a range of markets, the success of HRI will determine the benefit society derives from robotics technology. The field of HRI is facing a critical transformation that calls for clear and measurable ways to evaluate the effectiveness and outcomes of an HRI system, and to understand users' perceptions and experiences. To improve the overall quality of HRI studies to match other related disciplines, such as human-computer interaction or psychology, high-quality and scientifically rigorous methods are required to ensure the robustness of the results and draw conclusions across studies. In addition, standardized methodologies improve the validity, repeatability, and generalisability of the findings, which prevents the replication crisis, and provides feedback for design of new systems. Thus, we are motivated to propose a set of HRI study methodology guidelines supported by reviews of existing methodological approaches in HRI, as well as in the broader fields of human-computer interaction, human-centred artificial intelligence, psychology and social science. Researchers are encouraged to follow these guidelines when designing and conducting future HRI studies, which leads to high-quality study designs and robust outcomes that advances the frontier of HRI research and applications.

Additional Key Words and Phrases: human-robot interaction, social robotics, user study, methodology, guidelines

ACM Reference Format:

Leimin Tian, Nicole Robinson, Pamela Carreno-Medrano, Wesley P. Chan, Maram Sakr, Tina Wu, Elahe Abdi, Elizabeth Croft, Dana Kulić. 2020. Designing and Conducting Human-Robot Interaction Studies: Guidelines and an interactive specification approach. *ACM Trans. Hum.-Robot Interact.* 0, 0, Article 0 (2020), 54 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Increased attention and importance has been placed on adopting rigorous study methodology in multiple scientific fields. For example, the field of behavioural and social science is experiencing a "replication crisis" [135], in which efforts to replicate existing studies often do not yield the same observations or conclusions compared to the original experiment. This replication crisis damages the validity of existing research findings, resulting in incorrect assumptions about phenomena and subsequent waste of resources to revise or verify earlier work. Social science researchers have been pushing towards awareness for good research practices that ensure the validity, repeatability, and generalisability of the studied phenomena, and human-robot interaction should follow in the same path.

Author's address: Leimin Tian, Nicole Robinson, Pamela Carreno-Medrano, Wesley P. Chan, Maram Sakr, Tina Wu, Elahe Abdi, Elizabeth Croft, Dana Kulić, Faculty of Engineering, Monash University, 14 Alliance Ln, Clayton, VIC, 3168, Australia, {Leimin.Tian,Nicole.Robinson,Pamela.Carreno,Wesley.Chan,Maram.Sakr,Lee.Wu,Elahe.Abdi,Elizabeth.Croft,Dana.Kulic}@monash.edu.

Manuscript submitted to ACM

The field of human-robot interaction (HRI) is a rapidly growing discipline of robotics research, which focuses on how robots and humans interact or collaborate with one another, and identifying effective robot designs that facilitate such interaction or collaboration. HRI studies often adopts a multidisciplinary theoretical and experimental framework, including but not limited to robotics, behavioural and social science. As a emergent field, existing HRI research contains mostly exploratory studies with a less rigorous methodology as an initial review of concepts and outcomes, for example, an uncontrolled pilot with a small sample size or one key measurement to assess a factor or variable. As the field begins to mature, it is vital to improve the quality of research through tests of validity and generalisability in HRI studies, which can allow findings to be applied to a wider domain. Thus, we are motivated to identify a set of methodological guidelines that ensured the quality and robustness of HRI studies, which is urgently needed to prevent a "replication crisis" similar to the psychology discipline from happening in the HRI field.

There are a range of limitations and challenges in current HRI studies. A number of reviews have been conducted to review and address these challenges in HRI, such as evaluation metrics (e.g., Belhassein et al. [20], Leichtmann and Nitsch [124]). However, an overview of the HRI study life cycle, which captures the whole process from conceiving a research question to post-study follow-ups has not yet been addressed. This is important because researchers following rigorous methodological practice can help to reduce the impact and prevalence of lower-quality scientific studies, which can contribute to a possible replication crisis for HRI in the future. Moreover, HRI researchers, especially those new to the field, can benefit from having a set of general HRI study methodology guidelines that recommends best practices for each aspect of the study life cycle. This can help to teach earlier career researchers about good practice for their research, improving its capacity to contribute meaningful results to the domain. This paper proposes such guidelines with the support of existing HRI studies and research from related fields. For HRI researchers new to the field, we offer an overview of the HRI study life cycle that introduces them to good practices for conducting HRI studies that they can follow. For advanced or experienced HRI researchers, we provide in-depth analysis of strengths and weaknesses of existing HRI methodological approaches illustrated with recent HRI studies, as well as offering insights on novel methods that may be adopted in future studies.

1.1 Life cycle of an HRI study

We propose an overview of the life cycle of a typical HRI study as shown in Figure 1. An interactive version of the guideline can be accessed at https://tianleimin.github.io/HRI-Methodology-Guidelines/. This tool is provided to assist researchers in developing their study design and to present relevant guidelines at each step of the study. Each step in this HRI study life cycle will be explained in this paper and we will discuss specific recommendations for key steps in this life cycle, together with examples of existing HRI studies to illustrate these recommendations. As HRI is a highly interdisciplinary research field, we will support our discussion both with existing HRI studies and with methodological approaches from complementary fields, including human-computer interaction, human-centred AI, psychology and cognitive science.

1.2 Aims and scope

Unlike existing HRI methodology reviews that summarise current approaches, we focus on identifying best practices for various types of HRI studies, and aim to provide specific guidelines applicable to key steps during an HRI study cycle. We begin with an overview of existing reviews and guidelines for HRI study methodology in Section 2 to provide the rationale for such HRI methodology guidelines. We then cover a variety of HRI studies, methodological approaches, and application domains in Sections 3, 4, 5, and 6 in order to gain a broad view of the field. For each aspect of HRI

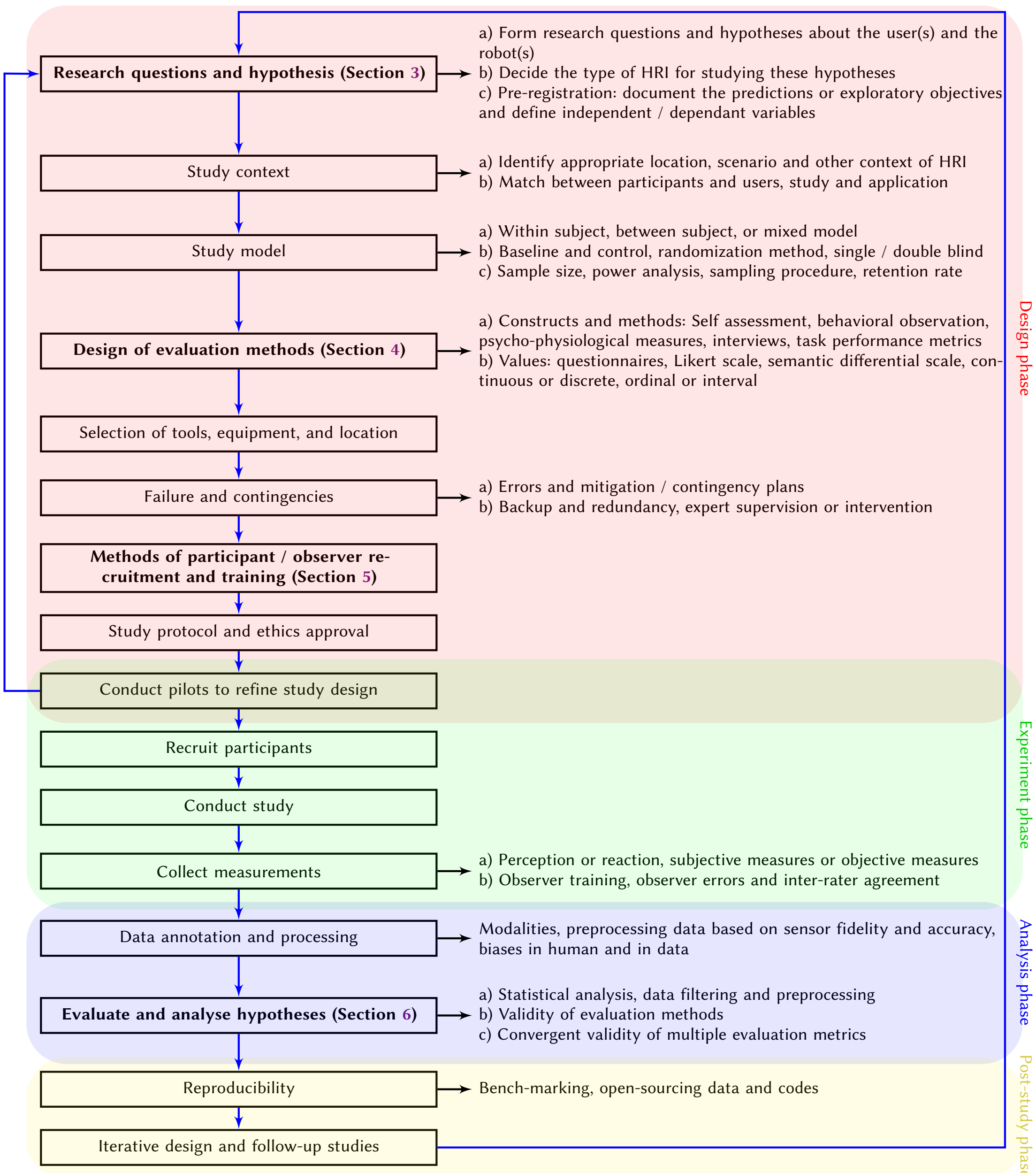

Fig. 1. Life cycle of an HRI study (click for an interactive version online).

study methodology, our discussion is illustrated with examples from existing work. Thus, our work provides a set of applicable and practical guidelines to improve the quality of HRI studies, particularly beneficial for researchers new to the field.

2 BACKGROUND

Existing guidelines for conducting HRI studies either introduce researchers to an overview of the study process (e.g., Belhassein et al. [20]), or focus on reviewing a specific aspect of HRI study, such as one style of interaction (e.g., Wizard-of-Oz (WoZ) style studies [163]), one type of assessment approaches (e.g., psychophyciological assessment [103]), one key step during the HRI study life cycle (e.g., evaluation [181]), one application domain (e.g., healthcare [166]), or one cohort of participants (e.g. elderly [232]). In this section, we will discuss representative examples of existing reviews and guidelines for HRI methodology, and provide motivation for this work. In particular, the HRI methodology guidelines we offer drew from a variety of HRI studies to provide a holistic view, while focusing on recommending suggestions applicable to core steps during an HRI study cycle.

2.1 Existing reviews and guidelines on HRI study methodology

Current HRI research is in urgent need of a scientifically rigorous methodology that ensures the validity and generalisability of the findings. A review of literature reveals that many HRI studies are designed and executed based on the preferences of researchers and available resources. For example, less than 10% of studies on social robots for healthcare have deployed randomized controlled trials for addressing potential biases when testing the effectiveness of the social robots, with even fewer conducting follow-up studies [166]. In the elder care domain, two-thirds of the studies had less than 30 participants, while the majority lasted a day, and only 11% lasted 30 days or more. Most importantly, 90% of the studies lacked a theoretical framework derived from human research [232]. That said, systematic approaches to HRI studies with detailed study design methods are gaining more attention in recent years.

Categorisation of HRI studies: HRI studies have been categorised in multiple review papers based on a number of attributes. Various taxonomies have been proposed in the literature for classifying HRI. These include taxonomies classifying HRI by the type of study [46], the role of the robot [42], its morphology [228], level of autonomy [19, 207], behaviour [188] and characteristics [156], the role of the human [180], type of task [193, 228] and interaction [147, 184] as well as human-robot proximity [4, 76] and coordination [91, 105]. The HRI study themes, types, and challenges as of the year 2008 were reviewed by Goodrich and Schultz [76]. Five attributes of HRI were highlighted including autonomy, task type and structure of the human-robot team. A surge in longitudinal studies, interdisciplinary research and blended simulated/physical experiments has begun to emerge. In a similar approach, Baxter et al. [18] categorised papers accepted at the ACM/IEEE International Conference on Human-Robot Interaction between years 2013-2015 based on level of robot autonomy, participant populations, evaluation environments, length of empirical studies, approach to statistics and reproducibility. In Baxter et al. [18] the authors provide a set of recommendations, highlighting the importance of clarity of the research goal, level of robot autonomy and statistical analysis, as well as justification of the number of participants, length of interaction and reproducibility of the study. Establishment of standards and common metrics is reported as an emerging effort, however, some years later, the HRI community is still moving towards this goal.

Hypothesis, evaluation metrics and use of human-human interaction theories: Hypotheses in HRI studies are often grounded by psychology theories and other relevant disciplines, such as cognitive science. Thus, it is inevitable that the recent replication crisis in psychology has also influenced the field of HRI. For example, Irfan et al. [103] presented their failures to replicate the social facilitation theory in HRI. Based on this experience, they recommended HRI researchers who are consumers of psychology literature avoid older psychology literature with weak methods, and encouraged bench-marking efforts in HRI. Their work highlighted the challenges of observing and measuring social interaction in HRI. Similarly, Leichtmann and Nitsch [124] reviewed HRI studies on human-robot physical and social distance, and investigated whether it is appropriate to generate hypotheses for HRI studies based on human-human interaction (HHI) theories. They found that HRI study hypotheses are often derived from HHI under the assumption that robots are treated as social actors in a manner similar to humans. However, this assumption may not always be valid. Based on their review, the authors recommended pre-registration of hypotheses, methodology, and analysis protocol, as well as transparent reporting of HRI studies grounded with HHI theories.

Design of HRI studies: Most reviews of HRI methodology have focused on the evaluation aspect, and provide recommendations regarding the design phase of the HRI study life cycle. A decade ago, Bethel and Murphy [27] identified study size and lack of multiple assessment methods as the primary issues in existing studies. They argued that a large sample size better represents the target population and exhibits a higher probability for statistically significant results. They also suggested that the deployment of three or more evaluation metrics for the same construct produces more reliable and accurate results. Currently, many metrics used in HRI studies remain "observed" with a lack of functional/generalizable measurement mechanisms. For example, Murphy and Schreckenghost [143] proposed an HRI metric taxonomy based on a review of 29 studies. A total of 42 metrics were categorised into three groups: human (e.g., reliability, productivity and awareness), robot (e.g., plan state and time in manual/autonomous operation), and system (e.g., efficiency, safety and coactivity). Psychological assessment metrics such as cardiovascular, electro-dermal and brain activity are increasingly used to evaluate human responses in HRI [208]. When studying social interactions and human responses in HRI, the most commonly adopted approach is to use the Likert scale to measure subjective judgements of participants. However, a recent review on the use of the Likert scale in HRI studies found that only 3 out of the 110 papers performed their analyses correctly [181]. This demonstrated the urgent need for methodological guidelines to ensure the quality and robustness of HRI studies, especially regarding the validity of evaluation methods and appropriate analyses of different metrics.

Post-study and reproducibility: In contrast to reviews focusing on the design phase of HRI studies, Belhassein et al. [20] provided recommendations for the experiment and post-study phases of HRI user studies with a focus on participants and the replication crisis. Regarding the experiment phase, they suggested that the outcome of the HRI study can be improved by recruiting more users with diverse backgrounds, rigorous implementation of the protocol, allowing time for the user to get accustomed to the robot, ensuring physical and psychological safety, assessing suitability of the robot for the task, choosing the right and preferably standardized measurements, deployment of theoretically solid tools and measures and finally providing required details for reproducing the study. Regarding the post-study phase, they suggested standardization of the technical components to address study reproducibility.

Evaluation of HRI systems' usability and technical limitations: Many HRI systems are designed with the goal of application to education, health, and elderly care [47]. This relies on advancements in sensory technologies and algorithms in robotics [205]. Furthermore, to evaluate the outcomes of such HRI systems, it is critical to evaluate HRI in-situ. For example, Weiss et al. [219] proposed the usability, social acceptance, user experience, and societal impact (USUS) framework. To envision new research questions beyond existing technical restraints, the Wizard of Oz (WoZ) approach is often applied in HRI studies, especially when studying complex social interactions. The WoZ approach facilitates experiment safety and control of variable factors in HRI. Riek [163] reviewed how WoZ has been used in HRI in order to identify valid WoZ designs. They stressed the importance of operator training and instruction, as well as the formation of hypotheses on user and robot behaviors.

One-off HRI studies: Existing reviews of HRI methodology often focus on one particular aspect or evaluation method. For example, Lasota et al. [121] reviewed factors influencing safety in HRI and different measurements of safety; Young et al. [231] reviewed the use of biological features and means to incorporate user variances; Prewett et al. [153] reviewed the measurement of robot teleoperator workload in relation to task outcomes. It is important to have such in-depth investigation on one type of HRI or on one aspect of HRI evaluation. However, identifying a set of guidelines that inform on core steps of an HRI study cycle is crucial to improve the overall quality of HRI studies.

Comprehensive overviews of HRI studies: Bartneck et al. [16] provided a comprehensive overview of HRI studies. In this report of good practices in HRI studies, research questions (explanatory vs. confirmatory), study designs (qualitative vs. quantitative vs. mixed methods), participants (choice of population and sample size), context (location, time and social unit), robot appearance and functionality, mode of interaction (WoZ vs. physical vs. simulated) as well as direct and indirect metrics, statistical analysis and ethical considerations are covered. Their discussions on (dis)advantages of each approach and analysis of the best practices provides a comprehensive introduction to HRI study. Similarly, Hoffman and Zhao [93] proposed an introduction to HRI methodology and a tutorial for researchers new to the field. A hypothetical HRI study was used as an example to illustrate the process of how an HRI study is conceived is conducted.

2.2 Summary

While there have been a number of exemplar contributions in this area, the existing HRI study methodology review papers are mainly focused on reporting the current common practices. They lack an overview for the HRI study cycle and a overall guideline for identifying the best study protocols and evaluation metrics. Further, in these studies, the discussions on (dis)advantages of each approach and analysis of the best practices are generally limited.

Because of the wide variation in research questions and experimental settings in HRI, a one-size-fits all methodological approach across all HRI studies is not a reasonable expectation. However, we have identified five shared recommendations emerging from existing reviews of HRI studies covered in this section:

Shared recommendations emerging from existing HRI reviews

(1) **Relevance of psychology and social science theories to HRI**: Grounding HRI hypotheses with theories in psychology and social science helps to increase the strength of the HRI hypotheses. However, researchers should also be aware of the differences between human-human interaction and HRI, which can limit the applicability of psychology or social science theories in HRI.
(2) **The importance of context**: HRI is context dependant. Researchers are encouraged to have an inter-disciplinary perspective when conducting HRI studies, and to incorporate social and interactive context in their research, such as the target user population or the location of the interaction.
(3) **Pay attention to evaluation**: The validity of evaluation metrics used is directly related to the validity of an HRI study. Researchers are recommended to use multiple metrics and evaluation approaches to measure a parameter, and to investigate the convergence of these metrics. In addition, researchers are encouraged to use appropriate statistical tests and conduct power analyses when studying experimental results.
(4) **Ecological validity of a study**: Researchers are encouraged to conduct field studies and longitudinal HRI studies. In HRI experiments, it is recommended to recruit participants from the target populations and pay attention to sample size when analyzing the results.
(5) **Rigour and transparency are key**: To improve the reproducibility and generalizability of HRI studies, researchers are encouraged to pre-register their hypotheses. They should also consult their target publication venues on whether or not pre-registration is requested. Researchers should cover both significant and non-significant results in their discussion, and open-source the code and data used in their work.

In the following sections, we will investigate each of these aspects of HRI study and discuss the best practices for different types of study. In particular, we will cover hypotheses and study context in Section 3, evaluation methods in Section 4, participants and ecological validity in Section 5, and reporting metrics in Section 6.

3 HYPOTHESIS IN HRI STUDIES

3.1 From research question to hypothesis

The experimental design process begins with refining the research question(s) that the experimenter wishes to explore. An effective method for developing strong research questions begins with a problem statement describing a gap in knowledge. The identification of this gap can be developed through literature review [15], personal observation, drawing similarities to other fields [42] or exploratory/pilot studies, among other methods. Each research question aims to fill a knowledge gap such that others in the field can move forward building on this new information. Once a research question is defined, a hypothesis can be developed to allow experiments that elucidate specific aspects of the research question.

Research studies can be generally classified into exploratory or confirmatory studies [15]. In exploratory studies, the researcher aims to investigate a relatively new idea or area where there is not much prior knowledge to draw expectations from. For example, in a study exploring how humans would interact with robots when robots begin to enter people's home in the future, the investigators designed five likely domestic scenarios and studied how participants responded to different robot behaviours in these scenarios [116]. In confirmatory studies, also known as hypothesis testing, the researcher aims to confirm certain expectations about how their system or users would perform or behave. For example, based on the knowledge that non-verbal communications play an important role in human-human collaborations, the experimenters in [32] designed a study to confirm the importance of non-verbal communication in human-robot interaction. Focusing on confirmatory HRI studies, a key initial step is to develop clear hypotheses about the user(s) and the robot(s).

A hypothesis is a predictive statement on how a given condition will affect a certain aspect of the outcome. The hypothesis drives the selection of the type of the study and the method of evaluation [27]. It should clearly identify key constructs, be testable, and concise [164]. As an example, in an HRI study examining the effects of non-verbal communication on the efficiency in human-robot teamwork [32], the experimenters hypothesized that:

- "implicit non-verbal communication positively impacts human robot task performance with respect to understandability of the robot, efficiency of task performance, and robustness to errors that arise from miscommunication."

The above is a statement relating the given condition of communication method used, to the outcome of task performance. It names key constructs, including implicit communication and robot understandability, making a concise prediction that, by using implicit non-verbal communication, robot understandability, task efficiency, and robustness to errors from miscommunication will be improved. As the aim of user studies is to find support for the formulated hypotheses, a good hypothesis should be testable in a scientific experiment. The condition stated in the hypothesis should involve a factor, often referred to as independent variable, that the researcher can vary and control independently in an experiment. The outcome stated in the hypothesis, also known as the dependent variable, should be a parameter that is measurable or observable in an experiment setup. Considering the above example, the factor of robot communication method was controlled independently from other factors by the researchers through employing either explicit communication only, or implicit non-verbal behaviors in addition. The outcome of task performance was measured through questionnaires asking users to rate the effectiveness of the interaction, as well as video analysis. Note that constructs stated in the hypothesis become the independent and dependent variables. While formulating a hypothesis has the benefits of helping provide clarity to the focus of the research, and guiding the design of the experiment, researchers should also be careful not to be biased by the formulated hypotheses or over reliant on theory, such that they become blinded to unexpected interesting findings [164].

3.2 Building hypotheses based on HRI attributes

When formulating hypotheses, it is helpful to think from a designer's perspective as suggested by Goodrich and Schultz [76]. They defined HRI as "a field of study dedicated to understanding, designing, and evaluating robotic systems for use by or with humans". In their paper, they listed five attributes that the designer can change to affect HRI: (1) level and behaviour of autonomy, (2) nature of information exchange, (3) structure of team, (4) adaptation, learning, and training of people and robot, (5) nature of the task.

A common strategy for formulating hypotheses is then to consider which attribute is the main concern of the study. Once the main attribute of concern has been identified, the levels for testing this attribute can be determined. These attribute settings would then naturally become the independent variables in the hypotheses. The dependent variables stated in a hypothesis then often relate to what the designer hopes to achieve.

In the following paragraphs, we take a closer look at the five attributes mentioned above.

Level and Behaviour of Autonomy. Autonomy is defined as the robot's ability to accommodate variations in its environment [207] by sensing, planning and acting to reach a goal without external control [19]. Autonomy can range between manual teleoperation to full autonomy. Beer et al. have proposed a 10-point taxonomy to identify the level of robot autonomy along this spectrum [19]. This is directly related to when and how the human or robot should take the initiative in HRI. Seizing the initiative can happen in a reactive, deliberative or hybrid manner [105].

In a study related to this attribute, a hypothesis may choose the robot's level of autonomy as the independent variable and explore its effect on a range of dependent variables such as participant workload and robot perceived trustworthiness [159], which are further discussed in Sec 4.2.

Nature of Information Exchange. The nature of information exchange is primarily defined by the communication channel and communication format [76]. These can be affected by human-robot interaction distance - e.g., whether they are physically attached, in the same area, or in different areas [4]. Communication channels used for human-robot interaction often mimic those in human-human interaction. Such channels, and example formats, include visual, such as gaze [3] or body gestures [130]; auditory, such as speech [230] or audio-based expressions [68]; and haptic, such as shared load [37] or vibrations [179]. A common strategy for enabling intuitive/effective HRI is by programming robots to use communication channels and formats familiar to humans [75, 79, 148].

Considering the first example in this section [32], the attribute of main concern is the nature of information exchange. The hypotheses consider the communication method as the independent variable, comparing the combined use of implicit non-verbal behaviors and explicit communication, or the latter only. The dependent variables are robot understandability, task efficiency and robustness to errors.

Structure of Team. Structure of team concerns the number (type) of humans and robots, their roles, and the configuration of communication links among them. Directly relating to these factors, Yanco and Drury's proposed taxonomy included ratio of people to robots, composition of robot teams, level of shared interaction among teams, and human role [228]. The human may take a range of roles from supervisor and operator, to mechanic, bystander, or teammate [180]. The role of the robot can be identified by comparing HRI to human interaction with other agents, including other humans, animals and objects and defining the role of the robot based on the similarity of its role to those agents [42]. This can also be affected by the robot's morphology [228]. An anthropomorphic, zoomorphic or functional robot is more likely deployed in roles matching its shape. In social robots, the ability of the robot to deceive has been suggested as an additional role capacity categorised based on the deception object, goal, and method [188].

Structure of the team and role of team members are among independent variables that can potentially shape the basis of HRI hypotheses. Measurable dependent variables e.g. likability, trustworthiness and team performance, should then be chosen based on the study main research question(s). For example, [169] examined humans' emotional reaction to the positive/negative treatment a robot receives in videos and based its hypotheses on existing literature in psychology. The independent variables were either manipulated within the experiment (e.g. different videos) or based on user's personality traits. The outcomes (e.g., participants' emotional state and physiological arousal) were then evaluated via questionnaires or physiological measures.

Adaptation, Learning, and Training of People and Robot. Adaptation, learning, and training can happen for both human and robot in HRI. Most HRI designs aim to develop intuitive interfaces for interaction. Therefore, robotic systems for HRI are often designed to minimize the amount of user training required [76]. However, any system will have its own behaviours and limitations. Hence, there is always some human learning and adaptation taking place [77]. Short-term HRI studies commonly treat learning in humans as an undesirable carryover effect to be mitigated. However, in longitudinal studies, adaptation and learning warrant careful study and may be the foci of the studies [123]. Certain high risk application domains, such as bomb disposal, can benefit from or may require careful training of the user or operator [162].

Robot adaptation has been used to personalize interaction and promote longer term engagement with users [5, 209]. Robots can also learn tasks and skills from humans through various approaches, such as Learning from Demonstration [12]. In the context of education, robots have taken the role of teachers or peer tutors for teaching students [57], while robotic pets have also been used as therapeutic animals for treating anxiety by teaching patients to adapt their breathing to the robot's [183].

Training period and mode are examples of independent variables in a hypothesis which might affect various outcomes such as user engagement and receptiveness. For example, a study on personal space examined people's behaviour (adaptation) around robots depending on past experiences, among other factors [200]. Their four hypotheses presented clearly how these independent variables are expected to influence measurable outcomes (e.g. personal space around robots).

Nature of the Task. Introduction of new technologies may change how people fundamentally perform a task [76]. Hence, it is important to consider how people interact with the robot or perform the task when a new robot technology is introduced. Task type and task criticality (measuring importance of task outcome) [228] as well as its cognitive/physical requirements [156] are introduced as part of the HRI taxonomy. The robot's level of autonomy is an influential factor in determining the type of tasks it can perform [207]. The aspects of the task the robot should perform and the extend at which the robot can perform those tasks affect the choice of the robot and the study design [19]. A sound theoretical foundation ensures that the robot is selected based on its suitability for the task rather the expected outcome [42].

The introduction of new methods and technologies, such as kinesthetic teaching [226] and augmented reality [36], has fundamentally changed how human can carryout tasks or interact with robots. For example, in a study examining methods for a robot programming task, the hypotheses related the independent variable of task shape (e.g., teleoperation or kinesthetic) to the outcomes of interaction ease and skill performance [7].

3.3 Pre-registration of hypotheses

In certain fields, such as psychology, to improve rigour in studies, it is recommended, or common, to register the hypotheses prior to conducting the study, to ensure that the methodology of the study is justified based on hypothesis, rather than the expected outcome, to avoid bias. Researchers can register their hypotheses at websites such as the Center for Open Science (https://osf.io/prereg), AsPredicted (https://aspredicted.org), or the U.S. National Library of Medicine (https://clinicaltrials.gov). Results are then reported with reference to the initial hypotheses to avoid a selective report based on the desired outcome. Some journals (e.g., JMIR Research Protocols) may automatically generate an identifier for submitted and published research protocols or proposals. The identifier is then linked to all subsequent result papers. While registering hypotheses is not yet common practice in the field of HRI, it may eventually be adopted for the same reasons mentioned above.

3.4 Summary

Identifying the research question, marks the beginning of the experimental design process. The hypothesis should be based on sound background, and testable with an experiment. It should be deigned to shed light on certain aspects of the research question. A clear hypothesis should concisely relate the independent variables to the dependent variables. We suggest the five discussed HRI attributes as the starting point in identifying the independent variables. These variables are manipulated in the experiment to study their effects on the outcome. Once the independent variables have been identified, the set of values to be tested can then be also decided.

Formulating a hypothesis is a critical initial step in any confirmatory HRI study. The hypotheses set the ground for the next steps of the study, from experimental protocol and design of evaluation methods, to recruitment of participants, evaluation metrics, statistical analysis and reporting the outcomes. Each of these steps are discussed in detail in the following sections.

Recommendations for forming HRI study hypothesis

To summarize,

- A hypothesis is developed to allow experiments that elucidate specific aspects of the research question.
- A hypothesis is a predictive statement stating how an experimental condition(s) is expected to affect a measured outcome(s).
- A good hypothesis should be testable, concise, and name key constructs.
- When constructing experiment hypotheses for HRI studies, it is useful to consider the various design factors can be altered to affect HRI.
- To improve rigour, the experimenter may consider registering the hypotheses prior to conducing the study.

4 EXPERIMENT DESIGN AND EVALUATION

Nicole and Tina This section examines the design and implementation of an experimental study in human-robot interaction. Once the hypothesis is formulated, the next step is to determine the study design, parameters and select the appropriate evaluation methods that are suited to explore the proposed hypothesis. A non-exhaustive list of study components are presented below. Each of these components will be explored in more detail, identifying advantages, disadvantages, and considerations in the following section.

- Study design structure
- Baseline and control
- Choice of constructs and metrics
- Choice of measurements
- Trial location (laboratory, field, and online)
- Session frequency and duration (single-session or multiple follow-up sessions)
- Type and number of robot(s) (humanoid, zoomorphic, nonbiomimetic)
- Task behaviour of the robot (actions or behaviours)

4.1 Study Model

Researchers must determine the type of study design and number of groups needed to explain the hypothesis. The study design is the core of the trial. The choice of design methodology must be carefully considered, as each approach will produce different results or answer different research questions. Common study design structure in a research trial can involve one of the following: within-subjects, between-subjects, or mixed-model factorial approach [41]. This includes the consideration of a control group in the study design and will be explained in more detail below.

4.1.1 Study Design Structure.

Within-subjects Design

In a within-subjects design, there is one group of participants. Each participant is exposed to all of the experimental conditions. An experimental condition can be made up of a series of tasks, sessions or interaction patterns. An advantage of a within-subjects design is that a smaller sample size can be used, which is helpful in situations where participants are difficult to recruit [122]. Other advantages include the possibility of performing experiments in a single session sequentially (i.e., more time efficient) and resulting in smaller statistical noise (e.g., the scores of the same participant is compared between each experimental condition) [41].

However, a longer experiment duration might increase the likelihood of participants being affected by non-intended incidents, such as program glitches [117]. Long experimental duration and repeated task exposure can also lead to participants anticipating the experiment outcome, habituation, practice effect, and/or fatigue, resulting in careless task performance, or increased performance due to experience rather than the task at hand. Within subject design is further prone to cross-condition contamination, where exposure to one condition context affects the participant's response during the second condition, sometimes known as the Halo effect (e.g. [227]). Sequence effects can also occur based on the sequence of conditions presented [174]. Researchers can consider using a Latin square design and/or randomization to control the impact of habituation and other confounding variables, and incorporate necessary breaks between experiments [122].

Between-subjects Design

In between-subjects designs, participants are allocated to one group only from two or more groups. Participants are then exposed to one experimental condition. This could be one task, session or interaction pattern. The number of participant groups is determined based on the number of conditions. A strength of between-subjects design is that it minimises confounding variables such as learning effect, fatigue, and frustration that occur as the byproducts of running long experimental sessions [41, 122].

However, since participants are different between each experimental condition, collected data will be affected by individual differences without randomization (i.e., random allocation of participant groups into experimental conditions). As a result, it might be more difficult to detect significant differences and type II errors (i.e., false negatives) are more likely to occur [26, 107]. In a between-subjects design, the experimental conditions should also be distinct from each other for between-group differences to be perceived, otherwise the researcher runs the risk of detecting no differences. Furthermore, larger sample size is required to mitigate the impact of individual differences.

Mixed-model Factorial Design

A mixed-model factorial design utilizes both between-subjects and with-in subjects designs, where the between-subjects aspect is used to investigate multiple independent variables while other random effects are explored using within-subjects designs. This study design can simultaneously explore the effect from individual variables and the interaction effects Lazar et al. [122]. The downsides of each study design structure discussed above still apply and additionally, mixed-model designs often require more subjects to properly randomize the study.

4.1.2 Baseline and Control.

A control is used to demonstrate how another group or condition performs against a neutral condition. This can be important for experiments that aim to evaluate the impact of a new technology from different perspectives such as usability, functionality, and engagement level. An example includes an immediate session compared to a delayed session. Interested readers can learn more about control conditions in experimental methodology books [8, 154], and here we provide a brief overview. Commonly used control conditions include waitlist/delayed (e.g., participants receiving no robot/therapy/treatment) and active control (e.g., participants interact with a monitor instead of a robot). The control scenarios can take on many forms such as robot compared to human, animal, or conventional treatments.

Recommendations for study model

Study Design Structure.

(1) **Within-group designs** are best used when researchers want to evaluate a measure across multiple time points, or test the effects of different conditions. It is best used with limited sample size. Researchers must note the possibility of sequence effects, fatigue and practice effects.
(2) **Between-group designs** are best used when researchers want to test the difference between two conditions, behaviours or tasks, and they do not want participants to be influenced by similar conditions. The design can strongly identify effects across-groups compared to within-group. Between-subjects design requires a larger sample size to produce statistically significant results.
(3) **Mixed-Model designs** are best used when researchers are seeking to assess both between and within-group effects. Mixed-model designs require a larger sample size and more research skill to implement correctly, but produce rigorous results.

Control conditions help to provide a comparison for the research condition of interest. In HRI studies, there are many options to benchmark the robot performance. A control may not be necessary in an initial study, but later studies that are looking to substantiate the claims in more detail should consider using one.

4.2 HRI Evaluation Methodology: Constructs and Metrics

Choice of constructs (what to measure) and metrics (how to measure it) are important given that these will determine the research focus and the data to collect. Constructs and metrics will help assess the validity of the research hypothesis and the relevance of the variables controlled by the researchers, but also shed insights on the implications for the other agents involved in the interaction process or the interaction itself. We provide a summary of the common constructs and their related metrics in Table 2. Researchers can choose whichever constructs or metrics most relevant to their study, whether it is trust, acceptance, likability, response time or collaborative efficiency. However, they should carefully evaluate the quality and soundness of the chosen measurements as it affects the statistical analysis of the study [196].

4.2.1 Common Constructs in HRI Studies. HRI studies are often interested in validating theoretical concepts, known as constructs. Most of these constructs are not directly observable but can be linked to observable and/or measurable metrics and events [102]. Given the highly diverse range of applications in human-robot interaction, more than 110 different metrics have been proposed or used in the HRI literature [46]. These metrics can be grouped by the agents taking part or being measured during the interaction process, which includes the robot, the human, and the team (often referred to as the system) [143]. Given the wide scope of HRI studies, the metrics listed in this section might not apply to all types of user studies.

Robot-Related Constructs

Robot-related constructs can help evaluate the attributes in the hypothesis relating to the level and behaviour of autonomy as defined in Sec 3.2. Neglect tolerance [147] (i.e., how the robot effectiveness declines when the human is not attending the robot) and attention demand (percentage of time the human must control the robot) are often used as overall measures of a robot's **autonomy**. Beer et al. [19] suggested to also include a subjective rating of the human intervention as a supplementary measurement of a robot's level of autonomy.

Productivity is another construct that is related to the robot's performance at a given task and/or the robot's technical capabilities (e.g., accuracy at object recognition). Time-based and error metrics such as task completion time and number of unsuccessful actions are often used to assess the former [147]. Finally, the third robot-related construct is the **social attributes**, which aim to capture how traits (e.g., warmth and competence) and characteristics (e.g., anthropomorphic appearance) associated with robots affect the social perception of people interacting with them. Since these attributes mostly rely on people's judgements, subjective scales such as the Robotic Social Attributes Scale (RoSAS) [35], Robot Incentives or Robot Self-Efficacy Scale [165, 167], and Godspeed questionnaires [17] are often employed as metrics.

Human-Related Constructs

Human-related constructs are related to the attributes in the hypothesis involving adaptation, learning, and training as well as the experimental task. For instance, situation awareness and cognitive workload are two intertwined constructs that have been identified as particularly relevant in both automation and robotics [19, 193]. Situation awareness defines the awareness and understanding an individual has of a situation [175]. Cognitive workload is a product of the mental resources demanded by a task and the capacity of the person performing the task. In addition, several other constructs have been used to investigate people's perception of robots, their responses to different robots under different contexts and types of interaction [16], or even predicting future use of the robot [19]. For instance, constructs such as willingness to interact with the robot ([87, 165]), trust, perceived or physiological safety, engagement, and affective states have been shown to be potential predictors of robot use [19]. The measurement or assessment of these constructs is frequently done using questionnaires (e.g., RoSAS [35], Discrete Emotions Questionnaire [82]), physiological measurements such as changes in heart rate and skin conductivity [10] and behavioral measurements.

System-Related Constructs

System-related constructs are related to the team structure and task nature attributes of the hypothesis. Most of the constructs associated with the system are defined in the context of task-oriented applications, which aim to assess how well the human(s) and robot(s) perform as a team. For instance, while productivity and efficiency [143] are associated with how well the task is completed and the time and effort required to complete the task, fluency characterizes the coordination of joint activities and actions between members of a well-synchronized team [91, 92]. Metrics such the time elapsed between the end of an agent's action and the beginning of the other agent's action (i.e., functional delay), number of unplanned human interventions or interactions, and the time required to interact with the robot (i.e., interaction effort) are often used to measure how different experimental conditions, human factors or new robot skills affect the system's performance [46].

4.2.2 Common Measurements in HRI Studies.

Multiple constructs can be of interest in an HRI study. However, existing studies often attempt to capture the human-robot interaction experience with multiple constructs using a single measurement, which is usually insufficient Bethel et al. [26]. Choosing the right measurements often requires specific expertise about the phenomenon under investigation to bridge the knowledge gap in an interdisciplinary HRI study. For instance, facial expression alone might not be the best indicator for evaluating emotions as expressions can be culture dependent or displayed out of context (e.g., adults learn to smile to hide disappointment or awkwardness) [14]. The current section highlights the five common evaluation methods listed in Bethel et al. [26] and discusses the advantages, disadvantages, and application of each.

Task Performance Metrics

Task performance metrics are objective measurements of how well the robot, the participant, and the overall system accomplish the task under investigation. The performance can be measured with respect to five different aspects of the robot's capability, which are navigation, perception, management, manipulation, and sociability [193]. Examples of these metrics include task completion time, error rate, and efficiency. More examples can be found in Table 2.

Performance measurements are useful for comparing technologies as it gives a numerical value indicating how much improvement can be achieved by adopting the technology (e.g. introducing warehouse robots to improve the sorting efficiency of packages). However, task performance metrics alone are insufficient to capture the entire HRI experience and the result might be heavily influenced by the population. For instance, people in engineering schools with more exposure to a range of technology might have a shorter task completion time in a human-robot collaboration task versus people who are less experienced with technology. The effect of population can be reduced by selecting a large, diverse group of participants and introducing a control/baseline to measure relative changes instead of the absolute values.

Behavioural Measurements

Behavioural data can be recorded under four different study contexts:

- Naturalistic observation involves observing people's behaviour in a natural environment where it typically happens (quantitative or qualitative),
- Participant observation where researchers are part of the experiment they are conducting (quantitative or qualitative),
- Structured observation, where researchers only focus on gathering quantitative data of the behaviours under investigation in a particular experimental setup, are typically more efficient than naturalistic/participant observation as researchers will already know what to look for during the experiment,
- and finally, case studies, which are detailed investigation of an individual, social unit, or events [154].

A metric in HRI studies might include recording how long participants look at the robot. Behavioural measurements do not rely on the participant recollection or self-interpretation of their behavior [26]. However, behavioural measurements are susceptible to the Hawthorne effect, where participants might behave differently from how they normally would because they know they are being watched. Analysis of behavioural data in the form of audio/videotape requires significant time and expertise. The cost of data annotation is often too high and many researchers might not process the data. With annotation comes the problem of inter-rater and intra-rater variability. Finally, behavioural measurements alone are insufficient in explaining the participant behaviour [106] and should be used in conjunction with other measurements, such as interviews, quantitative or qualitative data.

Psychophysiological measures

Psychophysiological signals are physiological measurements that are affected by the participant's mental state. This includes heart rate, blood pressure, brain activity, skin conductance, muscle activity, and more [28]. These measurements are good for quantifying abstract concepts such as physiological arousal using non-invasive and objective measurements. The time domain aspect of these measurements helps pinpointing exactly when in the experiment a response was elicited [49]. In addition, an advantage of using psychophysiological measures compared to behaviour measures is that participants cannot easily manipulate an automatic response.

Psychophysiological measurements are informative when applied properly. Since the signal gives very specific information about the person's state (e.g. heart rate and respiration rate are indicators for arousal levels), it is important to understand what the signal is measuring and not to infer meanings that may not be accurate. One challenge of obtaining the measurement is that the signal is often influenced by various confounding variables and noise (e.g. health status, room temperature). Health conditions can prevent changes with respect to the baseline to be observed due to the

"Law of Initial Values". The law demonstrates that there might be a narrower range that the measurement can possibly increase or decrease if the initial measurement is already high or low [28]. Other confounds include orienting response, defensive response, startle response, and habituation [28]. For example, a loud noise made by the robot during the experiment could make the participant nervous, resulting in a temporary increase in heart rate that is unrelated to the actual experiment. There may be scenarios where the measurement does not provide any useful information, for example, if the task might not provide a significant change from the baseline or when different tasks might elicit the same psychophysiological readings [49].

Interviews

Interviews can take on the form of informal conversational interviews, interview guide approaches, standardized open-ended interviews, and closed quantitative interviews [106, 138]. An interview provides information that might not be collected in self-assessment and it does not force people to rate the robot on scales that are not relevant to them (e.g., how much did you trust this robot?). Qualitative responses can also be used to find new research questions and correlations between response answers and new theories. An informal interview can be tailored towards the individual and the questions would be more relevant, whereas a structured interview makes the responses more comparable. A guided interview is more comprehensive compared to an informal interview as all the topics are defined in advance. Finally, data collected in closed quantitative interviews are the easiest to analyze and compare Johnson [106].

Interviews provide a wealth of information about what were the most important factors of the interaction from the participants' perspective. However, the quality of the data is influenced by the participant's response style, which can make the interaction appear overall negative or positive, or only getting socially acceptable answers. Volunteer participants might respond differently to non-volunteers, as they might inherently be more interested in robots/new technology [149]. Data collected in informal interviews are more difficult to analyze, potentially requiring training in particular qualitative methodologies to ensure rigor (i.e., interpretive phenomonological analysis or content analysis) [58, 190]. Consistency between unstructured interviews are difficult to maintain between experiments. Qualitative data analysis can be conducted on programs such as NVIVO, AtlasTI, and other natural language packages available online [44].

Self-assessment

Self-assessments in HRI studies can be both quantitative and qualitative, and often come in the form of paper/computer based scales, questionnaires, or surveys. A list of commonly used surveys can be found in Table 2. The assessments are easy to administer, but influenced by many factors and limited by human performance. A factor that affects the accuracy of the measurement is the timing of the survey. Since surveys are often administered at the end of the study, it means the data can only be collected on a reflective level, researchers will not be able to corroborate information from the participant directly, and participants will need to remember the feelings [28, 49]. The assessment is also influenced by societal or cultural norms where participants might answer questions based on what they think is acceptable by the research team or society [28]. Some self-assessment measures require substantial fees to be paid for their use, increasing the cost of research. Others have copyright that must be approved by the owner for its use, which might be difficult to obtain.

Custom Questionnaires

Existing surveys might not always be sufficient to capture the variable at hand or the research direction of interest. Researchers can consider designing their own questionnaire, but it must be noted that proper validation requires extensive testing, statistical measurement and sampling. Adopting existing questionnaires is recommended if they have been validated. Questionnaires should not be altered once validated, as they can become invalid when the order of the items or wording is changed [172].

To create a new questionnaire, the first step is to select the appropriate measurement scale (i.e., nominal data, ordinal data, interval data, or ratio; see Table 1 for more information). Once the scale is selected, types of options and the number of options need to be considered.

Table 1. Choice of measurement scales

Type of value	Definition and example	Stats. average
Nominal data	Categorized data e.g. male participants are assigned the value of 1 and females are assigned the value of 2	Mode
Ordinal data	Ranked data e.g. participants ranking their engagement level during a HRI interaction between a discrete scale from 1 to 10	Median
Interval data	Data measured along a scale e.g. participants rating their enjoyment level on a continuous scale from 0 to 100	Mean

As a general guideline for the design of question options as stated in Rust and Golombok [172]:

- A personality/mood questionnaire often includes options such as “not at all”, “somewhat” and “very much”.
- An attitude questionnaire might have “strongly disagree”, "disagree", “neutral”, "agree", and “strongly agree” as the options.
- A clinical symptom evaluation scale might include “always”, “sometimes”, “occasionally”, “hardly ever”, and “never” as the options.

The number of options for each questionnaire item depends on the nature of the questionnaire. The important factor is to give sufficient options to the participants so that they can express their opinions. A general rule is to include at least 4 options for rating scales and use consistent number and type of options across the survey [172]. When possible, test different survey wording and choice of the variables using a pilot study to improve the quality of the survey. Johnson [106] listed a detailed procedure to construct a questionnaire in educational research and the same principles can be adopted to HRI studies. In general, researchers should note that the customised questions have not yet been tested outside the experiment, and when feasible, provide evaluation metrics for validating the survey (see Sec 4.2.3 below). For more detailed information on how to design a questionnaire, researchers should refer to Litwin and Fink [129], Rust and Golombok [172]. Examples where customised questionnaires are designed and validated for HRI experiments can be found in [56, 92].

Table 2. Summary of Constructs and related Metrics

Agent	Construct	Measurement Type	Metrics
	Autonomy [193]	Task-performance metrics	Neglect tolerance and attention demand [147]
Robot	Social Attributes [35]	Self-reporting questionnaires	RoSAS [35]
	Productivity [193]	Task-performance metrics	Task completion time, percentage of successful actions [193]

Table 2. Summary of Constructs and related Metrics

Agent	Construct	Measurement Type	Metrics
Human	Situation awareness [180]	Task-performance metrics, self-report questionnaires, physiological measurements	Situation Awareness Global Assessment Technique, Situation Awareness Rating Technique [59], gaze analysis [52], secondary task performance [175]
	Cognitive workload [193]	Self-reporting questionnaires, task-performance metrics, physiological measurements	NASA-Task Load Index (TLX), Social Avoidance and Distress Scale (SADS) [132], task errors and reaction time [152], respiratory rate, heart rate and skin conductance and temperature [146]
	Acceptance [87]	Self-reporting questionnaires	Godspeed questionnaires [17], Technology Acceptance Model based questionnaires [87], Almere Model [132]
	Trust [193]	Self-report questionanires, task-performance metrics	Human-Robot trust scale [178], trust in automation scale [104], task allocation [168], Almere Model, BEHAVE-II, Multi-Dimensional Measure of Trust, Negative Attitude toward Robots Scale (NARS) [132]
	Affective state [119]	Self-report questionnaires, physiological measurements	Discrete Emotions Questionnaire [82], Pleasure-Arousal-Dominance ratings [24], cardiovascular and electromyogram (EMG) activities [119]
	Stress [121]	Self-report questionnaires, physiological measurements	Skin potential response, semantic differential (SD) questionnaire [11], NASA-TLX, NARS, SADS [132]
	Perceived safety [121]	Self-report questionnaires, physiological and behavioral measurement	Godspeed (safety questionnaire) [17], NARS [145], distance between human and robot [142], cardio-vascular and electrodermal activity [121]
	Engagement [193]	Physiological and behavioural measurements, self-assessment	Changes in heart rate and skin conductivity changes, non-verbal gestures [10], Self Assessment Manikin Instrument [132]
System	Productivity [193]	Task-performance metrics, self-report questionnaires	Fan out and interaction effort [147], System Usability Scale (SUS), Multidimensional Robot Attitude Scale (MRAS) [132], productivity time [46]
	Fluidity[193]	Task-performance metrics	Agent idle time, concurrent activity, functional delay [91]
	Efficiency[143]	Task-performance metrics, self-reporting questionnaires	Human-robot action or time ratio, time to complete the task [193], perceived quality of interaction [13], MRAS, SUS [132]
	Reliability[193]	Task-performance metrics	False alarms, number of interventions [143]

4.2.3 *Other Considerations.*

It is important to consider the quality of a metric, which is the extent to which it can accurately (i.e., validity) and consistently (i.e., reliability) capture the construct of interest. Donmez et al. [55] identified several other factors that should be evaluated when selecting the type of measurements and the specific metrics to be included in a HRI user study. These factors are:

- a) Experimental constraints: such as the temporal and monetary resources associated to collecting and analyzing a specific metric, and the characteristics of the testing environment (e.g. gaze-based metrics make sense in controlled, in-the-lab settings but are impossible to obtain in a field context)
- b) Comprehensive understanding: how much each selected metric explains the hypothesis of interest and the amount of additional insight that can be obtained from the casual relationship between metrics (e.g. a decrease in task performance can be explained by an increase in the human-partner mental workload)
- c) Statistical efficiency: the selected metric should provide the researcher with a good number of measurements such that requirements on the statistical power needed to detect potential effects are met; and
- d) Measurement efficiency: unless otherwise required, the type of measurement used to collect a specific metric should not be intrusive or distracting to the participants and to the nature of the study, task and interaction.

Reliability

Reliability is concerned with the degree to which metrics are free from measurement error. Formally, it is defined as the proportion of variance of a metric that is attributed to the true score variance (i.e., the metric that would be obtained if there were no measurement errors) [196]. Since true score variance cannot be calculated directly [86], in practice, reliability is often described in terms of consistency over four potential sources of error (time, forms, items and raters) and involves some type of correlation computation.

Consistency over time is often referred to as *test-retest reliability* and can be measured as the correlation coefficient between the metric values obtained under similar circumstances at two different moments in time (stability). Consistency between forms is known as *parallel form reliability* and can be obtained by measuring the correlation between the metrics collected using a different form of the original instrument in an immediate retest procedure. Consistency across items, also known as *internal consistency reliability*, is the most commonly reported type of reliability and corresponds to the correlation between people's responses across the items on a multiple-item instrument. Cronbach's alpha is the most commonly used test to determine the internal consistency of an instrument [86].

When internal raters are used to collect information that serves as the metric of interest, *inter-rater reliability* can be applied to determine the extent to which different observers are consistent in their judgements. Inter-rater reliability coefficients are sensitive to consistency in terms of relative agreement, but they do not determine absolute agreement (e.g., two observers agree on the relative rank ordering of their scores/metrics but attribute different values) [196]. Kappa is frequently used to test inter-rater reliability and detailed instructions around its use can be found in [137]. In addition to the inter-rater reliability, *intra-rater reliability* measures the consistency of an observer's rating over time and is also calculated using a correlation coefficient. Researchers can capture this measurement by asking a rater to complete the same rating scale at two different time points and a correlation coefficient greater than 0.7 is considered to have good agreement [129].

Independent of reliability type, it is important to note that reliability describes the quality of a set of metrics produced by a testing instrument and not the quality of the instrument itself [196]. Similarly, reliability scores are dependent on the participant sample being measured and hence researchers should always include it when reporting their studies' results.

Validity

Validity is considered the most important quality of a measured dependent variable. It is defined as the degree to which empirical evidence, i.e., the scales, metrics and instruments employed in a study, actually measure and support the theoretical properties and constructs they are supposed to measure [74]. The general recommendation regarding construct validity is to select metrics that have been the target of extended validation (e.g. NASA TLX questionnaire) [55]. In cases where non-validated metrics are chosen, construct validity and the degree to which valid inferences can be made based on the proposed metrics are assessed by considering the following four aspects [154, 157]:

- *Face validity* is the extent to which a metric or measurement method appears to be intuitively reasonable and represent that which the researcher is attempting to measure. Since face validity is based on people's intuitions about human behavior, it is considered to be a very weak evidence of construct validity [74].
- *Content validity* relates to the extent to which a measurement reflects and covers all the content that it presumably samples. In other words, does the selected measurement encompass all aspects of the construct it was designed to measure? [86]. Content validity is often assessed using expert judgements.
- *Criterion validity* relates to the extent to which different measurements (also referred to as criteria) reflect the same construct and is measured in three ways [86]: *convergent validity*, which shows that the metric of interest is highly correlated with other criteria measuring the same variable or construct; *concurrent validity*, which indicates the extent to which the measurement or metric being evaluated is related to other criteria or some other construct measured at the same moment in time; and *predictive validity*, which assess the degree to which the metric of interest predicts things they theoretically ought to predict.
- *Discriminant validity* is evaluated by the degree to which the metric being evaluated is not correlated with measures of variables that are conceptually distinct.

Recommendations for constructs and metrics for HRI evaluation

We recommend referral to Table 2 for a quick overview of the common constructs and measurements to select what is best for each trial.

(1) **Task performance metrics** should not be used alone to substantiate claims about the entire HRI experience.
(2) **Behavioural measurements** quantify human-related constructs. When using these measurements, psychology effects should be taken into account and sufficient training should be given to the raters. Behavioural measurements can be used in conjunction with other metrics.
(3) **Psychophysiological measurements** give detailed information about the participant's internal state. They require significant expertise to ensure the correct measurements are taken and the right interpretations/conclusions are drawn. Researchers should complete relevant training and read manuals around its use before using it in trial.
(4) **Interviews** are useful when researchers are interested in developing a deeper understanding of participants' responses during the experiment. Best practice is to use a research team member not invested in the final outcome of the study to conduct the interview to avoid unconscious bias towards reporting of experiences that are more favourable towards the intended outcome. If possible, interviews should be used alongside other measurements to better quantify the experience.
(5) **Self-assessments data** are often collected through the use of surveys and questionnaires. Similar to interviews, self-assessments can also be influenced by various human factors. Whenever possible, it is better to use existing questionnaires as they have been validated.
(6) **The design of a questionnaire** involves selecting the what to measure and how to measure them. Psychometric validation is lengthy, complex and time-consuming. Researchers should carefully consider the need to create a new scale, and investigate other literature for suitability. If the scale is the first time it has been used, this should be reported as part of the measures section. Reports on survey validity and reliability should be included when reporting the scale in a research trial.

Other considerations to be reported when describing measurements:

(1) **Metric reliability** captures the quality of the data with respect to measurement errors. Test-retest reliability, internal consistency reliability, inter-rater reliability, and intra-rater reliability are the common measures that are applicable to HRI studies.
(2) **Validity** is the extent to which the data-gathering tool measures the intended measurement. Face validity, content validity, and criterion validity are three important metrics to consider when exploring metric validity in the HRI setting.

4.3 Context of HRI study

4.3.1 Trial location.

Researchers will need to select an appropriate trial location. A variety of different trial locations can be considered when conducting an experiment in human-robot interaction.

Laboratory Testing

Laboratory testing involves conducting the trial in a research laboratory or research team office space. Most studies are conducted in a laboratory setting due to factors such as ease of testing, convenient access to robotic equipment, and rigorous control and observation capabilities. Laboratory testing involves requesting participants to attend a testing session and to use a set-up that has been arranged for the participant to interact with. It is generally easier to control for unpredictable variables in laboratory studies, but observed interactions might be less natural and lack real-world relevance.

Field Testing

Field testing involves deploying the robot in a setting outside of a laboratory, such as a public setting. There is an increased trend to deploy robots into more realistic settings that better represent their final deployment use case. Field testing sites can include public spaces such as museums [151], shopping malls [108], and schools [125]), or involve non-public environments such as industrial spaces [155]. Field trials can involve many uncontrolled variables and while experiments conducted in the field better replicate real engagement patterns, subtle evaluations can be lost in noisy data.

Online Environment

Online testing involves conducting the experiment or interaction through a digital medium, such as using tele-presence, online or computer-based mediums. There is growing interest to use robots within virtual environments given the ease of deployment and capacity to engage large samples [26]. This method is described further in section 5.3.

4.3.2 *Session Frequency and Duration.*

To determine duration and frequency of testing session, it is important to consider the following parameters:

- How many conditions?
- How many sessions?
- How long is each session?
- How long is the duration of the whole experiment?
- How frequently to run the sessions?
- What timing is best for the follow-up studies?

All parameters related to duration and frequency can be determined given the hypothesis and independent variables of the experiment. In addition, researchers should run a pilot proof of concept trial with a limited number of participants to best determine the frequency and duration of the session before larger sample recruitment. An increase in the number of sessions and duration of the study can help reduce the impact of the novelty effect (i.e., perception of a new technology spikes on first encounter). Specifically, it is recommended that a study should persist for at least two consecutive months to minimize the impact of novelty effect [51, 195]. On the other hand, trials that run for a long time can see challenges with participant retention (i.e., Fernaeus et al. [67]).

4.3.3 *Type of Robot.*

Researchers should consider the choice of robot for their experiment by taking into consideration different capabilities and requirements. Examples include selecting a robot based on its appearance, mobility, processing power, and level of functionality. There is a limited number of robotic systems that are commercially available and not all the systems have the features that one might be interested to explore (e.g., facial features). Researchers often select the robot model first, choosing a robot that best fits their current constraints such as budget, followed by tests to perform or data to collect. Selection of robots should take into consideration how the target sample will identify the appropriate social behaviour from the robot. Robot appearance can affect how people perceive the robot's sense of intelligence, sociability, likability, credibility, and submissiveness, along with other attributes, [150]. A humanoid robot can be more intuitive but physical resemblance to humans might set a higher expectation in robot capabilities [149]. Zoomorphic robots that look like animals can be less intuitive, but free from the user's preconceived expectations. Decisions on what robot model to use must be made based on the intended use case. Patten and Newhart [149] provides a list of design principles if researchers desire to build a customized robot for the study.

4.3.4 Task Behaviour of the Robot.

Researchers must decide what will be the task behaviour of the robot in the experimental design. There are a variety of different options, and the role of the operator can vary depending on the experimental tasks and the capability of the robot. Robots can act autonomously, or operators can have full control over the interaction, but in some cases, they might only intervene when the robot is unable to make a decision (e.g. [108]). The operators in a WoZ study is also known as the wizard who remotely controls the robot. WoZ is commonly used when when current robotic systems are insufficient to handle a fully autonomous interaction in a safe or socially acceptable matter. The wizard can replace natural language processing or other sensing requirements, generates non-verbal behaviour, navigates, localises, and performs manipulation or classification tasks as defined by the experiment [163].

Researchers should specify the details of operator training and the duration of the training during experiment preparation, regardless of the task. In addition, the capability of the operator/wizard, and interaction scope need to be fully defined and stay consistent across experiments. For example, if a participant asks a question to the wizard that is outside of the scope of the interaction, the wizard should answer by mimicking how a machine would react (e.g., I don't know), instead of inserting additional information. Any operator error should be noted as it can affect experiment results. More information about this technique can be found in [163]. Special considerations should be given when WoZ is used, as it does not reflect how a robot can behave on its own in that instance. Therefore, the experiment answers a research question about a proposed robot system that is not yet available or possible. Other ethical considerations regarding this technique both from the operator and participant's perspectives will be explored in 5.

4.3.5 Planning for Failure and Contingency.

Experimental planning must also take into consideration a plan for the possibility of failure and contingency. Failures can happen at any time during the experiment and good trial design can help to minimise the impact of these failures. Prior to the experiment, researchers may face failure in participant recruitment, which is discussed in Sec 5.3. Common failures that occur during the experimental conduct, such as equipment failure or the researcher failed to follow the experiment protocol, can be mitigated through the use of pilot studies (see Sec 6.1) and having detailed protocols on what to do should research team members become ill, resign or withdraw without notice from the project. Finally, comprehensive data management plans should be in place to avoid any challenges related to trial evaluation, including backup equipment and data capture whenever possible to avoid data loss. Steps to mitigate failures in trial evaluation through data cleaning and analysis recommendations are also discussed in Sec 6.

Recommendations for HRI study context

(1) HRI studies can take place in the laboratory, in public, or online. There is always a trade-off between a controlled laboratory environment and the real-world and there will be differences between online versus real-life interactions. Researchers should select the location that is appropriate to the hypothesis under investigation. They can also work incrementally, by first conducting studies in the laboratory then moving towards a more realistic setting. Trials conducted in one setting should not be generalised in the abstract, introduction or discussion section across all settings (i.e., laboratory, public, and online).

(2) When possible and feasible, it is best to use robot techniques and behaviours in trial that a robot can deliver on its own to avoid deception or assessment of effects that are not yet achievable. Instead, wizard of Oz is best used if the researcher must evaluate a technique that cannot be delivered on its own for the purpose of theoretical exploration, or the researcher is using the robot as an intentional avatar for experiments.

(3) Planning for failure is an important step in experimental design due to the dynamic nature of experiments involving human participants. We have identified specific failure modes that could happen during participant recruitment, experiment, and data analysis. The best approach is to perform multiple test runs prior to the start of the experiment to minimise the risk of potential challenges and failures, by being able to identify them before participant recruitment. Researchers may choose to use a small sub sample of participants that are discarded from the final analysis (e.g., the first 3 test runs with live participants) or request other staff to complete the session without collecting their data formally in the experiment.

4.4 Examples

An example of a HRI study that focuses on a robot-related construct (i.e., social intelligence) can be found in [186]. The authors aimed to build an adaptive system allowing the robot to adjust its social behaviour based on the user's affect. To evaluate the participant's affect, the authors utilized both physiological measurements (EEG) and the participant's self-assessment data, demonstrating the need of using multiple evaluation metrics to quantify the experience.

Towards system-related constructs, Hoffman and Breazeal [92] pioneered the first fluency evaluation in HRI with a physical robot. The authors hypothesized that the human-robot collaboration fluency can be improved by replicating the anticipatory behaviour in human-human collaborations. The authors implemented a cognitive framework in a non-anthropomophic robot and designed an experiment to evaluate the framework. As part of the experiment, the authors developed and evaluated a customised survey using proper techniques as described in Section 4.2.3.

5 PARTICIPANTS

Tina and Dana

HRI research envisions robots in a variety of settings, interacting with users. A major requirement to validate proposed approaches is the participation of users during experiments. Therefore, recruiting the appropriate participant population is crucial. In this section, we first identify the types of participants involved in HRI studies and summarize the important steps in participant selection, including sampling, randomization, and blinding. We then discuss the procedure for participant recruitment, retention, and highlight various ethical considerations pertaining to the people involved in HRI studies.

5.1 Types of Participants

Participants of a HRI study can be classified into four categories: observers, WoZ operators, teleoperators, confederates, and interactors. The role of the observer might involve viewing videos of HRI interactions and interpreting the content based on HRI evaluation metrics (refer to Section 4.2.2). Wizards, or operators, are participants who operate the robot and may require special training. Confederates are actors or researchers who pretends to be experiment participants, but in reality work for the research team (see [8] for more information). The use of wizards of confederates sometimes allow researchers to measure effects that would only arise in the artificially created scenario. Finally, the interactors are the participants who interact with the robot.

5.2 Participant Selection

5.2.1 Sampling.

Sampling is the process of identifying a representative group of participants suitable for the research study. An unbiased sample where participants are randomly drawn from a population is usually difficult to achieve; hence, biased (i.e., non-random) sampling techniques are often used, such as convenience sampling and purposive sampling [61]. Convenience sampling is often used when participant recruitment is limited by practical constraints, such as participant accessibility, proximity, and availability [61]. The majority of existing HRI studies use convenience samples involving university students [16, 18]. The use of convenience samples generally reduces the external validity of the research findings as the participant population is not representative of the user population in terms of social traits, cognitive performance, and attitudes towards technology [16, 18]. Geographic proximity is often another reason for using a convenience sample, which comes with the trade-off of limiting the cultural diversity of the participants. In general, convenient samples introduce sampling bias (e.g., excluding certain groups) or sampling error (e.g., including a higher proportion of certain groups) [149]. When using convenience sampling (e.g., university students), researchers might consider balancing potential confounding variables, such as gender or educational background, during participant selection [16].

In purposive sampling, participants are deliberately chosen for certain characteristics that they possess, such as knowledge or experience that would enable them to assist with the research. There are different types of purposive sampling methods including maximum variation sampling, homogeneous sampling, typical case sampling, extreme/deviant case sampling, total population sampling and expert sampling. Specific examples can be found in Etikan et al. [61]. Sampling affects the statistical power of the study, which is determined based on the combination of the number of trials, the size of the study (number of participants), and the number of random or uncontrolled factors. A lack of statistical power can lead to both type I errors (false positives) or type II errors (false negatives) and cause the wrong conclusions to be made. Sampling more participants from a diverse population is the standard and most straightforward way to increase statistical power [53]. More discussion on the sample size can be found in Section 6.3.3.

5.2.2 Randomization.

Randomization and stratification can help assign participants to groups, conditions, and observers while minimising biases. Randomization involves assigning participants with an equal chance to experimental groups to explore the effects of the condition and prevents researchers from predicting the results of the conditions. Simple randomization is a basic form of condition assignment, which can involve methods such as a coin toss or simple computer-generated function [21]. Simple randomization can be used for sample sizes above >200 participants, given their statistical likelihood to even out in terms of allocation, but should not be used for <100 participants as there might be an unequal number of participants in each group [22, 109, 110].

Stratified randomization can be used when the characteristics of the participants need to be equalized across the experimental groups. For instance, if age is a factor that can affect the HRI experiment, researchers can stratify the participants to ensure that similar age groups are present in each participant group. However, factors relevant to the phenomenon need to be identified and normalised, which might not always be possible. There are other stratification techniques such as block randomization [54] or response adaptive randomization [88], which is a technique to adjust the condition assignment favoring conditions with better performance based on ongoing data collection. While these techniques are more common in the medicine field and less common in current HRI studies, they can be useful in longitudinal studies with socially assistive robots.

5.2.3 Blinding and Allocation Concealment.

Blinding and allocation concealment are techniques that are used to further remove bias effects in experimental condition assignment, and must be decided on in the study design [173, 182, 225]. Blinding is the process of preventing the participants, researchers or both parties from knowing what intervention or experimental content participants are allocated to. Conventional blinding strategies include single-blinded or double-blinded studies. Blinding is often not reported Hróbjartsson et al. [100] and researchers should include this information as part of the study design.

Allocation concealment is when the researcher does not know the upcoming experimental condition until the moment of assignment. Blinding reduces the effect of observer bias (i.e., the observer's judgement is influenced by the knowledge of the group), whereas allocation concealment reduces allocation bias (i.e., how participants are assigned to each group) Ruxton [173]. Both blinding and concealment can be executed cheaply. For instance, researchers studying the effect of different robot appearances can ask a researcher unrelated to the project to randomly assign letters to each robot. The mapping between the robot and the coded letter is concealed from the researchers who are analysing the data until the data has been processed. Allocation concealment can be implemented using paper-based techniques Doig and Simpson [54] or using online tools such as SEPTRE.

5.3 Participant Recruitment and Incentivization

In participant recruitment, the required sample size, the best recruitment venue, and ways of motivating participants to join the study are among the key considerations. The right sample size will depend on the study design structure and the desired statistical power as discussed in Section 6.3.3.

Researchers may face challenges in recruiting sufficient participants to take part in the study. Significant recruitment difficulties should be reviewed as soon as possible, as this may indicate that the study design needs to be re-evaluated (e.g., due to the choice of task difficulty, lack of marketing material, accessibility of the trial location). While study design may be an hidden factor, some participants are inherently more difficult to recruit, e.g., participants for long-term studies or studies that require special participant groups (e.g., patients, children). For HRI studies, participants, such as patients, are often recruited through external collaborators. For instance, in [125], the authors recruited children with handwriting difficulties through the collaborating therapist. Once the participants who are more difficult to recruit have enrolled, researchers might inquire about the participant's future availability and willingness to partake in other relevant studies. However, researchers might also need to consider when not to reuse participants. Scenarios where participants should not be reused include when the study requires first exposure to the system. Participants who have prior experience in other studies should be excluded to prevent cross-condition contamination [1].

Participants can be recruited through contacts (phone contacts, email, social media, other web-based techniques), community groups, mailing lists, volunteer banks, schools, external collaborators [122] and crowd-sourcing platforms such as Amazon Mechanical Turk [38, 39]. The techniques and the corresponding recommendations are summarized in Table 3.

Table 3. Participant recruitment methods

Method	**Advantages**	**Disadvantages**
Contacts	Participants can be recruited via the researcher's external collaborators, such as recruiting patients through collaborating therapists [125] or organizations [202]. Researcher contacts are more likely to agree to participate.	This technique might be subjected to snowball sampling, which is a biased sampling technique where one participant refers other potential participants [149]. Participants who are contacts of the researchers might be less likely to provide negative feedback.
Community groups, professional organizations	This method makes it easier to contact a large amount of specialized participants at once (good for purposive sampling).	Response rates might be low and lead to a biased sample. Researchers can consider running a short presentation at the corresponding organization to raise awareness and gauge interest.
Mailing list	This recruitment method is cheap, fast, and easy to execute.	Response rates might be low and lead to a biased sample. Cold-emailing might solicit negative perception towards the study and hence it is preferred to recruit through existing mailing lists.
Volunteer banks	Volunteer banks enable access to large number of participants.	Volunteers might be fundamentally different from non-volunteers leading to sample bias. For instance, volunteers might be more excited about technology, have prior experience with robots, or be more willing to adopt a new technology compared to non-volunteers [149].
Crowd-sourcing platforms	This method is also cheap, fast, and easy to conduct. Different filters on the platforms or sites targeted at different academic projects also enable researchers to include a more diverse population or find participants with specific skills [38]. The experimental tasks typically involve viewing pictures or videos online to replace in-person HRI [48].	In addition to issues with the data consistency and reliability, researchers might neglect the fact that crowd-sourcing workers can have prior experience with similar studies (i.e., they are non-naïve), bring their pre-assumed knowledge into the study, which in turns reduces the effect size of the study [39]. Workers might also discuss the experiment in the worker discussion boards where they might obtain foreknowledge about the experiment and be influenced by the opinions of other people [38]. Another HRI specific concern is that online surveys and passive observation may not reflect participants' perception and interaction with the robot in real life. Guidelines on crowd-sourcing platforms can be found in [204].

Regardless of the recruitment method, researchers should always pay special attention when recruiting vulnerable people. Vulnerability is not limited to disabilities but also includes instances where people might be exploited and abused due to an existing unequal or dependent relationship [191]. For instance, students participating in their supervisor's study or clinicians recruiting patients for trials. This also includes recruiting close friends or family, who may comply with the experimental manipulation to please the researcher. A simple way to minimise the effect of potential unequal relationship is through information disclosure (i.e., who is involved with the project, how privacy will be protected, how patients' right to medical care will be reserved regardless of the study results).

Motivating participants to enroll in the study might be another challenge. Different incentives might be provided, such as food for students, gifts for children, raffles, and participant honoraria. Participant compensation should be proportional to the amount of time required and the type of participants involved. However, biases might be introduced when a reward system is set up based on time, as financial incentives might cause data from a subset of participants to be over-represented in the entire sample [197], and interfere with the participants' ability to weigh the risks and benefits of the study [70]. When compensation is involved, the rule of thumb is that the payment should not be provided as a reward for the risk or harm [70]. Researchers should follow the standard procedure defined by the host research institute and maintain transparency. Ethics committees might only allow small amount of incidental expenses to be reimbursed. In addition, researchers can motivate participants by providing accommodations depending on the context of the study, such as offering to conduct the experiment at the participant's home to accommodate for the elders who might not be able to travel. Finally, incentives must be suitable and not cause additional risk for participation or unfairly encourage people to participate who are then not voluntarily providing consent, e.g., they are financially challenged and participate in the trial only to receive the incentive.

5.4 Participant Retention

Longitudinal studies are particularly susceptible to the problem of participant retention. Participants dropping out over time can result in a type of sampling bias known as mortality. In clinical research, participants can drop out of the experiment if they do not feel the benefit of the therapy, causing the remaining sample to be biased towards the groups with more effective treatment [149]. In the HRI context, mortality is observed when participants stop using the technology over time, especially in the longitudinal, in-home studies [51, 67]. The reasons for the participant's departure may not be random and it is important to record mortality in longitudinal studies and its occurrence in each group.

5.5 Participant Preparation and Training

Once the appropriate participants are recruited for the study, researchers should consider the need for participant preparation and training especially in experiments that involve observers or operators. In HRI literature, commonly the details of the training program including its length and approach are not reported [163]. Experimenters should consider the following questions when planning for participant preparation and training. A full description of participant debriefing is available in [8].

- What information does the participant need prior to the start of the experiment in order to give informed consent? Is deception involved?
- What training does the participant need to undergo?
- How will the observers be trained (e.g. what scale is being used to evaluate the HRI session and what are the evaluation standards)?

- Do the initial results present good inter-rater/intra-rater agreement? For instance, some experiments might require observers to undergo training, so that they can label affects from people's facial expressions. Researchers can administer a test prior to the start of the actual experiment to determine whether the participants had been trained sufficiently. If the initial results show poor rater agreement, researchers might consider extending the training or removing some observers.

5.6 Privacy and Ethical Considerations

This section discusses other considerations about participants from the procedural and ethical aspects of the experiment including privacy protection, informed consent, and special considerations for different experiment setups.

5.6.1 Privacy protection.

HRI studies follow guidelines similar to other research that deals with human participants. Maintaining the privacy of the participants is important as it protects them from potential harms, including embarrassment or fear of being judged by others. The basic guidelines noted by Lazar et al. [122] include:

- collect the minimum amount of information,
- limit the number of times data is used,
- aim for full information disclosure whenever possible,
- provide a secure data storage medium (especially when cloud-services are used),
- remove personal identifiers whenever possible,
- follow standard data disposal procedures recommended by the institution,
- and provide channels to address concerns and enforce accountability.

5.6.2 Informed consent.

Obtaining informed consent in HRI studies is similar to any other research involving human participants. Participants must understand the purpose, procedure and risks involved with their participation and their rights to data and information. For experiment setups that require deception (e.g., WoZ studies), informed consent cannot be obtained and researchers have to purposefully withhold information in the study. While the use of deception raises ethical concerns, it is sometimes needed especially when the phenomenon under investigation is prone to reactivity, which is when the measurement changes the participant's behaviour Price et al. [154]. It is important to devise a post-study debriefing procedure during ethics application for this type of HRI studies to fulfill the researchers' duty in information disclosure.

5.6.3 Special considerations.

In addition to being aware of the participants who might require special care, accommodation, or the region specific standards when preparing the experiment protocol, this section is dedicated to highlight potential concerns relating to other special participant groups and research practices.

As HRI is a relatively new field, many phenomena are not well understood and researchers should be aware of the potential long term impacts on the participants across all age groups. For instance, Lemaignan et al. [125] conducted a study with a social robot in a classroom setting. They pointed out the potential impact of the robot, changing the dynamics between the teachers and the students in the classroom. In addition, de Graaf [50] noted that long term human-robot interactions can potentially foster human-robot relationships, where humans start forming emotional bonds towards the technology. Considerations should be given beyond risk assessment from the safety perspective. The role and utility of the technology in the user's life should also be taken into account, especially in settings such as home, long term care, workplaces, and schools. Researchers should be aware of the ethical implications involved when developing robots with human-like qualities or high level of autonomy. These functions might fundamentally be a form of deception, as the users might wrongly believe that the robots have human characteristics when they do not.

Researchers should also be aware of the potential for participants, such as Wizard of Oz operators, to have a negative experience during the study. For instance, in Rea et al. [160], the authors reviewed social HRI experiments that put the WoZ operators in stressful situations, such as having to cheat [189], enduring verbal/physical abuse, or simulating other behaviours that would otherwise be socially awkward (e.g. long periods of silence). Researchers should minimise the duration of the necessary social conflicts in the interaction script, allow the operators to have positive interactions with the participants in post experiment debriefing, and have means for the operators to be trained and express their concerns Rea et al. [160].

COVID-19 has highlighted the importance of hygiene in every procedure including experimental studies to protect both the participant and the researcher. Experimental procedures should take into account the appropriate institution and government guideline relating to social distance and minimize contact whenever possible. As some robots cannot be sanitized using alcohol-based methods, procedures should be established between the host institution and the lab to ensure proper hygiene (e.g., have people who need to work with the robots sanitize their hands before and after working with the robot or using near UV light for equipment sterilization [114]).

5.7 Ethics application

Most academic research projects are constrained under guidelines established by different governments or institutions. For instance, the most recent guidelines for EU projects are defined in the Horizon 2020 framework and the original principals are derived from the Charter of Fundamental Rights of the European Union, and the European Convention on Human Rights. Projects in the United States are fundamentally based on the Belmont report and Hippocratic Oath. In Canada, the Tri-Agency Policies and Guidelines govern all ethical conducts for human research. In Australia, the standards are defined under the National Statement on Ethical Conduct in Human Research. While most research projects in collaboration with industry often go through the university ethics application system, some may choose to go with commercial ethics review boards. This practice is not recommended as these commercial review boards might be subjected to conflict of interests [126]. While pilot studies involving only internal researchers may often begin during development stage, any formal user studies involving recruited participants should only commence after ethics approval has been granted.

5.8 Examples

The study conducted by Moyle et al. [141] demonstrates sample randomization and stratification techniques. The authors investigated the use of PARO robots for treating the behavioural and psychological symptoms of dementia. The experiment took place in 28 different long term care facilities where the facilities were first stratified based on the organization type (private vs. nonprofit), then the experimental conditions were randomly assigned in blocks. A cluster randomized control trial was employed to minimize the effect of between-group contamination, as patients in different care facilities are inevitably exposed to different activities and treatment environments.

Lemaignan et al. [125] highlights the procedure for participant recruitment and study design in a study aiming to integrate robots in childhood education for both normal children and children with disabilities. This study demonstrated how an iterative study design can be used to address the challenges of validation with a vulnerable population group. The system was first validated with a more accessible population and then with the clinical group. Finally, the authors conducted two in-depth case studies at the lab with 2 children over the period of a month. Since the study took place under multiple experimental settings (i.e.: schools, the clinic, and the lab), the authors were able to identify differences in the children's behaviour in the lab compared to other scenarios.

The work described in [108] showcases participant recruitment in a target field environment. The authors released a semi-autonomous robot in a shopping mall, where it interacted with visitors every weekday for four hours over five weeks. During the first three weeks of the study, 332 participants were recruited by flyers that were distributed around the mall or approached by the experimenters who were onsite. At the end of the study, a questionnaire was mailed to each participant and 70% of the participants responded.

Wang et al. [216] illustrates the importance of considering cultural context during recruitment. The study explored how the cultural background of the participants (Chinese vs. US) affects the human-robot collaboration process when using two different communication styles (implicit vs. explicit). The research team was able to find representative samples by recruiting participants directly from two comparable universities in China and the US and conducting the experiments in the respective countries.

Yanco et al. [229] provides an example showing recruitment of specialised participant groups. The study evaluated HRI interfaces in-situ, developed for expert users for use during the DARPA trials. The authors used the unique event to collect data with expert users. They observed the teams participating during the trials and analyzed the HRI methodologies used for robot teleoperation.

Recommendations for participants of HRI studies

Participant selection:

(1) Balancing potential confounding variables can help with noises introduced by the participant selection process, especially when non-random sampling techniques are used.
(2) Simple randomization such as a coin toss or random number generator can be used for a large number of participants, but should not be used for a small participant group. Stratified randomization is best used for equal cross across multiple factors, such as age. Block randomization and response adaptive are more sophisticated randomization techniques and should be implemented on a needed basis.
(3) Blinding should be reported in the final report as it represents an important methodological step.Allocation concealment is an important step for demonstrating that the experiment is not influenced by the researcher's bias.

Recruitment and incentivization: Each recruitment platform comes with different advantages and trade-offs (see Table 3). Researchers need to consider the platform in relation to the participants relevant to the HRI study. In addition, suitable incentivization can be provided given that it does not affect the participant's consent.
Retention: Participant retention is important especially in longitudinal studies and participant drop-out should be noted in the final study report.
Training: Participant preparation, such as training for WoZ operators/teleoperators, should be noted in the study. In addition, researchers should devise strategies to ensure that the participants are sufficiently trained.
Privacy and ethics:

(1) Participant privacy should be maintained through data anonymization and data management.
(2) Informed consent may not be possible for WoZ studies. Appropriate procedures need to be put in place to debrief the participants after the study
(3) Given HRI is an emerging field, there are many long term consequences of the experience that are unknown. Considerations need to be given to all possible participants, from interactors such as children vulnerable groups, to the researchers participating in the study.

6 DATA COLLECTION AND DATA ANALYSIS

Once researchers have decided on the experimental design of their user study, the next step is to collect and analyze the data to test the research hypothesis. To ensure that the conclusions drawn from the analysis are valid and generalizable, researchers must carefully identify any threat to the validity and generalizability of their results (e.g., extreme metric values or insufficient statistical power, take the appropriate actions (e.g., reduce error variance), and choose an adequate statistical analysis for their study design and data. Before diving into the data analysis and statistical methods that can be used for this purpose, we first cover general recommendations on how to:

- validate the data collection process before the main study is conducted;
- identify common threats to the quality, validity and reliability of the data; and
- clean, better understand, and post-treat the data prior to any statistical analysis.

The recommendations listed in this section are common practices in fields such as psychology (e.g., [127]) and human-computer interaction (e.g., [98, 122]). The section next provides an overview of the main statistical methods found in the literature as well as some recommendations on how to best use these methods. We end this section with some examples that showcase how statistical analyses are commonly done in HRI studies.

6.1 Pilot Studies

Pilot studies are small-scale versions of a full study often conducted with one of two objectives: evaluate the feasibility of a major study or pre-test a particular research instrument or procedure. Besides helping researchers identify and correct potential deficiencies with their research design prior to the implementation of the full study, pilot studies can also help all the members of the research team to familiarize themselves with the study protocol as well as assess the appropriateness of the planned data collection and analysis techniques [212].

In the case of HRI studies, pilots are undoubtedly helpful in identifying potential problems with the primary data collection methodology. For example, Bethel and Murphy [27] were able to identify that participants often misunderstood questions relating to the dominance dimension when using the Self-Assessment Manikin (SAM) [31] to report their affective state. Based on this observation, the authors later decided to exclude participants' dominance ratings from the analysis. Pilot studies can also sometimes help identify serious omissions or mistakes in data collection [98]. This is of particular importance when psychophysiological measurements are included as part of a study since the quality and reliability of these measurements are dependent on factors such as noise, lighting, sensor placement, usage of appropriate amounts of conducting gel or paste, etc. [28, 208]. Similarly, when a new questionnaire is developed or modifications are applied to existing validated questionnaires, it is often recommended to pilot test all questionnaires. In this context, pilot studies will help researchers to test the duration of the survey as well as identify ambiguous or difficult questions [170].

Pilot studies can also help determine whether the independent variable or conditions being manipulated by the researchers work and are perceived as intended. For instance, if the aim of a study is to investigate how robots can convey an affective state through their movement, researchers should make sure that participants can clearly recognize the robot's gestures and postures [96]. Pilot studies can assist researchers in the detection of confounds, that is, factors that adversely affect the relationship between the independent and dependent factors. For example, in HRI studies, researchers should consider confound factors such as novelty, physical embodiment or the environment itself [18, 65].

It is important to keep in mind that pilot studies are smaller versions of the main study. Thus, they should closely follow the research design and protocol of the main study. For instance, the participants involved in the pilot study should be representative of the target study population and their selection should also be based on the same inclusion/exclusion criteria as the main study [203]. Similarly, pilot studies should be executed using the same administration and measurement procedures that would be used to carry out the real study [174]. Finally, the success of the main study is not guaranteed by having a pilot study. It is still possible that problems not previously encountered during the pilot might arise during the actual study [212].

Recommendation: In addition to the careful design and planning of a research study, researchers should aim to conduct a good pilot study prior to the main study. This will help them identify critical problems and deficiencies in the study protocol as well as plan and take all the corrective actions.

6.2 Common Sources of Errors and Bias in User Studies

As in the case of HCI, HRI studies are challenging because measurements of human behavior and social interactions can potentially suffer from higher fluctuations [122]. These fluctuations, often referred to as measurement errors, can arise from the testing conditions (e.g., robot's speech recognition kept failing during the interaction, which negatively impacted acceptance ratings [222]), the type of measurements and instruments (e.g., questions were ambiguous and difficult to answer), the participants themselves (e.g., participants with higher technical affinity often have more realistic expectations about robot capabilities and thus feel less anxious when interacting with robots [99]), among other factors [174].

Measurement errors can be random or systematic. While random errors are due to chance and cause measurement fluctuations in either direction, systematic errors are often due to external, predicable factors and result in measurements being biased. For example, participants might consistently under-perform during all trials of a task because of tiredness or nervousness. The impact of random errors can be reduced by increasing the observed sample size or averaging multiple measurements. When dealing with systematic errors, [122] recommend that researchers should: **1.)** aim to eliminate or control all possible sources during the study and data collection, and **2.)** isolate their impact from the main effect of interest when analyzing the data. Possible sources of systematic errors or biases in HRI studies are:

- Inappropriate, inaccurate, or incorrectly configured measurement instruments. For example, physiological measurement instruments such as EMG, EGG or IMU sensors require an accurate placement on the body [208].
- Inappropriate or unclear experimental procedures. For example, when a within-subject design is employed, the randomization of all conditions is critical. Otherwise, conditions tested later may be consistently better or worse than conditions tested earlier due to learning or fatigue effects. In the case of WoZ studies, if poorly designed, it is very likely that participants ratings on the perceived intelligence of the robot are reflections of the intelligence of the human operator rather than the robot[218].
- Characteristics of the participants may also introduce systematic errors into the results of a HRI study. As previously mentioned in Sec. 5, since many HRI studies employ university students as their main pool of participants, often from a computer science or engineering background, it is likely that the measurements will be biased by the high levels of robotics and technological familiarity of the students [16].
- Non-intended robot errors due to lack of technical robustness and functionality. Multiple studies have shown that participants' evaluation of their interaction experience with a robot can be strongly influenced by technical issues. For instance, [81] observed that robot reliability plays a substantial role in how much trust participants attribute to a robot. Similarly, [222] report on a study in which a poorly designed robot frustrated participants and hence biased their acceptance ratings.

In addition to these sources of systematic bias, there are also interaction errors and failures that are often observed when a human and a robot interact with each other. From the human perspective, Reason [161] divides human errors into slips, lapses, mistakes and violations. Slips represent the situations when the action is not what was intended (e.g., accidentally pressing the wrong button). Lapses occur as a result of lapses of user's memory and/or attention (e.g., forgetting to turn the robot off). Mistakes represent the situation when the intention is not appropriate and consequently the performed action is wrong. In other words, slips and lapses are execution failures while mistakes represent planning failures. Violations represent intentional illegitimate actions (e.g., directing the robot to run into a wall). From the robot perspective, errors can be classified into two classes [97]: 1) *technical failures* and 2) *interaction failures*. Technical failures are caused by problems in either the robot's hardware or software. Interaction failures are caused by the uncertainties in the robot interaction with the environment, other agents, and humans [192]. These errors, either from human or robot, can negatively influence the quality of the collected data. The data might be corrupted and/or missed. Hence, early detection of the possible sources of interaction errors that might happen during a HRI study is of high importance.

Recommendation: Although systematic errors cannot be completely eliminated, it is important to properly identify and control them because they represent potential threats to the validity and reliability of the data collected during a user study. Knowledge about potential sources of bias can be very helpful during data cleaning and analysis. Pilot studies can be used to identify potential biases.

6.3 Considerations Prior to Data Analysis

No matter how well designed and implemented a user study is, researchers frequently have to deal with errors and their effects on study results. As a first step in identifying these errors, researchers should employ appropriate visualization techniques to better understand and examine their data. Next, they should leverage a priori defined data cleaning strategy to make the data as free of errors as possible. Finally, researchers should take into consideration of the type of measurements included in their data, the statistical power of their design and participant sampling scheme when deciding on an appropriate statistical method for their data.

6.3.1 Data Cleaning.

Data cleaning can be described as an iterative and repeated three-stage process in which data errors and abnormalities are screened, diagnosed and edited. In most studies, data cleaning is mainly done as a post-data collection process. However, careful data monitoring and visualization during the collection process can also help to improve the quality of the data prior to the data cleaning.

Researchers must keep in mind that it is not always immediately clear whether a suspected data point is erroneous and why it is so. Similarly, missing values could be due to interruptions during data collection or unavailable information [211]. Thus, researchers should utilise their knowledge about potential technical errors (e.g., measurement errors) or previously identified systematic biases (.e.g., novelty or learning effects) and expected ranges of normal values to define the rules to follow during data cleaning. For more information about detecting and handling failures, we refer the reader to [97, 224].

Screening: A good starting point for detecting invalid data is to do a visual scan of each participant's responses and measurements as they are collected. Patterned responses, or responses completed in less time than what was observed across all participants are often employed as indicators of invalid data [127]. If instructional manipulation checks or attention check questions, that is, questions with obvious answers, were included among the measurements collected during the study, they can also be employed to identify potentially erroneous data points. This screening method is frequently used in studies in which participants were recruited through crowd sourcing services [85]. During screening, researchers should also look for lack or excess of data (e.g., missing values), outliers, and strange patterns in the data distributions.

Most of the the methods used to evaluate possible outliers are based on the idea that a particular proportion of valid data points will exist within a given k number of standard deviations from the population mean. By setting a threshold on the number of standard deviations considered as valid, data points above and below that threshold are considered to be outliers. Statistics such as a the first ($Q1$) and third ($Q3$) quartiles of the data as well as the interquartile range ($IQR = Q3 - Q1$) are also used to define outlier detection bounds. Data points outside these bounds are considered to be outliers and are thus eligible to be excluded from any posterior data analysis. In the case of a sample size that is relatively small, alternative methods such as the Grubb's test (see [127] for more information) are recommended instead. When looking for outliers, it is important to keep in mind that although all outliers are characterized as extreme data points, not all extreme data points fall into the outlier category. Similarly, researchers should be aware of erroneous inliers, that is, data points generated by error but that fall well within the expected range of values.

Diagnosis: In this phase, researchers clarify the nature of potentially problematic data points identified during data screening. Data points should be categorized as: erroneous, true extreme, true normal (i.e., prior expectations about normal data ranges were incorrect), or idiopathic (i.e., there is no explanation, but the data point is still suspected of error). For the suspected points for which diagnosis is less straightforward, i.e., they do not fall into the expected ranges of true extreme or outlier values, the application of a combination of diagnostic procedures is recommended [211]. Examples of these procedures are: determine whether data points were consistently the same throughout the whole data collection procedure or collect additional information, e.g., question the experimenter in charge of data collection, and if possible, repeat the measurement.

Editing: After the identification of errors, missing values and true values, the next step is to decide what to do with these problematic data points. Overall, researchers should decide whether problematic data points should be corrected, excluded or left unchanged. In the case of impossible values, the general recommendation is to correct them if a correct value can be found, or exclude them otherwise [127]. This editing rule also applies when missing values are due to hardware or software failures. In the case of missing data, the researcher should decide on the amount of missing data that is acceptable before leaving a participant's data out of the analysis as well as on what acceptable substitute values should be used instead (e.g., samplewide median values are often employed to substitute a missing variable or metric) [127]. Additionally, researchers should declare excluded data and the reasons for excluding this data.

Whatever the data cleaning strategy researchers employ prior to data analysis, they should provide detailed documentation of the data-cleaning methods, error types and rates, error deletion and correction rates in their study report. Similarly, researchers should report on the differences in the outcomes of their studies with and without outliers [211].

6.3.2 Data Visualization.

Simple visualizations can be made for both individual participants and the aggregated responses with the goal of seeing the trend (the direction of data progression over time), level (changes in relative value of the data over the dependent variable), and stability/variability of the data between experiment conditions and participants [120]. Four common techniques are highlighted in this section: cumulative records, semi-logarithmic charts, bar graphs and line graphs. Cumulative records are generated by summing the participants' responses across sessions. The method is useful when there is a progression to the experimental conditions (e.g. in each experiment, the robot learns to do a new skill in addition to what it can do previously) and a survey is administered at the end of each condition. This type of graph is a simple way to illustrate trends (increasing, decreasing, no change) in the total response within a study.

Semi-logarithmic charts are commonly used to display the rate of change or proportional changes in performance. These graphs are common in machine learning to track the accuracy of the algorithm over the training iterations. Both cumulative records and log charts are for continuous data, whereas bar graphs are common for discrete data or continuous data for studies with small sample sizes [120, 221]. Bar graphs are typically used for presenting summary statistics between experimental conditions (e.g. user ratings between the control condition versus the robot condition). However, summary statistics can be problematic as many distributions can lead to the same graph while other important features of the dataset are hidden [73, 221]. In the case of a small sample size, researchers should plot the data distributions or use scatter plots instead of the traditional bar graphs [220]. Weissgeber et al. Weissgerber et al. [221] implemented a free, online tool for data visualization, which can serve as a starting point for researchers when deciding between different visualizations.

6.3.3 Small Sample Size Studies.

A study's statistical power is defined as the probability of detecting a significant effect of the factor being manipulated on the dependent variables being measured if it exists. Statistical power is directly tied to the number of samples (or measurements) being analyzed. Thus, the likelihood of detecting small or even moderate effects vanishes as the sample size decreases [140]. Similarly, the likelihood of detecting an effect when there is none (Type I error) or failing to detect an effect (Type II error) can be reduced with greater statistical power. As a result, statistical analyses are more reliable when the sample size is large and the subsequent statistical power is greater [134]. The general recommendation is to perform a pre-study power analysis to estimate the appropriate number of participants and measurements required in order to achieve adequate statistical power or accurate parameter estimates [26, 149]. Tools like G*Power 3 can be used [64] for this purpose.

When data from small sample size studies still needs to be analyzed, [140] provide the following recommendations. First, for categorical data, alternative exact tests such as the Fisher's test [136] should be preferred since more common tests (e.g., the χ^2-test) are known to lack accuracy with a small sample size. Second, for continuous, interval data, it is often advised to verify key assumptions such as normality and sphericity before employing classic parametric methods such as $t-$tests or Analysis of Variance (ANOVA). However, in small sample sized data, the tests employed to verify these assumptions are likely to be under-powered and may lead to incorrect conclusions about the validity of these assumptions. In these cases, Welch's extensions to the t-test and ANOVA or non-parametric options such as the Mann-Whitney or Kruskal-Wallis tests are recommended. It is important to notice however that non-parametric tests are known to have less statistical power than their parametric counterparts. Third and last, when visualizing small sample size data, the use of common visualization methods such as histograms and box plots can be misleading and hard to interpret. Scatterplots are suggested as the best choice for showing the distribution and trends of the data in this case [220].

6.3.4 On the Type of Measurement Scales.

Putting interviews aside, most of the measurements obtained in HRI studies can be categorized into continuous, interval values (e.g., behavioral measurements such as the distance between a human and a robot or physiological data such as heart rate) or ordinal data (e.g., Likert items and scales often used on self-reporting questionnaires). While the choice of statistical methods to analyze the former is a more or less straightforward process[1], there is an ongoing debate on what is the correct way of analyzing the latter.

On the one hand, some researchers argue that ordinal data obtained using Likert scales can be treated as interval data and thus standard parametric statistical methods can be used. On the other hand, more conservative researchers choose to employ non-parametric tests instead even though it is known that these methods are less powerful and lack sensitivity to detect smaller effects [194].

Recently, Schrum et al. [181] provided a set of recommendations on the use and analysis of Likert scales in HRI studies. These recommendations are: i) employ summary statistics such as mode, median, range and skewness when reporting individual Likert items; ii) although there is compelling evidence suggesting that parametric tests such as ANOVA can be used on Likert scale data [194] as long as the appropriate assumptions have been tested and validated, a conservative approach in which non-parametric tests are used instead, is recommended; iii) analysis should be performed on a multi-item scale instead of on single Likert items.

Recommendation: Data cleaning and post-processing is an important step prior to data analysis, since the validity and generalizability of the observations and conclusions made in a study strongly depend on the quality of the data being analyzed. Researchers should formulate a set of predefined rules for dealing with data errors, and missing or

[1]The literature suggests that classic parametric tests (e.g., ANOVA) are preferred if assumptions of normality and sphericity are valid, otherwise non-parametric tests (e.g., Kruskal-Wallis) are used instead [214].

extreme values beforehand. It is also important to provide detailed documentation of all the data-cleaning methods employed, the types and sources of error identified in the data as well as the excluded data and the reasons for exclusion.

6.4 Data Analysis

This section provides an overview of the the most common statistical analyses found in the literature: hypothesis testing methods, confidence intervals and Bayesian inference methods.

6.4.1 Hypothesis Testing.

Hypothesis testing is the most common use of statistics. We usually see in the papers or reports that a *null hypothesis* is rejected or retained with a *p-value* < 5% or *p-value* < 1%. This hypothesis test is a formal approach for deciding between two interpretations of a statistical relationship in a sample. One interpretation, called the *null hypothesis* (often symbolized H_0), refers to no relationship in the target population. The other interpretation is called the *alternative hypothesis* (often symbolized H_1) which refers to the existence of relationship in the population which is reflected in the population sample as well.

Significance level is the threshold at which it is decided whether the null hypothesis should be rejected or retained. This significance value is also known as *p-value* and it is related to the probability statement made about the observed sample in the context of a hypothesis, not about the hypotheses being tested [9]. The smallest significance level that is normally considered as a reasonable evidence is 5%. In practice, if the probability of observing the *null hypothesis* statement H_0 is less than 5%, the *null hypothesis* is rejected in favor of the *alternative hypothesis* H_1. The *p-value* represents the probability of observing a similar statistical relationship if the data being analyzed was generated from random samples [112]. P-values do not give the probability of a hypothesis being true or false for this particular experiment, they only provide a description of the long term Type I error rate for a class of hypothetical experiments. Similarly, p-values do not indicate whether the means of the samples being analyzed are either equal or not equal [9].

Hypothesis testing methods can be classified into a) parametric and b) non-parametric tests with respect to the satisfaction of some assumptions. These two classes are discussed in detail in the following sections.

Standard Parametric Methods

Parametric tests are widely used in analyzing user data. These tests include t-test, analysis of variance (ANOVA), and ordinary least squares regression. Here, we focus on the ANOVA test as it is the most known and (mis)used test. The ANOVA test assesses the potential difference between two or more groups on a *continuous* measurement. Before using ANOVA, there are some assumptions that need to be met in the tested data [34]. These assumptions are as follows:

(1) Normality of data samples
(2) Homogeneity of variance (sphericity)
(3) Independence of samples

There are different tests for normality such as Chi-square, Kolmogorov-Smironov, Shapiro-Wilk, Jarque-Barre, and D'Agostino-Pearson. Some researchers argue that if the sample size of all groups are equal (balanced model) and sufficiently large, the normality assumption can be relaxed provided the samples are symmetrical [29, 66]. The homogeneity assumption can be tested by comparing variance values or through specific statistical tests such as Levene's, Fligner Killeen, and Bartlett's tests. ANOVA is not robust against unequal variance [78], thus if this assumption is violated, it may alter both the type I error, i.e., detecting an effect when there is none [84] and the statistical power of the test [144]. For the independence assumption, there is no specific test that proves its validity, but it should be accounted for in the design of the study. For instance, the observation data may be dependent if repeated measurements are collected from the same subject.

From the study design perspective (Section 4), ANOVA can be categorized based on the study model as well as the number of independent variables considered in the study [199]. In the following, we list the different types of ANOVA:

(1) **Between-subjects ANOVA** is used when examining the differences between two or more independent groups. Within this type of ANOVA there are one-way ANOVA and factorial-ANOVA. The main difference between these two branches is the number of independent variables to be tested.
   (a) **One-way ANOVA** is used when assessing the difference between groups with respect to one independent variable. In practice, one-way ANOVA is used when studying the difference between at least three groups as in the case of two groups, t-test is usually used.
   (b) **Factorial-ANOVA** is used when examining multiple independent variables.
(2) **Within-subjects ANOVA**, also called repeated measures ANOVA, is used when the same subjects were tested for each experiment condition. It is frequently used with pre- and post-test experiment design. However, it is not limited to only two time periods, it can be also used when examining the differences over two or more measurements in time. When using this type of ANOVA, it is also important to test the sphericity assumption, that is, the variance in the differences between all pairs of groups are equal, is valid [133]. This assumption can be be tested using Mauchly's test.
(3) **Mixed-model ANOVA**, also called within-between ANOVA, is used when studying the differences by group and time. That is, when the study follows both a between-subjects and within-subjects design.
(4) **Multivariate analysis of variance (MANOVA)** is used when studying the differences between groups on multiple dependent variables. We can not replace it with simply performing multiple ANOVA's for each dependent variable as ANOVA would not account for the correlations between the dependent variables.

Post-hoc Tests and Corrections

While the ANOVA test determines whether the differences observed among groups are due to chance, it does not identify which particular differences between pairs of groups are significant/not-significant. This is what *post-hoc* tests do, they are used after the ANOVA test to assess the differences between all possible group pairs (multiple comparisons) [101]. The more groups in an ANOVA test, the higher Type I error rate. Type I error rate (i.e., false positive) is defined by the significant level in the case of one comparison (two groups). The error rate inflates with the increase in the number of groups which in turn increase the number of comparisons [206]. A post-hoc test constrains the experiment-wise error rate to the significant level with an adjusted p-value. A variety of methods exist in the literature for conducting post-hoc tests. Here, we focus on the most common tests available in statistical software.

- **Bonferroni test** is simply a series of t-tests performed on all possible pairs of the tested groups. In order to limit the Type I error rate to the significant level, Benferroni test sets the significance cutoff at the significant level divided by the number of comparisons. This is called *Benferroni correction.* Thus, Benferroni test tends to be more conservative. Tanevska et al. [201] utilizes Bonferroni test after the significance was found in mixed-model ANOVA analysis. They compared people perception of iCub robot when it uses a personalized behaviour versus when it uses adaptive one in three different settings.
- **Tukey's Honest Significant Difference (HSD)** is used to measure the the pairwise differences in groups. It gives an estimate of the difference between the groups and a confidence interval for the estimate. This test is used with balanced data; the sample size is equal between groups. Akalin et al. [6] employed Tukey HSD test,

after significance was found in one-way ANOVA analysis, for comparing different feedback styles given by a social robot to older people performing arm exercises.

- **Tukey-Kramer HSD** the same as Tukey HSD except it can be used with unbalanced groups.
- **Fisher's Least Significant Difference (LSD)** is the most 'liberal' method as it has a high probability of Type I error rate. This test is computationally identical to the Bonferroni test except it does not employ any adjustment in Type I error [187]. Bethel [25] employed Fisher's LSD after significance was found in a two-way ANOVA analysis for comparing people perception of different robot's operating modes.
- **Dunnett's test** is used when we care about the difference between control group and the other groups; not considering all the pairwise comparisons. Also, it can be used with unbalanced groups as well as if the homogeneity assumption is violated. Szafir [198] employed Dunnett's test after significance was found in one-way ANOVA analysis comparing human interaction quality with a drone using different interfaces.

For more details about the post-hoc test and types, we refer the reader to [95], [89], and [34].

Non-Parametric Methods

When deviation from parametric assumptions is a concern, sample size is small or there is uncertainty about the type of distribution, non-parametric methods and modern robust statistical tests offer viable alternatives [128].

Non-parametric methods require minimal assumptions about the underlying distribution generating the observed data. This is done through the use of the ranks of the original data instead of the data itself, and the assessment of test statistical significance through randomization tests. Ranks have the advantage of not being affected by the presence of outliers or skewed distributions and allow for test statistics distributions that do not depend on assumptions such as normality. Similarly, randomization tests, in which all ways to permute the data are considered, allow to assess a test statistic significance without making explicit assumptions regarding distributions [69]. Some of the most commonly used non-parametric tests are:

- **Wilcoxon Mann Whitney Test**, also known as the non-parametric analog of the t-test. This method is a rank-based test for comparing two populations on a continuous outcome using independent samples, e.g., compare the weight of males to that of females. Abd et al. [2] employed the wilcoxon mann whitney test for comparing the users' trust towards a Baxter robot operated in different modes.
- **Wilcoxon Signed Rank Test** is the non-parametric equivalent of the paired samples t-test. This method examines whether two samples were drawn from the same population and can be used when comparing two related samples, matched samples, or repeated measurements on a single sample. Mirnig et al. [139] utilized wilcoxon signed rank test for comparing the frequencies of different social cues during human interaction with a faulty robot.
- **Kruskal-Wallis Test** is the non-parametric analog to a one-way between-subjects ANOVA. The method analyzes the population medians of the ranked form data. Harriott et al. [83] employed Kruskal-Wallis test for comparing the mental workload in human-human interaction versus in human-robot interaction scenarios.
- **Friedman's Test** is the non-parametric equivalent of the repeated-measure ANOVA. This test examines the data based on its rank properties. Hypothesis testing for the Friedman's test may be expressed by the medians or the average ranks. Walters et al. [215] used the friedman's test for comparing the user's comfort rating for different approaching techniques of a robot in a fetch and carry task.

In the context of HRI studies, [181] advocate for the use of non-parametric tests when analyzing categorical data suck as Likert scales. Similarly, recent studies seem to indicate an increase in the use of non-parametric tests. For instance, in a user study comparing the difference in the effectiveness of different learning aids (e.g., robots, tablets and traditional teachers) in the teaching of the Latin script [176], the authors opted for using non-paramterics tests after confirming that their collected data violated the normality assumption required for an ANOVA analysis. In [72] and [217], where all collected data consisted of Likert-type self-assessment data, the authors decided on the use of non-parametric tests such as the Kruskal-Wallis and Wilcoxon Mann Whitney tests.

Robust Statistical Methods

Robust methods, also known as modern methods, are defined as statistical methods particularly designed to provide adequate control of Type I error rates, increase the likelihood on discovering relevant differences between groups [60] and deal with issues such as non-symmetric distributions, outliers, differences in group variances, among others [223].

Robust methods replace traditional regression methods (i.e, ordinary least square), measures of location (i.e., the mean) and measures of association (i.e., Pearson correlation coefficients) with robust alternatives such as the sample median, trimmed means or *Kendall* and *Spearman* coefficients combined with bootstrap methods. These alternative measures can be later used to perform hypothesis testing.

Bootstrapping methods provide precise estimates of population distributions by iteratively resampling cases from a set of observed data. They can be used to obtain accurate confidence interval estimates in the presence of outliers or strongly skewed data [171]. Rank-based methods such as rank transform ANOVA-type statistic and Wilcoxon analysis offer extensions to classic non-parametric methods and are known to produce valid results when analyzing data that is non-normally distributed and/or with different variance between groups [60].

There are still some limitations associated to use of these robust alternatives. For instance, there is a reduction in the number of degrees of freedom available for statistical tests and the estimation of sample size required for sufficient statistical power is more complex [115]. We refer the reader to [60], [223], and [115] for a more detailed introduction to these robust alternative methods.

As an example of the application of robust statistical methods, in their study of how attributions and perceptions of trust are affected by the presence of intelligent artificial mediation during human-computer communication, Hohenstein et al. observed that most of their participants ratings followed a skewed distribution [94]. Instead of a standard mean as a measure of central tendency, the authors computed and reported trimmed mean estimates. Bootstrap methods were also used to compute confidence intervals.

6.4.2 Confidence Interval.

Confidence interval (CI) use the same underlying mathematical methods as the hypothesis testing, but instead of giving a probability of a single value, it gives a range of values [53]. It is powerful for showcasing the level of uncertainty around an estimate or prediction. It can be defined as a range of values, calculated by statistical methods, that includes the desired true parameter with a probability defined in advance; called the *confidence level*. The size of the confidence interval depends on the sample size, the standard deviation, and the selected level of confidence [71]. For instance, if the sample size is large, the confidence interval will be narrow. On the other side, the larger confidence level is, the wider the confidence interval. The most commonly used value for the confidence level is 95% - similar to a 5% level of statistical significance in hypothesis test.

CIs can be one or two-sided. A two-sided CI defines the population parameter from both lower and upper bounds. A one-sided CI provides either an upper or a lower limit to the population parameter. Calculation of the CI of a sample parameter takes the general form of CI = Point estimate ± Margin of error, where the margin of error is given by the product of the critical value, selected as per the required confidence limits, and the standard error of point estimate. In descriptive statistics, the CI is reported with the point estimate of the concerned parameter, indicate the reliability of the estimate. Hancock et al. [81] utilized the CI for analyzing the effects of human, robot, and environmental factors on the perceived trust in HRI. They used a two-sided CI for each factor and they found that most of the factors comparison did not include zero. This suggested that the identified relationship is consistent and substantive.

6.4.3 Bayesian Inference Methods.

In recent years, cautionary recommendations against the use of classical statistical approaches based on null-hypothesis testing have been frequently issued [213], [112], [45]. Alongside these warnings, there have been recurrent calls to shift from classic frequentist methods such as null-hypothesis testing to Bayesian inference methods. While hypothesis testing provides a $p-$value that indicates the probability of a given observation (i.e., an estimate or the difference between groups) is due to chance, Bayesian approaches provide a relative comparison of how well a null hypothesis and an alternative hypothesis account for the actual data [118]. That is, Bayesian methods allow researchers to make strong observations about the probability of a given phenomenon based on the collected evidence.

Bayesian methods offer multiple advantages such as robustness in low-power situations (e.g., small sample size) and quantification of uncertainty. Moreover, they allow for the inclusion of relevant context or domain information through the choice of informative priors, and they offer a powerful framework to build and test complex models [118]. It is important to notice however, that Bayesian methods require expertise to be used correctly since they are highly dependent on a good model specification [177]. A general recommendation is to consult with an expert especially when working with continuous data [53]. Furthermore, researchers should be aware that, compared to classical frequentist approaches, Bayesian methods also require larger sample sizes [33].

Although still new, we have started seeing the application of Bayesian Methods in the analysis of HRI studies. For instance, while analysing the relation between participants' extroversion and their tendency to view a robot as anthropomorphic, Kaplan et al. [111] performed a Bayesian regression analysis to determine the model that best predicted the participant's ratings of the robots, based on their personality variables. In [30], one-way Bayesian ANOVA was used to confirm whether the workload manipulation proposed by the authors had an effect on the participants' perceived workload as measured by the NASA-TLX questionnaire. Hamilton et al. [80] employed a Bayesian repeated measures analysis of variance to determine whether different categories of mixed-reality deictic gestures used to augment the gestural capabilities of non-humanoid robots had an effect of task-related metrics such as accuracy and reaction time.

We refer the reader to [118], [62], [131], and [63] for a detailed introduction to Bayesian statistics. A concrete and detailed example on how to use such methods can be found in [177].

6.4.4 Regression Analysis.

Sometimes the aim of a user study is not determine the validity of a hypothesis, but rather to analyze the relationship among a dependent variable of interest (e.g., a task performance metrics or the final score of a subjective questionnaire) and a number of independent variables or *predictors* (e.g., the factors being manipulated during a user study). Regression analysis is a statistical method that, under the assumption of the existence of a linear relationship between the dependent variable and the predictors, allows to: *i.)* quantify this relation, i.e., how much of the variance observed for the dependent variable is explained by the predictors, and *ii.)* make predictions of the values of the dependent variable based on the observed values of the predictors [122].

Linear regression models are constructed differently depending on the purpose of the statistical analysis [122]. If the purpose of the linear regression analysis is to find the relationship between the dependent variable and all predictors, the linear model is built by simultaneously considering all predictors. However, if the purpose of the analysis is to build a model that explains the relationship between the dependent variable and each individual predictor, a hierarchical procedure in which predictors are added to the model one at the time is more appropriate.

Linear regression models can be also used to measure the strength of the relationship between the dependent variable and the predictors, make inferences about the significance and predictive power of the predictors, and compute prediction and confidence intervals for a given a set of values of the independent predictors [40]. As with other statistical methods, for a linear regression model to be appropriate and valid, in addition to the linearity assumption previously mentioned, the following assumptions about the data being modelled must also be respected [40, 158]:

- The expected difference between the observed and the fitted values of the dependent variable must be equal to zero. This difference is also known as error.
- The variance of the errors remains constant across all values (homoscedasticity).
- None of the independent variables must be a constant or a linear combination of other predictors (colinearity).
- The errors associated with one observation should not be correlated with errors of any other observations (auto-correlation)
- The errors must follow a normal distribution (normality). This assumption is critical when sample size is small.

Violations of the first assumption can serve as indications that a current linear model has been incorrectly specified and relevant predictors are missing or irrelevant predictors should be removed. Deviations from the other assumptions (i.e., homoscedasticity, colinearity, auto-correlation, and normality) can potentially result in biased errors and regression coefficients and thus lead to incorrect inferences and misleading significance tests or confidence intervals [23].

Although often underutilized in the HRI literature compared to other behavioural research communities as human-computer interaction and affective computing, we have seen the application of linear regression models and analysis in HRI studies. Traeger et al. [210] used a multi-level regression analysis to determine how a robot's social behaviours influences the conversation dynamics between the human members of a human-robot group.

We refer the reader to [23], [40], and [158], for a detailed introduction to Regression Analysis.

Recommendation: Researchers should not blindly apply statistical hypotheses tests, such as ANOVA, seeking for statistical significance. As researchers, we can potentially fall into the trap of determining the success or failure of our studies by whether the p-values given by an ANOVA analysis are under the significance criterion of $\alpha = 0.05$. If a study is well-designed, even if the final p-value is not significant, it conveys meaningful results that suggest there is no strong evidence of the casual relation being investigated. For a well-designed experiment, researchers should pay more attention to statistical power analysis that includes the relationship between the sample size and effect size. Proper experiments should ensure that the power will be reasonably high to detect the significance with sample size calculation [113]. Despite the potential misuse, there is a better general understanding of classic parametric methods. Researchers are more likely to use them correctly and readers are more likely to understand the results better. It is important to report p-values as well as confidence intervals and effect sizes when possible [53]. Finally, it is important to include all the data necessary to reproduce your results (e.g., means, variances, number of participants, demographics, number of conditions, excluded data points, etc.) [16].

6.5 Examples

Sena et al. [185] propose a visual feedback framework in robot learning from demonstration that allows the novice user to observe robot learning progress and provide better distribution for the demonstrations over the task space. They compare between three conditions for the visual feedback methodology using a within-subjects study. They conduct power analysis prior to the data collections that indicated that for a medium effect size (Cohen's $f = 0.3$) and a type I error rate = 0.05, the required sample size is 30. Then, prior to performing a repeated measures ANOVA analysis, they checked and verified that the collected data are normally distributed dependent variable, continuous dependent variable, absence of outliers and sphericity of the data (Homogeneity of variance). They also have another study that compare between four conditions and they used mixed-factor ANOVA analysis. The power analysis indicated the required sample size is 36 or 9 per condition. However, when they check the ANOVA assumptions, they found that three out of four conditions data are not normally distributed using Anderson-Darling test. The excess Kurtosis for the four groups was 1.37, 1.60, -0.64 and 0.98, where a normal distribution would have an excess kurtosis value of zero. They chose to go with the argument about the robustness of the ANOVA for the violation of the normality assumption especially with a *minor* violation. They also tested for the homogeneity of the variance using Mauchly's test and they got $\chi^2(2) = 1.24, p = 0.54$ which indicate the validity of this assumption and no need for data correction.

On the other hand, Correia et al. [43] used non-parametric methods for analyzing their data after they found that both the normality and the homogeneity assumptions are violated. They used the Shapiro-Wilk test for checking the normality assumption and the Levene's test for checking the homogeneity assumption. They compare between two condition using the Mann-Whitney-U test. This test is similar to t-test with which a comparison between two groups can be conducted. For pairwise comparison for three conditions and more, the Kruskal–Wallis-H test can be used which is analogous to ANOVA test.

Kaplan et al. [111] used both correlation coefficient and Bayesian regression analysis for comparison between eight groups. They first conduct a pairwise correlation analysis to find how much each pair of the eight groups are correlated. Then, a Bayesian regression analysis was conducted to determine the model which best predicts the data. They used the Jeffrey's Awesome Statistical Program (JASP) for Bayesian analysis. The regression model only includes the significant predictors of the dependent variable. Any predictor variable that was not significant at the $p < 0.05$ level was excluded from the final regression model.

Recommendations for data collection and analysis of HRI studies

(1) **Piloting**: Researchers should aim to conduct a good pilot study prior to the main study. This will help them identify critical problems and deficiencies in the study protocol as well as plan and take all the corrective actions.
(2) **Biases**: Although all sources of systematic errors cannot be eliminated, it is import to properly identify and control them because they represent potential threats to the validity and reliability of the data collected during a user study. Knowledge about potential sources of bias can be also very helpful during data cleaning and analysis. Pilot studies are a good manner in which potential biases can be identified.
(3) **Quality of data**: Data cleaning and post-processing is an important step prior to data analysis, since the validity and generalizability of the observations and conclusions made in a study strongly depend on the quality of the data being analyzed. Researchers should formulate a set of predefined rules for dealing with data errors, and missing or extreme values beforehand. It is also important to provide detailed documentation of all the data-cleaning methods employed, the types and sources of error identified in the data as well as the excluded data and the reasons for exclusion.
(4) **Statistical analysis**: Researchers should not blindly apply statistical hypotheses tests, such as ANOVA, seeking for statistical significance. For a well-designed experiment, researchers should pay more attention to statistical power analysis that includes the relationship between the sample size and effect size. It is important to not only report p-values, but also, where possible, report confidence intervals and effect sizes.
(5) **Reproducibility**: It is important to include all the data necessary to reproduce your results.

7 LIMITATIONS AND FUTURE DIRECTIONS

This manuscript describes a systematic approach for designing and executing an HRI-study over its lifecycle. The main steps and applicable guidelines at each stage are provided, together with an interactive tool to serve as a guide and checklist, and facilitate study development and monitoring. For each step we also provide illustrative examples drawn from the HRI literature.

While the paper provides a comprehensive overview of current best practices, not all methodologies have been covered. For example, additional constructs and measurements, or recruitment and analysis methodologies may be available in the literature, which have not been covered in this article. Furthermore, while we have focused on the HRI (and to some extend the HCI) literature, appropriate methodologies continue to be adapted from complementary fields, e.g., psychology, neuroscience, and biomechanics. The HRI field is young and rapidly evolving, and adaptation from other fileds, where beneficial and highly relevant for HRI, should be considered in experimental design. No doubht, new methodology and technology will emerge that demonstrates that previous theories or ways of measuring phenomena are no longer accepted or conventional. Researchers should continue to monitor the HRI and related literature to remain up to date with advancements in methodology.

One example where best practice is likely to change in the future is study registration. While the importance of identification and registration of hypotheses has been acknowledged by the HRI community, it is not yet a common practice in the literature. In the near future, this might become a requirement with more publication venues recognising the value of this approach for rigour in HRI studies. Another area is the use of online studies and interaction tools, particularly given the large shift to online interaction during the Covid-19 pandemic.

Considering the majority of HRI studies to date, and as noted by previous reviews [16, 26, 93], the following guidelines are highlighted to improve future studies:

Increase awareness and use of approaches in related fields: The HRI community should closely follow and incorporate findings in related fields to improve HRI study methodology. This includes adapting theories from psychology and social sciences to develop hypotheses, constructs and measures, and incorporating innovations for measurement, data processing and data analysis from engineering, statistics and machine learning.

Increase transparency: A detailed description of the study process, including any methodology or data processing errors, should be reported in the paper to avoid misleading researchers on trial results, which can further contribute to the replication crisis as seen in other fields. Authors should consider pre-registering their hypotheses, and open-source the code and data used in their work.

Increased rigour: Authors should strive towards improved practices in recruitment, aiming for larger cohorts with appropriate demographics. Rigorous statistical methods for data analysis should be employed.

Adoption of best practices and increased rigour in the design and deployment of HRI studies will inevitably lead to better reproducability of findings and ensure that findings are relevant in application deployments, avoiding costly failures [90] and increasing the impact of the field.

ACKNOWLEDGMENTS

TBA

REFERENCES

[1] 2018. When and how you should reuse research participants. https://www.simpleusability.com/inspiration/2018/07/when-and-how-you-should-reuse-research-participants/ Library Catalog: www.simpleusability.com.

[2] Moaed A Abd, Iker Gonzalez, Mehrdad Nojoumian, and Erik D Engeberg. 2017. Trust, Satisfaction and Frustration Measurements During Human-Robot Interaction. In *30th Florida Conference on Recent Advances in RoboticsMay.* 11–12.

[3] Henny Admoni and Brian Scassellati. 2017. Social Eye Gaze in Human-Robot Interaction: A Review. *Journal of Human-Robot Interaction* 6, 1 (Mar 2017), 25. https://doi.org/10.5898/JHRI.6.1.Admoni

[4] Arvin Agah. [n.d.]. Human interactions with intelligent systems: research taxonomy. 27, 1 ([n. d.]), 71–107. https://doi.org/10.1016/S0045-7906(00)00009-4

[5] Muneeb Imtiaz Ahmad, Omar Mubin, and Joanne Orlando. 2017. Adaptive Social Robot for Sustaining Social Engagement during Long-Term Children–Robot Interaction. *International Journal of Human–Computer Interaction* 33, 12 (Dec 2017), 943–962. https://doi.org/10.1080/10447318.2017.1300750

[6] Neziha Akalin, Annica Kristoffersson, and Amy Loutfi. 2019. The Influence of Feedback Type in Robot-Assisted Training. *Multimodal Technologies and Interaction* 3, 4 (2019), 67.

[7] Baris Akgun and Kaushik Subramanian. [n.d.]. Robot Learning from Demonstration: Kinesthetic Teaching vs. Teleoperation. ([n. d.]), 7.

[8] Mike Allen. 2017. *The SAGE encyclopedia of communication research methods.* Sage Publications.

[9] Naomi Altman and Martin Krzywinski. 2017. Points of significance: interpreting P values.

[10] Salvatore M Anzalone, Sofiane Boucenna, Serena Ivaldi, and Mohamed Chetouani. 2015. Evaluating the engagement with social robots. *International Journal of Social Robotics* 7, 4 (2015), 465–478.

[11] Tamio Arai, Ryu Kato, and Marina Fujita. 2010. Assessment of operator stress induced by robot collaboration in assembly. *CIRP annals* 59, 1 (2010), 5–8.

[12] Brenna D. Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. 2009. A survey of robot learning from demonstration. *Robotics and Autonomous Systems* 57, 5 (May 2009), 469–483. https://doi.org/10.1016/j.robot.2008.10.024

[13] Jimmy Baraglia, Maya Cakmak, Yukie Nagai, Rajesh PN Rao, and Minoru Asada. 2017. Efficient human-robot collaboration: When should a robot take initiative? *The International Journal of Robotics Research* 36, 5-7 (2017), 563–579. https://doi.org/10.1177/0278364916688253 arXiv:https://doi.org/10.1177/0278364916688253

[14] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest* 20, 1 (2019), 1–68.
[15] Belpaeme T. Eyssel F. Kanda T. Keijsers M. & Sabanovic S. Bartneck, C. 2020. *Human-Robot Interaction – An Introduction.* Cambridge: Cambridge University Press.
[16] Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Šabanović. 2020. *Human-Robot Interaction: An Introduction.* Cambridge University Press.
[17] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1, 1 (2009), 71–81.
[18] Paul Baxter, James Kennedy, Emmanuel Senft, Severin Lemaignan, and Tony Belpaeme. 2016. From characterising three years of HRI to methodology and reporting recommendations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 391–398.
[19] Jenay M Beer, Arthur D Fisk, and Wendy A Rogers. 2014. Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction. *Journal of Human-Robot Interaction* 3, 2 (Jun 2014), 74. https://doi.org/10.5898/JHRI.3.2.Beer
[20] Kathleen Belhassein, Guilhem Buisan, Aurélie Clodic, and Rachid Alami. 2019. Towards methodological principles for user studies in Human-Robot Interaction. In *Test Methods and Metrics for Effective HRI in Collaborative Human-Robot Teams Workshop, ACM/IEEE International Conference on Human-Robot Interaction.*
[21] Elaine M Beller, Val Gebski, and Anthony C Keech. 2002. Randomisation in clinical trials. *Medical Journal of Australia* 177, 10 (2002), 565–567.
[22] Elaine M Beller, Val Gebski, and Anthony C Keech. 2002. Randomisation in clinical trials. *Medical Journal of Australia* 177, 10 (2002), 565–567.
[23] Henning Best and Christof Wolf. 2014. *The SAGE handbook of regression analysis and causal inference.* Sage.
[24] Alberto Betella and Paul FMJ Verschure. 2016. The affective slider: A digital self-assessment scale for the measurement of human emotions. *PloS one* 11, 2 (2016).
[25] Cindy L Bethel. 2009. Robots without faces: non-verbal social human-robot interaction. (2009).
[26] Cindy L Bethel, Zachary Henkel, and Kenna Baugus. 2020. Conducting Studies in Human-Robot Interaction. In *Human-Robot Interaction.* Springer, 91–124.
[27] Cindy L Bethel and Robin R Murphy. 2010. Review of human studies methods in HRI and recommendations. *International Journal of Social Robotics* 2, 4 (2010), 347–359.
[28] Cindy L Bethel, Kristen Salomon, Robin R Murphy, and Jennifer L Burke. 2007. Survey of psychophysiology measurements applied to human-robot interaction. In *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication.* IEEE, 732–737.
[29] María J Blanca, Rafael Alarcón, Jaume Arnau, Roser Bono, and Rebecca Bendayan. 2017. Non-normal data: Is ANOVA still a valid option? *Psicothema* 29, 4 (2017), 552–557.
[30] Indu P Bodala, Bing Cai Kok, Weicong Sng, and Harold Soh. 2020. Modeling the Interplay of Trust and Attention in HRI: An Autonomous Vehicle Study. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction.* 145–147.
[31] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59.
[32] C. Breazeal, C.D. Kidd, A.L. Thomaz, G. Hoffman, and M. Berlin. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems.* 708–713. https://doi.org/10.1109/IROS.2005.1545011
[33] Marc Brysbaert. 2019. How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition* 2, 1 (2019).
[34] Rudolf N Cardinal and Michael RF Aitken. 2013. *ANOVA for the behavioral sciences researcher.* Psychology Press.
[35] Colleen M Carpinella, Alisa B Wyman, Michael A Perez, and Steven J Stroessner. 2017. The Robotic Social Attributes Scale (RoSAS) Development and Validation. In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction.* 254–262.
[36] Wesley P. Chan, Geoffrey Hanks, Maram Sakr, Tiger Zuo, H. F. Machiel Van der Loos, and Elizabeth A. Croft. 2020. An Augmented Reality Human-Robot Physical Collaboration Interface Design for Shared, Large-Scale, Labour-Intensive Manufacturing Tasks. *IROS* (2020).
[37] Wesley P Chan, Chris AC Parker, HF Machiel Van der Loos, and Elizabeth A Croft. 2013. A human-inspired object handover controller. *The International Journal of Robotics Research* 32, 8 (Jul 2013), 971–983. https://doi.org/10.1177/0278364913488806
[38] Jesse Chandler, Pam Mueller, and Gabriele Paolacci. 2014. Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior research methods* 46, 1 (2014), 112–130.
[39] Jesse Chandler, Gabriele Paolacci, Eyal Peer, Pam Mueller, and Kate Ratliff. 2015. Non-Naïve Participants Can Reduce Effect Sizes. *ACR North American Advances* NA-43 (2015). https://www.acrwebsite.org/volumes/1020052/volumes/v43/NA-43
[40] Samprit Chatterjee and Jeffrey S. Simonoff. 2012. *Handbook of Regression Analysis.* Wiley.
[41] Larry B. Christensen. 2007. *Experimental methodology* (10th ed. ed.). Pearson/Allyn & Bacon, Boston.
[42] Emily C. Collins. 2019. Drawing parallels in human-other interactions: a trans-disciplinary approach to developing human-robot interaction methodologies. 374, 1771 (2019), 20180433. https://doi.org/10.1098/rstb.2018.0433
[43] Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco S Melo, and Ana Paiva. 2018. Exploring the impact of fault justification in human-robot trust. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems.* 507–513.
[44] António Pedro Costa, Francislê Neri de Sousa, António Moreira, and Dayse Neri de Souza. 2017. Research through design: qualitative analysis to evaluate the usability. In *Computer supported qualitative research.* Springer, 1–12.

[45] Geoff Cumming. 2014. The new statistics: Why and how. *Psychological science* 25, 1 (2014), 7–29.

[46] Praveen Damacharla, Ahmad Y Javaid, Jennie J Gallimore, and Vijay K Devabhaktuni. 2018. Common metrics to benchmark human-machine teams (HMT): A review. *IEEE Access* 6 (2018), 38637–38655.

[47] Kerstin Dautenhahn. 2007. Methodology & themes of human-robot interaction: A growing research field. *International Journal of Advanced Robotic Systems* 4, 1 (2007), 15.

[48] Kerstin Dautenhahn. 2018. Some brief thoughts on the past and future of human-robot interaction.

[49] Kerstin Dautenhahn and Joe Saunders. 2011. *New frontiers in human robot interaction.* Vol. 2. John Benjamins Publishing.

[50] Maartje MA de Graaf. 2016. An ethical evaluation of human–robot relationships. *International journal of social robotics* 8, 4 (2016), 589–598.

[51] Maartje MA de Graaf, Somaya Ben Allouch, and Jan AGM van Dijk. 2016. Long-term evaluation of a social robot in real homes. *Interaction studies* 17, 3 (2016), 462–491.

[52] A. Dini, C. Murko, S. Yahyanejad, U. Augsdörfer, M. Hofbaur, and L. Paletta. 2017. Measurement and prediction of situation awareness in human-robot interaction based on a framework of probabilistic attention. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 4354–4361.

[53] A. Dix. 2020. *Statistics for HCI: Making Sense of Quantitative Data*. Morgan & Claypool. https://ieeexplore-ieee-org.ezproxy.lib.monash.edu.au/document/9070780

[54] Gordon S Doig and Fiona Simpson. 2005. Randomization and allocation concealment: a practical guide for researchers. *Journal of Critical Care* 20, 2 (2005), 187–191.

[55] Birsen Donmez, Patricia E Pina, and Mary L Cummings. 2009. Evaluation criteria for human-automation performance metrics. In *Performance evaluation and benchmarking of intelligent systems*. Springer, 21–40.

[56] Anca D Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S Srinivasa. 2015. Effects of robot motion on human-robot collaboration. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 51–58.

[57] Bosede I. Edwards, Idris O. Muniru, Nosiba Khougali, Adrian D. Cheok, and Rui Prada. 2018. A Physically Embodied Robot Teacher (PERT) as a Facilitator for Peer Learning. In *2018 IEEE Frontiers in Education Conference (FIE)*. 1–9. https://doi.org/10.1109/FIE.2018.8658445

[58] Satu Elo and Helvi Kyngäs. 2008. The qualitative content analysis process. *Journal of advanced nursing* 62, 1 (2008), 107–115.

[59] Mica R Endsley, Stephen J Selcon, Thomas D Hardiman, and Darryl G Croft. 1998. A comparative analysis of SAGAT and SART for evaluations of situation awareness. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 42. SAGE Publications Sage CA: Los Angeles, CA, 82–86.

[60] David M Erceg-Hurn and Vikki M Mirosevich. 2008. Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. *American Psychologist* 63, 7 (2008), 591.

[61] Ilker Etikan, Sulaiman Abubakar Musa, and Rukayya Sunusi Alkassim. 2016. Comparison of convenience sampling and purposive sampling. *American journal of theoretical and applied statistics* 5, 1 (2016), 1–4.

[62] Alexander Etz, Quentin F Gronau, Fabian Dablander, Peter A Edelsbrunner, and Beth Baribault. 2018. How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review* 25, 1 (2018), 219–234.

[63] Alexander Etz and Joachim Vandekerckhove. 2018. Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review* 25, 1 (2018), 5–34.

[64] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.

[65] David Feil-Seifer and Maja J Mataric. [n.d.]. Human RobotInteraction. ([n. d.]).

[66] Betty J Feir-Walsh and Larry E Toothaker. 1974. An empirical comparison of the ANOVA F-test, normal scores test and Kruskal-Wallis test under violation of assumptions. *Educational and Psychological Measurement* 34, 4 (1974), 789–799.

[67] Ylva Fernaeus, Maria Håkansson, Mattias Jacobsson, and Sara Ljungblad. 2010. How do you play with a robotic toy animal? A long-term study of Pleo. In *Proceedings of the 9th international Conference on interaction Design and Children*. 39–48.

[68] Morten Roed Frederiksen and Kasper Stoey. 2019. Augmenting the audio-based expression modality of a non-affective robot. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 144–149. https://doi.org/10.1109/ACII.2019.8925510

[69] Rudolf Freund, Donna Mohr, and William Wilson. 2010. *Statistical Methods*. Elsevier.

[70] Craig L Fry, Wayne Hall, Alison Ritter, and Rebecca Jenkinson. 2006. The ethics of paying drug users who participate in research: A review and practical recommendations. *Journal of Empirical Research on Human Research Ethics* 1, 4 (2006), 21–35.

[71] Martin J Gardner and Douglas G Altman. 1986. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J (Clin Res Ed)* 292, 6522 (1986), 746–750.

[72] Ilaria Gaudiello, Elisabetta Zibetti, Sébastien Lefort, Mohamed Chetouani, and Serena Ivaldi. 2016. Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers. *Computers in Human Behavior* 61 (2016), 633–655.

[73] Christopher H George, S Clare Stanford, Steve Alexander, Giuseppe Cirino, James R Docherty, Mark A Giembycz, Daniel Hoyer, Paul A Insel, Angelo A Izzo, Yong Ji, et al. 2017. Updating the guidelines for data transparency in the British Journal of Pharmacology–data sharing and the use of scatter plots instead of bar charts. *British journal of pharmacology* 174, 17 (2017), 2801.

[74] Darren Gergle and Desney S Tan. 2014. Experimental research in HCI. In *Ways of Knowing in HCI*. Springer, 191–227.

[75] Michael J. Gielniak and Andrea L. Thomaz. 2012. Enhancing interaction through exaggerated motion synthesis. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 375–382. https://doi.org/10.1145/2157689.2157813

[76] Michael Goodrich and Alan Schultz. 2007. Human-Robot Interaction: A Survey. *Foundations and Trends in Human-Computer Interaction* 1 (01 2007), 203–275. https://doi.org/10.1561/1100000005

[77] A. Green and K.S. Eklundh. 2003. Designing for learnability in human-robot communication. *IEEE Transactions on Industrial Electronics* 50, 4 (Aug 2003), 644–650. https://doi.org/10.1109/TIE.2003.814763

[78] Robert J Grissom. 2000. Heterogeneity of variance in clinical data. *Journal of consulting and clinical psychology* 68, 1 (2000), 155.

[79] Amir Haddadi, Elizabeth A. Croft, Brian T. Gleeson, Karon MacLean, and Javier Alcazar. 2013. Analysis of task-based gestures in human-robot interaction. In *2013 IEEE International Conference on Robotics and Automation*. 2146–2152. https://doi.org/10.1109/ICRA.2013.6630865

[80] Jared Hamilton, Nhan Tran, and Tom Williams. 2020. Tradeoffs Between Effectiveness and Social Perception When Using Mixed Reality to Supplement Gesturally Limited Robots. In *Proceedings of the 3rd International Workshop on Virtual, Augmented, and Mixed Reality for HRI*.

[81] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[82] Cindy Harmon-Jones, Brock Bastian, and Eddie Harmon-Jones. 2016. The discrete emotions questionnaire: A new tool for measuring state self-reported emotions. *PloS one* 11, 8 (2016).

[83] Caroline E Harriott, Tao Zhang, Glenna L Buford, and Julie A Adams. 2015. Measuring human workload in a collaborative human-robot team. *Journal of Human-Robot Interaction* 4, 2 (2015), 61–96.

[84] Michael R Harwell, Elaine N Rubinstein, William S Hayes, and Corley C Olds. 1992. Summarizing Monte Carlo results in methodological research: The one-and two-factor fixed effects ANOVA cases. *Journal of educational statistics* 17, 4 (1992), 315–339.

[85] David J Hauser and Norbert Schwarz. 2016. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods* 48, 1 (2016), 400–407.

[86] Roberta Heale and Alison Twycross. 2015. Validity and reliability in quantitative studies. *Evidence-based nursing* 18, 3 (2015), 66–67.

[87] Marcel Heerink, Ben Krose, Vanessa Evers, and Bob Wielinga. 2009. Measuring acceptance of an assistive social robot: a suggested toolkit. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 528–533.

[88] Trevor Higgins. 2003. Randomization in Clinical Trials: Theory and Practice. *Current Medical Research and Opinion* 19, 1 (2003), 67. http://search.proquest.com/docview/207967010/

[89] Anthony Hilton and Richard A Armstrong. 2006. Statnote 6: post-hoc ANOVA tests. *Microbiologist* 2006 (2006), 34–36.

[90] Guy Hoffman. 2019. Anki, Jibo, and Kuri: What We Can Learn from Social Robots That Didn't Make It. *IEEE Spectrum* (2019).

[91] Guy Hoffman. 2019. Evaluating fluency in human–robot collaboration. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 209–218.

[92] Guy Hoffman and Cynthia Breazeal. 2007. Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*. 1–8.

[93] Guy Hoffman and Xuan Zhao. 2020. A Primer for Conducting Experiments in Human–Robot Interaction. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 1 (2020), 1–31.

[94] Jess Hohenstein and Malte Jung. 2020. AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior* 106 (2020), 106190.

[95] Susan R Homack. 2001. Understanding What ANOVA Post Hoc Tests Are, Really. (2001).

[96] Nadine Homburg. 2018. How to Include Humanoid Robots into Experimental Research: A Multi-Step Approach. In *Proceedings of the 51st Hawaii International Conference on System Sciences*.

[97] Shanee Honig and Tal Oron-Gilad. 2018. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in psychology* 9 (2018), 861.

[98] Kasper Hornbæk. 2013. Some whys and hows of experiments in human–computer interaction. *Foundations and Trends in Human-Computer Interaction* 5, 4 (2013), 299–373.

[99] Aike C Horstmann and Nicole C Krämer. 2019. Great expectations? Relation of previous experiences with social robots in real life or in the media and expectancies based on qualitative and quantitative assessment. *Frontiers in psychology* 10 (2019), 939.

[100] Asbjørn Hróbjartsson, Julie Pildal, An-Wen Chan, Mette T Haahr, Douglas G Altman, and Peter C Gøtzsche. 2009. Reporting on blinding in trial protocols and corresponding publications was often inadequate but rarely contradictory. *Journal of clinical epidemiology* 62, 9 (2009), 967–973.

[101] Jason Hsu. 1996. *Multiple comparisons: theory and methods*. CRC Press.

[102] Michael Hyland. 1981. *The Nature of Hypothetical Constructs*. Macmillan Education UK, London, 42–58.

[103] Bahar Irfan, James Kennedy, Séverin Lemaignan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2018. Social psychology and human-robot interaction: An uneasy marriage. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 13–20.

[104] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics* 4, 1 (2000), 53–71.

[105] Shu Jiang and Ronald C. Arkin. [n.d.]. Mixed-Initiative Human-Robot Interaction: Definition, Taxonomy, and Survey. In *2015 IEEE International Conference on Systems, Man, and Cybernetics* (2015-10). 954–961. https://doi.org/10.1109/SMC.2015.174

[106] Robert Burke Johnson; Larry B. Christensen; Burke Johnson. 2016. *Educational Research: Quantitative, Qualitative, and Mixed Approaches* (hardcover ed.). Sage Publications, Inc.

[107] James M Johnston and Henry S Pennypacker. 2010. *Strategies and tactics of behavioral research.* Routledge.

[108] Takayuki Kanda, Masahiro Shiomi, Zenta Miyashita, Hiroshi Ishiguro, and Norihiro Hagita. 2009. An affective guide robot in a shopping mall. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 173–180.

[109] Minsoo Kang, Brian G Ragan, and Jae-Hyeon Park. 2008. Issues in outcomes research: an overview of randomization techniques for clinical trials. *Journal of athletic training* 43, 2 (2008), 215–221.

[110] PhD. Kang, Minsoo, PhD. A.T.C. Ragan, Brian G, and PhD. Park, Jae-Hyeon. 2008. Issues in Outcomes Research: An Overview of Randomization Techniques for Clinical Trials. *Journal of Athletic Training* 43, 2 (Mar 2008), 215–21. https://search.proquest.com/docview/206648600?accountid=12528 Copyright - Copyright National Athletic Trainers Association Mar/Apr 2008; Document feature - Illustrations; Tables; Charts; ; Last updated - 2017-11-09.

[111] Alexandra D Kaplan, Tracy Sanders, and Peter A Hancock. 2019. The Relationship Between Extroversion and the Tendency to Anthropomorphize Robots: A Bayesian Analysis. *Frontiers in Robotics and AI* 5 (2019), 135.

[112] Maurits Kaptein and Judy Robertson. 2012. Rethinking statistical analysis methods for CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1105–1114.

[113] Sung-Hee Kim, Ji Soo Yi, and Niklas Elmqvist. 2014. Oopsy-daisy: failure stories in quantitative evaluation studies for visualizations. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*. 142–146.

[114] Hiroki Kitagawa, Toshihito Nomura, Tanuza Nazmul, Keitaro Omori, Norifumi Shigemoto, Takemasa Sakaguchi, and Hiroki Ohge. 2020. Effectiveness of 222-nm ultraviolet light on disinfecting SARS-CoV-2 surface contamination. *American journal of infection control* (2020).

[115] Barbara Kitchenham, Lech Madeyski, David Budgen, Jacky Keung, Pearl Brereton, Stuart Charters, Shirley Gibbs, and Amnart Pohthong. 2017. Robust statistical methods for empirical software engineering. *Empirical Software Engineering* 22, 2 (2017), 579–630.

[116] Kheng Koay, Dag Sverre Syrdal, Michael Walters, and Kerstin Dautenhahn. 2009. Five Weeks in the Robot House - Exploratory Human-Robot Interaction Trials in a Domestic Setting. 219–226. https://doi.org/10.1109/ACHI.2009.62

[117] Joachim Krauth. 2000. *Experimental design: a handbook and dictionary for medical and behavioral research*. Vol. 14. Elsevier.

[118] John K Kruschke and Torrin M Liddell. 2018. The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review* 25, 1 (2018), 178–206.

[119] D. Kulic and E. A. Croft. 2007. Affective State Estimation for Human–Robot Interaction. *IEEE Transactions on Robotics* 23, 5 (2007), 991–1000.

[120] Justin D Lane and David L Gast. 2014. Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological rehabilitation* 24, 3-4 (2014), 445–463.

[121] Przemyslaw A. Lasota, Terrence Fong, and Julie A. Shah. [n.d.]. A Survey of Methods for Safe Human-Robot Interaction. 5, 4 ([n. d.]), 261–349. https://doi.org/10.1561/2300000052 Publisher: Now Publishers, Inc.

[122] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction.* Morgan Kaufmann.

[123] Kheng Lee Koay, Dag Sverre Syrdal, Michael L. Walters, and Kerstin Dautenhahn. 2007. Living with Robots: Investigating the Habituation Effect in Participants' Preferences During a Longitudinal Human-Robot Interaction Study. In *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication*. 564–569. https://doi.org/10.1109/ROMAN.2007.4415149

[124] Benedikt Leichtmann and Verena Nitsch. 2020. How much distance do humans keep toward robots? Literature review, meta-analysis, and theoretical considerations on personal space in human-robot interaction. *Journal of Environmental Psychology* 68 (2020), 101386. https://doi.org/10.1016/j.jenvp.2019.101386

[125] Séverin Lemaignan, Alexis Jacq, Deanna Hood, Fernando Garcia, Ana Paiva, and Pierre Dillenbourg. 2016. Learning by teaching a robot: The case of handwriting. *IEEE Robotics & Automation Magazine* 23, 2 (2016), 56–66.

[126] Trudo Lemmens and Benjamin Freedman. 2000. Ethics review for sale? Conflict of interest and commercial research review boards. *The Milbank Quarterly* 78, 4 (2000), 547–584.

[127] Frederick TL Leong and James T Austin. 2006. *The psychology research handbook: A guide for graduate students and research assistants.* Sage.

[128] Todd D. Little, Trent D. Buskirk, Lisa M. Willoughby, and Terry T. Tomazic. 2013. Nonparametric Statistical Techniques. In *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2: Statistical Analysis*. Oxford University Press, 1–68.

[129] Mark S Litwin and Arlene Fink. 2003. *How to assess and interpret survey psychometrics*. Vol. 8. Sage.

[130] Hongyi Liu and Lihui Wang. 2018. Gesture recognition for human-robot collaboration: A review. *International Journal of Industrial Ergonomics* 68 (Nov 2018), 355–367. https://doi.org/10.1016/j.ergon.2017.02.004

[131] Alexander Ly, Akash Raj, Alexander Etz, Maarten Marsman, Quentin F Gronau, and Eric-Jan Wagenmakers. 2018. Bayesian reanalyses from summary statistics: A guide for academic consumers. *Advances in Methods and Practices in Psychological Science* 1, 3 (2018), 367–374.

[132] Jeremy A Marvel, Shelly Bagchi, Megan Zimmerman, and Brian Antonishek. 2020. Towards Effective Interface Designs for Collaborative HRI in Manufacturing: Metrics and Measures. *ACM Transactions on Human-Robot Interaction (THRI)* 9, 4 (2020), 1–55.

[133] John W Mauchly. 1940. Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics* 11, 2 (1940), 204–209.

[134] Scott E Maxwell, Ken Kelley, and Joseph R Rausch. 2008. Sample size planning for statistical power and accuracy in parameter estimation. *Annual review of psychology* 59 (2008).

[135] Scott E Maxwell, Michael Y Lau, and George S Howard. 2015. Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist* 70, 6 (2015), 487.

[136] John H McDonald. 2009. *Handbook of biological statistics*. Vol. 2.

[137] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica* 22, 3 (2012), 276–282.

[138] Jane Mills and Melanie Birks. 2014. *Qualitative methodology: A practical guide*. Sage.

[139] Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. 2017. To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI* 4 (2017), 21.

[140] Charity J Morgan. 2017. Use of proper statistical techniques for research studies with small samples. *American Journal of Physiology-Lung Cellular and Molecular Physiology* 313, 5 (2017), L873–L877.

[141] Wendy Moyle, Cindy J Jones, Jenny E Murfield, Lukman Thalib, Elizabeth RA Beattie, David KH Shum, Siobhan T O'Dwyer, M Cindy Mervin, and Brian M Draper. 2017. Use of a robotic seal as a therapeutic tool to improve dementia symptoms: a cluster-randomized controlled trial. *Journal of the American Medical Directors Association* 18, 9 (2017), 766–773.

[142] Jonathan Mumm and Bilge Mutlu. 2011. Human-robot proxemics: physical and psychological distancing in human-robot interaction. In *Proceedings of the 6th international conference on Human-robot interaction*. 331–338.

[143] Robin R Murphy and Debra Schreckenghost. 2013. Survey of metrics for human-robot interaction. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 197–198.

[144] Kim F Nimon. 2012. Statistical assumptions of substantive analyses across the general linear model: a mini-review. *Frontiers in psychology* 3 (2012), 322.

[145] Tatsuya Nomura, Tomohiro Suzuki, Takayuki Kanda, and Kensuke Kato. 2006. Measurement of negative attitudes toward robots. *Interaction Studies* 7, 3 (2006), 437–454.

[146] Domen Novak, Matjaž Mihelj, and Marko Munih. 2011. Psychophysiological responses to different levels of cognitive and physical workload in haptic interaction. *Robotica* 29, 3 (2011), 367–374.

[147] Dan R Olsen and Michael A Goodrich. 2003. Metrics for evaluating human-robot interactions. In *Proceedings of PERMIS*, Vol. 2003. 4.

[148] Chris A. C. Parker and Elizabeth Croft. 2010. J-Strips: Haptic Joint Limit Warnings for Human-Robot Interaction. In *Volume 8: Dynamic Systems and Control, Parts A and B*. ASMEDC, 789–795. https://doi.org/10.1115/IMECE2010-40717

[149] Mildred L Patten and Michelle Newhart. 2017. *Understanding research methods: An overview of the essentials*. Taylor & Francis.

[150] Elizabeth Phillips, Xuan Zhao, Daniel Ullman, and Bertram F Malle. 2018. What is Human-like? Decomposing Robots' Human-like Appearance Using the Anthropomorphic roBOT (ABOT) Database. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 105–113.

[151] Karola Pitsch, Hideaki Kuzuoka, Yuya Suzuki, Luise Sussenbach, Paul Luff, and Christian Heath. 2009. "The first five seconds": Contingent stepwise entry into an interaction as a means to secure sustained engagement in HRI. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 985–991.

[152] Matthew S Prewett, Ryan C Johnson, Kristin N Saboe, Linda R Elliott, and Michael D Coovert. 2010. Managing workload in human–robot interaction: A review of empirical studies. *Computers in Human Behavior* 26, 5 (2010), 840–856.

[153] Matthew S Prewett, Kristin N Saboe, Ryan C Johnson, Michael D Coovert, and Linda R Elliott. 2009. Workload in human-robot interaction: a review of manipulations and outcomes. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 53. SAGE Publications Sage CA: Los Angeles, CA, 1393–1397.

[154] Paul C Price, Rajiv Jhangiani, I-Chant A Chiang, et al. 2015. *Research methods in psychology*. BCCampus.

[155] David Puljiz, Gleb Gorbachev, and Björn Hein. 2018. Implementation of augmented reality in autonomous warehouses: challenges and opportunities. *arXiv preprint arXiv:1806.00324* (2018).

[156] Mansour Rahimi and Waldemar Karwowski. [n.d.]. A research paradigm in human-robot interaction. 5, 1 ([n. d.]), 59–71. https://doi.org/10.1016/0169-8141(90)90028-Z

[157] Paul Ralph and Ewan Tempero. 2018. Construct validity in software engineering research and software metrics. In *Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018*. 13–23.

[158] Kandethody M Ramachandran and Chris P Tsokos. 2020. *Mathematical statistics with applications in R*. Academic Press.

[159] Pei-Luen Patrick Rau, Ye Li, and Jun Liu. 2013. Effects of a Social Robot's Autonomy and Group Orientation on Human Decision-Making. , e263721 pages. https://doi.org/10.1155/2013/263721 DOI: https://doi.org/10.1155/2013/263721.

[160] Daniel J Rea, Denise Geiskkovitch, and James E Young. 2017. Wizard of awwws: Exploring psychological impact on the researchers in social HRI experiments. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 21–29.

[161] James Reason. 1990. *Human error*. Cambridge university press.

[162] Sherri A. Rehfeld. 2006. The Impact Of Mental Transformation Training Across Levels Of Automation On Spatial Awareness In Human-robot Interaction.

[163] Laurel D Riek. 2012. Wizard of Oz studies in HRI: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction* 1, 1 (2012), 119–136.

[164] Alexander Rind. 2011. Some Whys and Hows of Experiments in Human–Computer Interaction. *Foundations and Trends® in Human–Computer Interaction* 5, 4 (2011), 299–373. https://doi.org/10.1561/1100000043

[165] Nicole L Robinson, Jennifer Connolly, Genevieve M Johnson, Yejee Kim, Leanne Hides, and David J Kavanagh. 2018. Measures of incentives and confidence in using a social robot. *Science Robotics* 3, 21 (2018), Article–number.

[166] Nicole Lee Robinson, Timothy Vaughan Cottier, and David John Kavanagh. 2019. Psychosocial health interventions by social robots: systematic review of randomized controlled trials. *Journal of medical Internet research* 21, 5 (2019), e13203.

[167] Nicole L Robinson, Teah-Neal Hicks, Gavin Suddrey, and David J Kavanagh. 2020. The Robot Self-Efficacy Scale: Robot Self-Efficacy, Likability and Willingness to Interact Increases After a Robot-Delivered Tutorial. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 272–277.

[168] A. Roncone, O. Mangin, and B. Scassellati. 2017. Transparent role assignment and task allocation in human robot collaboration. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. 1014–1021.

[169] Astrid M. Rosenthal-von der Pütten, Nicole C. Krämer, Laura Hoffmann, Sabrina Sobieraj, and Sabrina C. Eimler. 2013. An Experimental Study on Emotional Reactions Towards a Robot. *International Journal of Social Robotics* 5, 1 (Jan 2013), 17–34. https://doi.org/10.1007/s12369-012-0173-8

[170] Matthew Rueben, Shirley A Elprama, Dimitrios Chrysostomou, and An Jacobs. 2020. Introduction to (Re) Using Questionnaires in Human-Robot Interaction Research. In *Human-Robot Interaction*. Springer, 125–144.

[171] Craig J Russell and Michelle A Dean. 2000. To log or not to log: Bootstrap as an alternative to the parametric estimation of moderation effects in the presence of skewed dependent variables. *Organizational Research Methods* 3, 2 (2000), 166–185.

[172] John Rust and Susan Golombok. 2014. *Modern psychometrics: The science of psychological assessment*. Routledge.

[173] G. Ruxton. 2017. Allocation concealment as a potentially useful aspect of randomised experiments. *Behavioral Ecology and Sociobiology* 71, 2 (2017), 1–3.

[174] Neil J Salkind. 2010. *Encyclopedia of research design*. Vol. 1. Sage.

[175] Paul M Salmon, Neville A Stanton, Guy H Walker, Daniel Jenkins, Darshna Ladva, Laura Rafferty, and Mark Young. 2009. Measuring Situation Awareness in complex systems: Comparison of measures study. *International Journal of Industrial Ergonomics* 39, 3 (2009), 490–500.

[176] Anara Sandygulova, Wafa Johal, Zhanel Zhexenova, Bolat Tleubayev, Aida Zhanatkyzy, Aizada Turarova, Zhansaule Telisheva, Anna CohenMiller, Thibault Asselborn, and Pierre Dillenbourg. 2020. CoWriting Kazakh: Learning a New Script with a Robot. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 113–120.

[177] Daniel J Schad, Michael Betancourt, and Shravan Vasishth. 2019. Toward a principled Bayesian workflow in cognitive science. *arXiv preprint arXiv:1904.12765* (2019).

[178] KRISTIN E SCHAEFER. 2013. *THE PERCEPTION AND MEASUREMENT OF HUMAN-ROBOT TRUST*. Ph.D. Dissertation. University of Central Florida Orlando, Florida.

[179] Stefano Scheggi, Fabio Morbidi, and Domenico Prattichizzo. 2014. Human-Robot Formation Control via Visual and Vibrotactile Haptic Feedback. *IEEE Transactions on Haptics* 7, 4 (Oct 2014), 499–511. https://doi.org/10.1109/TOH.2014.2332173

[180] Jean Scholtz. 2003. Theory and evaluation of human robot interactions. In *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the*. IEEE, 10–pp.

[181] Mariah L Schrum, Michael Johnson, Muyleng Ghuy, and Matthew C Gombolay. 2020. Four Years in Review: Statistical Practices of Likert Scales in Human-Robot Interaction Studies. *arXiv preprint arXiv:2001.03231* (2020).

[182] Kenneth F Schulz and David A Grimes. 2002. Allocation concealment in randomised trials: defending against deciphering. *The Lancet* 359, 9306 (2002), 614–618.

[183] Yasaman Sadat Sefidgar. 2012. *TAMER: touch-guided anxiety management via engagement with a robotic pet efficacy evaluation and the first steps of the interaction design*. Ph.D. Dissertation. University of British Columbia. https://doi.org/10.14288/1.0052145

[184] Johanna Seibt. [n.d.]. Towards an Ontology of Simulated Social Interaction: Varieties of the "As If" for Robots and Humans. In *Sociality and Normativity for Robots: Philosophical Inquiries into Human-Robot Interactions*, Raul Hakli and Johanna Seibt (Eds.). Springer International Publishing, 11–39. https://doi.org/10.1007/978-3-319-53133-5_2

[185] Aran Sena and Matthew Howard. 2020. Quantifying teaching behavior in robot learning from demonstration. *The International Journal of Robotics Research* 39, 1 (2020), 54–72.

[186] Mingyang Shao, Matt Snyder, Goldie Nejat, and Beno Benhabib. 2020. User Affect Elicitation with a Socially Emotional Robot. *Robotics* 9, 2 (2020), 44.

[187] David J Sheskin. 2020. *Handbook of parametric and nonparametric statistical procedures*. crc Press.

[188] Jaeeun Shim and Ronald C. Arkin. [n.d.]. A Taxonomy of Robot Deception and Its Benefits in HRI. In *2013 IEEE International Conference on Systems, Man, and Cybernetics* (2013-10). 2328–2335. https://doi.org/10.1109/SMC.2013.398 ISSN: 1062-922X.

[189] Elaine Short, Justin Hart, Michelle Vu, and Brian Scassellati. 2010. No fair!! an interaction with a cheating robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 219–226.

[190] Jonathan A Smith and Pnina Shinebourne. 2012. *Interpretative phenomenological analysis*. American Psychological Association.

[191] Linda J Smith. 2008. How ethical is ethical research? Recruiting marginalized, vulnerable groups into health services research. *Journal of Advanced nursing* 62, 2 (2008), 248–257.

[192] Gerald Steinbauer. 2012. A survey about faults of robots used in robocup. In *Robot Soccer World Cup*. Springer, 344–355.

[193] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. [n.d.]. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction* (Salt Lake City, Utah, USA, 2006-03-02) *(HRI '06)*. Association for Computing Machinery, 33–40. https://doi.org/10.1145/1121241.1121249

[194] Gail M Sullivan and Anthony R Artino Jr. 2013. Analyzing and interpreting data from Likert-type scales. *Journal of graduate medical education* 5, 4 (2013), 541–542.

[195] JaYoung Sung, Henrik I Christensen, and Rebecca E Grinter. 2009. Robots in the wild: understanding long-term use. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 45–52.

[196] W Newton Suter. 2011. *Introduction to educational research: A critical thinking approach*. SAGE publications.

[197] David M Swanson and Rebecca A Betensky. 2015. Research participant compensation: a matter of statistical inference as well as ethics. *Contemporary clinical trials* 45 (2015), 265–269.

[198] Daniel Szafir. 2019. Mediating human-robot interactions with virtual, augmented, and mixed reality. In *International Conference on Human-Computer Interaction*. Springer, 124–149.

[199] Barbara G Tabachnick, Linda S Fidell, and Jodie B Ullman. 2007. *Using multivariate statistics*. Vol. 5. Pearson Boston, MA.

[200] Leila Takayama and Caroline Pantofaru. 2009. Influences on proxemic behaviors in human-robot interaction. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 5495–5502. https://doi.org/10.1109/IROS.2009.5354145

[201] Ana Tanevska, Francesco Rea, Giulio Sandini, Lola Cañamero, and Alessandra Sciutti. 2020. A Socially Adaptable Framework for Human-Robot Interaction. *arXiv preprint arXiv:2003.11410* (2020).

[202] Adriana Tapus, Cristian Tapus, and Maja J Mataric. 2009. The use of socially assistive robots in the design of intelligent cognitive therapies for people with dementia. In *2009 IEEE international conference on rehabilitation robotics*. IEEE, 924–929.

[203] Lehana Thabane, Jinhui Ma, Rong Chu, Ji Cheng, Afisi Ismaila, Lorena P Rios, Reid Robson, Marroon Thabane, Lora Giangregorio, and Charles H Goldsmith. 2010. A tutorial on pilot studies: the what, why and how. *BMC medical research methodology* 10, 1 (2010), 1.

[204] Kyle A Thomas and Scott Clifford. 2017. Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior* 77 (2017), 184–197.

[205] Andrea Thomaz, Guy Hoffman, and Maya Cakmak. [n.d.]. Computational Human-Robot Interaction. 4, 2 ([n. d.]), 105–223. https://doi.org/10.1561/2300000049

[206] Bruce Thompson. 1994. Common Methodology Mistakes in Dissertations, Revisited. (1994).

[207] Sebastian Thrun. 2004. Toward a framework for human-robot interaction. *Human–Computer Interaction* 19, 1-2 (2004), 9–24.

[208] Lorenza Tiberio, Amedeo Cesta, and Marta Olivetti Belardinelli. 2013. Psychophysiological methods to evaluate user's response in human robot interaction: a review and feasibility study. *Robotics* 2, 2 (2013), 92–121.

[209] Daniel C. Tozadore, Joao P.H. Valentini, Victor H.S. Rodrigues, Fernando M.L. Vendrameto, Rodrigo G. Zavarizz, and Roseli A.F. Romero. 2018. Towards Adaptation and Personalization in Task Based on Human-Robot Interaction. In *2018 Latin American Robotic Symposium, 2018 Brazilian Symposium on Robotics (SBR) and 2018 Workshop on Robotics in Education (WRE)*. 383–389. https://doi.org/10.1109/LARS/SBR/WRE.2018.00075

[210] Margaret L Traeger, Sarah Strohkorb Sebo, Malte Jung, Brian Scassellati, and Nicholas A Christakis. 2020. Vulnerable robots positively shape human conversational dynamics in a human–robot team. *Proceedings of the National Academy of Sciences* 117, 12 (2020), 6370–6375.

[211] Jan Van den Broeck, Solveig Argeseanu Cunningham, Roger Eeckels, and Kobus Herbst. 2005. Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS Med* 2, 10 (2005), e267.

[212] Edwin Van Teijlingen and Vanora Hundley. 2002. The importance of pilot studies. *Nursing Standard (through 2013)* 16, 40 (2002), 33.

[213] Joachim Vandekerckhove, Jeffrey N Rouder, and John K Kruschke. 2018. Bayesian methods for advancing psychological science.

[214] Renato Paredes Venero and Alex Davila. 2020. Experimental Research Methodology and Statistics Insights. In *Human-Robot Interaction*. Springer, 333–353.

[215] Michael L Walters, Kerstin Dautenhahn, Sarah N Woods, and Kheng Lee Koay. 2007. Robotic etiquette: results from user studies involving a fetch and carry task. In *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 317–324.

[216] Lin Wang, Pei-Luen Patrick Rau, Vanessa Evers, Benjamin Krisper Robinson, and Pamela Hinds. 2010. When in Rome: the role of culture & context in adherence to robot recommendations. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 359–366.

[217] Y. Wang, F. Guimbretière, and K. E. Green. 2020. Are Space-making Robots, Agents? Investigations on User Perception of an Embedded Robotic Surface. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 1230–1235. https://doi.org/10.1109/RO-MAN47096.2020.9223532

[218] Astrid Weiss and Christoph Bartneck. [n.d.]. Meta analysis of the usage of the godspeed questionnaire series. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 381–388.

[219] Astrid Weiss, Regina Bernhaupt, Michael Lankes, and Manfred Tscheligi. 2009. The USUS evaluation framework for human-robot interaction. In *AISB2009: proceedings of the symposium on new frontiers in human-robot interaction*, Vol. 4. 11–26.

[220] Tracey L Weissgerber, Natasa M Milic, Stacey J Winham, and Vesna D Garovic. 2015. Beyond bar and line graphs: time for a new data presentation paradigm. *PLoS biology* 13, 4 (2015).

[221] Tracey L Weissgerber, Marko Savic, Stacey J Winham, Dejana Stanisavljevic, Vesna D Garovic, and Natasa M Milic. 2017. Data visualization, bar naked: A free tool for creating interactive graphics. *Journal of Biological Chemistry* 292, 50 (2017), 20592–20598.

[222] Franz Werner. 2020. A Survey on Current Practices in User Evaluation of Companion Robots. In *Human-Robot Interaction*. Springer, 65–88.

[223] Rand R Wilcox and Guillaume A Rousselet. 2018. A guide to robust statistical methods in neuroscience. *Current protocols in neuroscience* 82, 1 (2018), 8–42.

[224] Michael S Wogalter. 2006. *Handbook of warnings*. CRC Press.

[225] Lesley Wood, Matthias Egger, Lise Lotte Gluud, Kenneth F Schulz, Peter Jüni, Douglas G Altman, Christian Gluud, Richard M Martin, Anthony J G Wood, and Jonathan A C Sterne. 2008. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ* 336, 7644 (2008), 601. http://bmj.com/content/336/7644/601.full.pdf

[226] Sebastian Wrede, Christian Emmerich, Ricarda Grünberg, Arne Nordmann, Agnes Swadzba, and Jochen Steil. 2013. A User Study on Kinesthetic Teaching of Redundant Robots in Task and Configuration Space. *Journal of Human-Robot Interaction* 2, 1 (Mar 2013), 56–81. https://doi.org/10.5898/JHRI.2.1.Wrede

[227] Yuki Yamashita, Hisashi Ishihara, Takashi Ikeda, and Minoru Asada. 2016. Path analysis for the halo effect of touch sensations of robots on their personality impressions. In *International Conference on Social Robotics*. Springer, 502–512.

[228] H.A. Yanco and J. Drury. [n.d.]. Classifying human-robot interaction: an updated taxonomy. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)* (2004-10), Vol. 3. 2841–2846 vol.3. https://doi.org/10.1109/ICSMC.2004.1400763 ISSN: 1062-922X.

[229] Holly A Yanco, Adam Norton, Willard Ober, David Shane, Anna Skinner, and Jack Vice. 2015. Analysis of Human-robot Interaction at the DARPA Robotics Challenge Trials. *Journal of Field Robotics* (2015).

[230] Deng Yongda, Li Fang, and Xin Huang. 2018. Research on multimodal human-robot interaction based on speech and gesture. *Computers & Electrical Engineering* 72 (Nov 2018), 443–454. https://doi.org/10.1016/j.compeleceng.2018.09.014

[231] James E. Young, JaYoung Sung, Amy Voida, Ehud Sharlin, Takeo Igarashi, Henrik I. Christensen, and Rebecca E. Grinter. [n.d.]. Evaluating Human-Robot Interaction. 3, 1 ([n. d.]), 53–67. https://doi.org/10.1007/s12369-010-0081-8

[232] Oded Zafrani and Galit Nimrod. 2019. Towards a Holistic Approach to Studying Human-Robot Interaction in Later Life. *The Gerontologist* 59, 1 (Jan. 2019), e26–e36. https://doi.org/10.1093/geront/gny077 Publisher: Oxford Academic.

Received 2020; revised 2020; accepted 2020