A Primer for Conducting Experiments in Human-Robot Interaction

GUY HOFFMAN, Cornell University
XUAN ZHAO, University of Chicago

We provide guidelines for planning, executing, analyzing, and reporting hypothesis-driven experiments in Human-Robot Interaction (HRI). The intended audience are researchers in the field of HRI who are not trained in empirical research but who are interested in conducting rigorous human-participant studies to support their research. Following the chronological order of research activities and grounded in updated research practices in psychological and behavioral sciences, this primer covers recommended methods and common pitfalls for defining research questions, identifying constructs and hypotheses, choosing appropriate study designs, operationalizing constructs as variables, planning and executing studies, sampling, choosing statistical tools for data analysis, and reporting results.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Computer systems organization** → **Embedded systems**; *Robotics*.

Additional Key Words and Phrases: experimental studies, statistical analysis, research methods

ACM Reference Format:
Guy Hoffman and Xuan Zhao. 2020. A Primer for Conducting Experiments in Human-Robot Interaction. *ACM Trans. Hum.-Robot Interact.* 1, 1 (January 2020), 31 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Human-Robot Interaction (HRI) research spans a wide variety of academic disciplines and scholarly practices. Some of the work focuses on empirical studies in the tradition of social psychology and other social sciences, investigating how humans perceive and interact with robots in various contexts. Another part of the HRI field is concerned with the design of new robots, algorithms, and interaction methods, rooted in robotics engineering, computer science, and design research. Yet another group of researchers are exploring the theoretical, cultural, and societal aspects of HRI, grounded in humanities traditions.

Despite the substantial diversity in researchers' backgrounds and methodological training, it has become a common practice in HRI to include an empirical evaluation of new systems, designs, or theories. Conducting studies that adhere to rigorous methodologies is critical to obtaining scientific knowledge about these systems, designs, and theories [27, 38]. In this paper, we provide a primer with principles and recommendations for each stage of the process of conducting experimental research in HRI: planning, execution, analysis, and reporting. The material presented here does not span the full spectrum of HRI methodologies. It focuses primarily on hypothesis-driven experimental research and does not, for example, cover context-grounded or observational research.

Authors' addresses: Guy Hoffman, Cornell University, 124 Hoy Road, Ithaca, New York, 14850; Xuan Zhao, University of Chicago, 5807 South Woodlawn Avenue, Chicago, Illinois, 60637.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2020 Association for Computing Machinery.
Manuscript submitted to ACM

This primer is mainly aimed at readers who may be new to conducting experiments. It is intended to help them understand the core principles and concepts, get up to speed in running their first studies, and avoid making common mistakes. To that end, it offers a hands-on guide on how to plan, execute, analyze, and report an HRI experiment from start to finish; it outlines important terminology, best practices, and common pitfalls in every essential step of conducting experimental research; and it features some of the most common designs and analyses used in HRI research. This tutorial aims to strike a balance between scope and depth and is therefore complementary to existing publications that discuss methods in HRI on a conceptual level [22] or provide targeted recommendations on specific methodological issues [7, 67]. Hence, it may serve as an accessible teaching material for the HRI community to train the next generation of experimental researchers.

Another motivation for this primer is that recent years have seen significant progress in the social sciences toward more rigorous methods and practices [53, 57, 61, 66]. This progress has been in part a response to a replication crisis that started in the early 2010s and has led to active discussion resulting in major changes in how empirical research is conducted in a variety of empirical disciplines [3, 5, 25, 30, 63, 71]. New practices have been introduced and are increasingly considered as necessary to improve research reproducibility [2, 42, 70]. Against this background, we hope to offer updated recommendations for HRI research methodology based on advances in related disciplines.

A third motivation of this paper is to clarify terminology and help researchers in the HRI community use a common language to describe their research goals, tools, and results. As an interdisciplinary research field, HRI has attracted scholars from a wide range of backgrounds and is successful in breaking down disciplinary boundaries, which has the potential to generate novel ideas with high impact [51]. At the same time, communication across disciplines can be challenging when colleagues expect different research practices and speak different scientific languages. We hope that this primer can assist researchers from different academic backgrounds to establish a common terminology around research processes, concepts, and practices, and to identify differences in disciplinary perspectives when they arise.

The recommendations put forward here are neither absolute nor exhaustive. There is active debate regarding research methods, and for every recommendation in this paper there may be well-founded scholarly disagreement. However, given the current state of empirical methods in many HRI publications, setting a baseline could help move the needle toward achieving better practices and more reproducible findings.

This primer is roughly organized in chronological order of empirical research activities (see: Fig. 1). It begins with clarifying research questions, constructs, and hypotheses (Sections 2 and 3), then moves to study design, which includes choosing variables and measures (Section 4), planning the study procedure (Section 5), and sampling participants (Section 6). It then goes on to cover data collection (Section 7), some commonly used experimental statistical methods and their appropriateness in various contexts (Section 8), as well as recommendations for reporting and discussing results (Section 9). This presentation order also corresponds to how research is usually presented in an academic paper. In addition, this primer discusses the limitations of the proposed recommendations (Section 10). We conclude by offering suggestions on how our field can take collective efforts to produce more rigorous and reproducible experimental research (Section 11).

Due to the limited scope afforded by a single paper, the methods presented here are biased toward experimental and quantitative research. We comment in less detail on observational studies and qualitative research. This should not be seen as an endorsement of specific methods (see: Section 10). In fact, we believe that qualitative research methods have been much underutilized in HRI research, and we encourage readers to familiarize themselves with these methods by seeking out other sources.

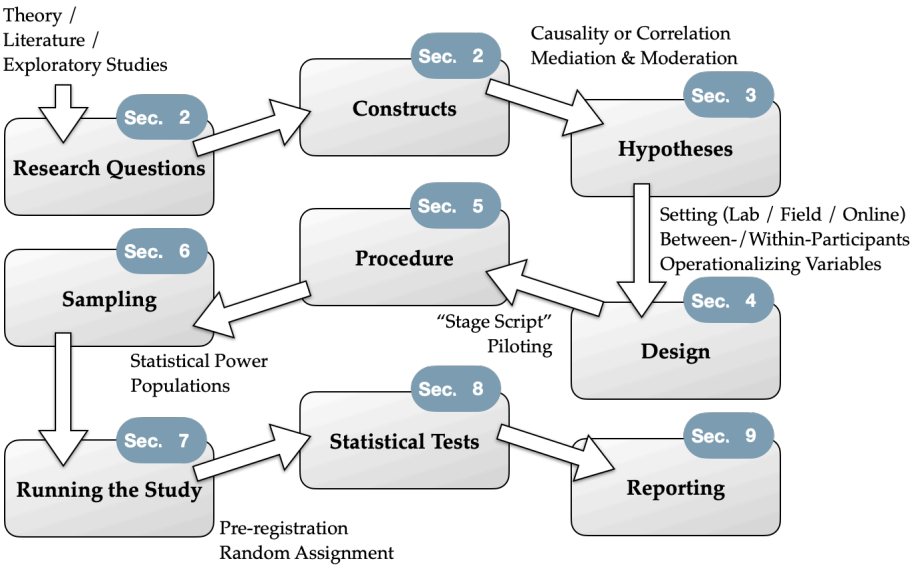

Fig. 1. The stages of empirical research, mapped onto the sections of this document.

1.1 Example Scenario: A Robot Walking Side-by-Side with a Human

Throughout this paper, we will use the following example scenario to illustrate the presented concepts: In this scenario, you are a roboticist and have designed a new algorithm for mobile robots to navigate alongside humans. Your work is based on the idea that effective human-robot joint navigation requires an algorithm that adjusts the robot's navigation path based on the human user's movement in real time. Therefore, your algorithm uses the latest machine learning methods to track human walking patterns and employs expertly crafted planners to stay by the human's heels without colliding with them.

When you test your algorithm, it seems to work well. You ask a few students in your lab for their opinions and everyone tells you it is brilliant. To get a more objective evaluation, you show a video of the robot using your algorithm to five students in the campus cafeteria and ask them whether they would trust your algorithm to be safe, and four of them say "yes." So you draw the conclusion that your algorithm is doing a good job in terms of safety and user trust.

But there are many problems with this approach: For example, it is unclear whether simply asking people if they trust your algorithm after watching a video is an accurate measure of their actual trust. Also, students might respond positively because they want to make you happy. Furthermore, you do not know if asking five people is enough to draw a sound conclusion. And finally, it is difficult to tell if 80% is a low or a high percentage of supporters. In the next sections, we will address these and other issues in this scenario by describing how to plan and run a rigorous empirical study that evaluates whether your algorithm actually works as intended.

2 RESEARCH QUESTION AND CONSTRUCTS

At the core of empirical research are research questions and constructs. While you might be excited to start planning a study, withstand the temptation to leap directly from your system, design, or theory to the details of a study. Begin by clearly formulating the research questions and constructs underlying your research project.

2.1 Clarify Your Research Questions

Any empirical study starts with one or more clearly defined research questions. The reason you are running a study is to answer a research question, and it is recommended that you state this research question explicitly. You may soon find yourself buried in the details of study design and execution; having clarity about what question you intend to answer will be a beacon that helps you make decisions throughout every step of the process.

One important thing to remember is that research questions should always be phrased as questions, not statements. They can be broad, for example: "To what extent, if any, is the new algorithm better than the current state-of-the-art algorithm?" or "Under what conditions is the new algorithm better?" Or they can be narrow, such as: "Can the new algorithm prevent collisions with humans while maintaining a minimal distance?" or "Do people trust the new algorithm to run on a suitcase-carrying robot?" The scope of your question is up to you but, regardless, it is crucial to have a clear question that your study is meant to address, because the choice of research question will drive everything downstream. If you are not sure where to start, many research questions can be phrased with the prefix "to what extent, if any...".

We will continue with our example and the following two research questions: To what extent, if any, will a human-adaptive path algorithm make people trust the robot to accompany them? And do people feel safe walking with a robot running the new algorithm?

2.2 Identify Your Constructs

Once you have a clear research question, you need to phrase the question in terms of a relationship between *constructs*. Constructs are the theoretical and abstract concepts that you intend to investigate in your empirical study. In running the study, you evaluate what kind of relationships, if any, exist between your constructs. In our example scenario, possible constructs could be "human-adaptive movement," "trust in the robot," and "psychological safety." Constructs should be meaningful theoretical concepts and not specific ways to measure them. Therefore, "trust in the robot" is a construct, whereas the trust questionnaire you decide to use is not a construct, but an operationalization of this construct. We will expand more on operationalizing constructs in a later section.

In many cases, such as the one in the example, you would like to test causal relationships, i.e., whether changes in one thing cause changes in another. Evaluating causal relationships is valuable because it serves to explain why an outcome happens and can provide useful insight on how to make a desirable outcome more likely to happen. In these cases, constructs can be described as having a *predictor-outcome* relationship[1]. In other cases, you might be interested in, or be limited to, testing correlational relationships, i.e., whether or not two constructs are related, even though you cannot identify which one is the cause, or if a causal relationship exists at all. Correlation does not necessarily imply causation, but knowledge about correlational relationships is still valuable in several important ways (see Stanovich [76] for a review). For instance, once you know two constructs or measures are correlated, you can use a score on one measure to make a more accurate prediction of another measure.

Our example scenario does involve testing causality, namely whether human-adaptive movement *causes* a sense of safety and/or trust. You therefore decide that the human-adaptive movement is the predictor construct and that the user's sense of safety and trust in the robot are the two outcome constructs.

Because the choice of research questions and constructs is going to affect the rest of your empirical project, it is worthwhile to spend time and care choosing, defining, and clarifying them. To identify good research questions and constructs, related literature in human behavior (in areas such as social psychology, cognitive science, sociology,

[1]Note, however, that the terms "predictor" and "outcome" are also used in regression analysis, where a causal relationship is not necessarily assumed.

communication, consumer behavior, public health, and so on) is often a good source of inspiration. Identifying a theory that would carry over to the HRI situation you are studying will give your research a solid base to stand on.

Good researchers usually start with a vague idea and then look into the literature and see what has been done in order to judge whether the idea is novel and consequential. Only if the idea seems interesting and the theoretical constructs seem important should you go on to the next step, coming up with specific hypotheses about your constructs.

3 HYPOTHESES

After identifying a construct-based research question, it is time to formulate your *hypotheses*. Hypotheses are affirmative statements about relationships between your constructs that your study can either support or refute. In the walking-companion robot example, a hypothesis could be "adapting to human walking patterns leads to higher trust in the robot than not adapting to the human." In this case, your predictor is categorical, i.e., either adapting the movement to the human or not. In contrast, the hypothesis can be made about relationships between continuous predictors and outcomes, like "the more the robot takes distance into consideration, the safer people feel around the robot."

3.1 Specifying a Baseline

When you are formulating a hypothesis about the effects of a new system or design, you need to be clear about the alternative baseline you are comparing to. In our example, the new algorithm could be more trust-evoking than a specific algorithm currently used by most researchers in the field. Or it could be better than a human teleoperating the robot, or maybe even an actual human walking with another human. If your hypothesis does not end with a subclause like "compared to..." ask yourself if you forgot to include a baseline in your hypothesis. In addition, try to find a fair baseline. It is not unusual for peers to reject an experimental study on the basis that the claimed innovation was compared to an unfair "strawman" baseline.

3.2 The Problem of HARKing

Hypotheses have to be clearly and explicitly defined *before* running your study. Hypothesizing after the results are known (HARKing for short) has been identified as a threat to the credibility of research results and researchers' cumulative knowledge [46, 58, 70]. Changing or introducing a research hypothesis based on the discovery of an effect critically confuses confirmatory and exploratory research—it is akin to shooting first and then drawing the target around where you shot [70].

This is not to say that there is no benefit to conducting additional analyses after first examining the data. Such exploratory analyses may reveal unexpected and interesting associations and enrich people's understanding of the research phenomenon; however, the final paper should fully disclose the exploratory nature of such analyses, and it is recommended that researchers follow up important insights from their exploration with confirmatory studies down the road, where they start with clear and explicit hypotheses.

Your hypotheses should be clearly stated in your write-up of the study in the same form that you wrote them before running the study. As a stylistic note, you may find it helpful to number your hypotheses, or give them short names, and refer to them by their numbers and names throughout your paper, although not all publications follow this convention. In our examples, hypotheses could be written as:

- **H1 (Trust):** Users trust a walking-companion robot that adjusts its navigation path to the human's movement more than they trust the famous navigation algorithm `STRAIGHT_WALKER`.

- **H2 (Distance-Safety):** The higher the minimal distance allowed in the algorithm, the safer people will feel around the robot.

3.3 Where do Hypotheses Come From?

In the example above, hypotheses H1 and H2 were mainly based on wishful thinking. Instead, you can make a stronger argument and have a greater chance of finding supporting evidence if your hypotheses are based on theory [73]. "Theory" is a loaded term and may seem intimidating. In reality, the word can mean different things. It could refer to a well-established theory from another field, such as social psychology, but it could also be an ad-hoc theory that you yourself constructed based on your observations of people's daily social interactions. It could also be your own theory based on a literature review of previous studies in HRI. A good practice is to survey a set of previous work and other related literature and generate hypotheses based on existing knowledge rather than out of thin air.

An additional great way to construct hypotheses is to run exploratory studies—often taking the form of *pilot studies* (see also: Section 5.2). In the navigation example, you could recruit a few users and let them walk with the robot while performing a number of activities, and then interview them about their experiences. If most people noted how safe they felt when the robot was further away, that can be a good basis for a hypothesis. You could also introduce quantitative measures in your pilot study, so that you can analyze your results and form your hypotheses based on emerging trends identified among a small group of participants.

3.4 Higher-Order Relationships between Constructs: Mediation, Interaction, and Moderation

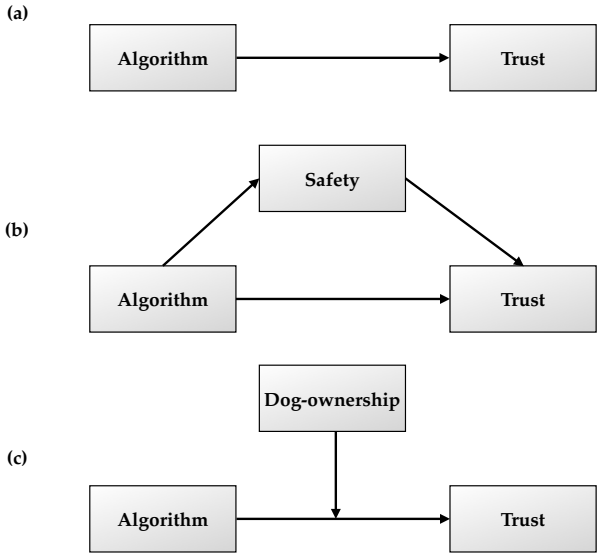

Fig. 2. Relationships between constructs, without (a) and with (b) mediation and (c) moderation.

In some cases, you may find it informative to hypothesize the relationship between more than two constructs. For instance, you might want to understand *how*, *why*, for *whom*, *when*, and *where* the adaptive algorithm increases trust. Two of the core methods to analyze these questions are called *mediation* and *moderation*.

Mediation analysis enables researchers to examine the "how" and "why" questions by hypothesizing an internal (often psychological) process through which one construct affects another [54]. There could be, for instance, a chain reaction among the three constructs we used in our example: Compared to a baseline algorithm, the human adaptive algorithm leads people to feel safer, which then leads to people's increased trust in the algorithm. Thus, the trust people feel in the human-adaptive algorithm is "mediated through their sense of safety." Fig. 2 (b) shows how mediation is usually represented in graphical form, compared to an unmediated relationship in Fig. 2 (a).

Moderation analysis enables researchers to examine the "for whom," "when," or "where" questions about the hypothesized relationship between predictor and outcome constructs. For instance, you speculate that dog ownership is related to people's sense of trust in the new algorithm, as shown graphically in Fig. 2 (c). You specifically hypothesize that the algorithm may elicit more trust in dog owners than in non-dog owners, because your algorithm was inspired by how dogs follow their owners and may have an intuitive appeal to those who are already accustomed to this movement pattern.

In this case, the "dog ownership" construct would be said to "moderate" the effect of "adaptive walking" on the outcome construct of trust. In some cases, a moderator can completely reverse the effect, i.e., the algorithm could work worse for non-dog owners than the baseline. In other cases, a moderator only affects the intensity of the effect, i.e., your algorithm always works better than the baseline algorithm, but *especially* for dog owners compared to non-dog owners. How one should conduct mediation and moderation analyses is beyond the scope of this paper; for a complete guide, see Hayes [37]. Finally, it is worth mentioning that, mathematically, a moderation effect is identical to an interaction effect in a factorial ANOVA test (see: Section 8.4); the critical difference lies in your research question and data interpretation—in our example, you are only interested in how the algorithm influences trust and how dog ownership moderates this relationship, but you are not interested in how dog ownership influences trust on its own.

4 DESIGNING THE STUDY

At this stage you have defined a research question, broken it down into a set of chosen constructs, and generated hypotheses about the relationships between these constructs. It is time to design your study.

4.1 Study Context: Laboratory, Field, or Internet

The same hypotheses can be tested in a number of contexts. The most popular options are testing in a laboratory, testing in the field, or testing on the Internet. Each has its benefits and drawbacks.

To start with, laboratory studies are conducted in a well-controlled environment, so they give you more control over your variables, allow you to conduct random assignment, establish causal relationships, and enable strict replication of conditions. Field studies, on the other hand, are conducted in the everyday environment of the participants and are more similar to real-life situations. This lends field studies *external* and *ecological validity*. External validity means that you stand on solid ground to claim that what you have found in your study can be generalized to other times, places, populations, and situations. In other words, good external validity suggests that your finding is not just a quirk of the specific procedure you used. An additional question is whether your conclusions are generalizable to the real world outside of the lab. This is called the *ecological validity* of your study. For instance, testing your adaptive-walking algorithm at a shopping mall can teach you more about how people would actually respond to this algorithm in their day-to-day lives than testing in the unfamiliar and constrained environment of a research laboratory.

That said, field studies are more difficult to set up and manage, and it is harder to control for *confounding variables* than it is in lab studies. Confounding variables (or simply "confounds" or "confounders") are factors that affect your

outcome constructs but are not part of your theory or hypotheses. In our navigation example, if you ran your study at a shopping mall, you could not control the noise level and the behavior of bystanders. These would be confounding variables, as they—in addition to the algorithm—might affect people's sense of safety.

Confounding variables are one of the factors that pose a threat to the *internal validity* of your study. Internal validity is defined as the extent to which a researcher can be "certain that the independent variable, or treatment, manipulated by the experimenter is the sole source or cause of systematic variation" in the outcome variable. [87] You may thus have an issue with low internal validity if your experiment produces "systematic sources of variance that are irrelevant to the treatment variable and not under the control of the researcher." [87]

In summary, laboratory and field studies often trade off external and ecological validity with internal validity, although you should attempt to take measures in both cases to minimize the negative effects of the chosen context.

A third option that has become increasingly common in recent years is running online studies. This option has been popularized by "crowdsourcing" platforms such as Amazon Mechanical Turk and Prolific [77]. In the context of HRI, this is usually done by showing participants photos or videos of robots or of humans interacting with robots and asking them questions via an online questionnaire. There are considerable advantages and disadvantages to this method: On the one hand, you can collect data from more participants more quickly with less financial cost, which allows you to easily increase the statistical power of your experiment (see: Section 6). In addition, the populations recruited online have a more representative demographic than the usual university participant pool that is made up mostly of high-achieving young adults. On the other hand, it is difficult to control online participants as they may be multitasking and are not as immersed in and committed to your study as face-to-face participants [88]. Furthermore, because some participants may participate in dozens of studies each day, they can be overly experienced in certain judgment tasks that are popular online and therefore might not show the same effects as naïve participants [13]. Perhaps the biggest problem, though, is that many studies require real-world interaction between participants and robots to get people's genuine responses, so watching a video lowers the external validity of online studies [11].

In light of the strengths and limitations of lab, field, and online studies, you could also consider combining them in the same project. For instance, you might conduct a quick and low-cost online study as a pilot study for exploratory purposes, then test the effect and its underlying processes in a well-controlled laboratory setting, and then go out to the field to corroborate your finding in natural environments. This approach has been coined "full-cycle research" [56]. Whichever context you choose, make sure you consider these options with their respective benefits and drawbacks in mind during study design, explicitly explain the reasoning that led to your decision, and discuss potential weaknesses and limitations in your paper.

4.2 Between- and Within-Participants Designs

If you are running a human-participant experimental study, you can choose to either use a *between-participants* or a *within-participants* study design. A between-participants study means that each participant is randomly assigned into one group to experience one variation of your predictor construct (which is also referred to as a *"condition"* or *"level"*), and you compare these different groups of participants. A within-participants design means that each person experiences more than one condition, and you compare the different experiences for each participant.

In our example, we have two conditions of the predictor construct: the adaptive robot walking algorithm and the baseline algorithm `STRAIGHT_WALKER`. In a between-participants design, half your participants experience a robot using your new algorithm and the other half walk with a baseline robot. In a within-participants design, each participant

would experience one algorithm and then the other. The same logic can scale to experimental designs where a predictor has three or more levels.

When both designs are feasible, a within-participants design is often preferred over a between-participants design due to a clear advantage: fewer participants are needed in a within-participants design to achieve the same statistical power (more on statistical power in Section 6). One way to think about this is that each participant serves as their own control, so you are less affected by *individual differences* between participants. In our example, people might have different baseline anxiety levels affecting their trust in everything. When you compare each person's responses in one condition to that same person's responses in the other condition, individual differences in anxiety level "cancel out," as they affect both sides of the comparison.

Why should you not always use a within-participants designs then? One crucial reason is that within-participants designs suffer from *order effects*. This refers to the possibility that the effects you find are confounded with the order in which people experience the different conditions. In our example, it may be that people get used to the robot, and trust it more the second time around, an effect completely unrelated to the algorithm used. This would be an example of a *familiarity* effect. Another important order effect is *fatigue* which increases with time spent on the experiment. A related order effect is the *novelty effect*, where people react differently to things they experience for the first time. Other factors such as learning and habituation may also lead to order effects.

Order effects to look out for
Ask yourself whether any of these order effects might be a substantial confound that can affect your outcomes and thus undermine the internal validity of your study.

- Familiarity - Participants know the task and have more information in later rounds.
- Novelty - Experiencing something for the first time is different, for better or worse.
- Habituation - People may get bored, or desensitized to the manipulation.
- Learning - Participants get better at a task when they repeat it.
- Fatigue - Participants' cognitive and physical abilities decline over time.

Fig. 3. Potential order effects in within-participants designs.

There are ways to mitigate order effects. The most popular one is called *counterbalancing*, meaning that you randomize the order of the conditions that your participants experience. This way, you can still use a within-participants design. In our case, half of the participants would be randomly assigned to the new algorithm first, and the baseline second, and half of them would do it the other way around. Make sure to rigorously randomize these assignments (see: Section 7.3).

When counterbalancing is not feasible, there are other ways to minimize order effects. These include giving people long breaks between conditions, sometimes up to several days, and including an initial training run to mitigate novelty and learning effects before officially starting your data collection.

You should always test for order effects on your collected data using statistical tests. Moreover, you can add the order of your manipulation as a separate variable and control for its effect in your statistical analysis.

There are situations where order effects cannot be mitigated by any of the above-mentioned techniques. For instance, it might not make sense for participants to evaluate the same stimulus with slight variation more than once, especially if you are interested in people's initial responses [87]. Perhaps it is impossible to administer two different conditions while keeping the same participants around in a field study. These situations require a between-participants design.

4.3 Operationalizing Constructs into Variables and Measures

The next step in designing your study is to convert your constructs into specific things you can manipulate and measure, a process called *operationalization*. This may seem like a trivial point, but it actually involves critical decision-making, because a construct can be operationalized in many different ways. Consider our construct of "trust in the robot" from the navigation example. To operationalize "trust" you need to start with a basic question: What does it mean to "trust" a robot? To ensure conceptual clarity and consistency, you need to examine the prior literature in order to understand how other researchers have conceptualized "trust" in the social sciences and in the context of HRI. With a clear definition, you can then consider how to measure trust.

For instance, you could ask people to rate how trustworthy they perceive the robot to be using a questionnaire [e.g., 24]. You could ask open-ended questions during interviews or online surveys and assess their level of trust by conducting a textual analysis on their verbal responses [e.g., 52]. You could ask them to read hypothetical scenarios that require trust and rate how willing they are to use the robot in each scenario [e.g., 80]. If you do not want to ask participants directly, you could measure physiological indicators of their emotional arousal, which can sometimes be related to trust [86]. Or, you could introduce tasks that measure actual trust-related behaviors. Such behavioral measures can vary widely from letting a robot enter a secure-access area [9], to complying a robot's instruction during an emergency [68], to entrusting money to a robot in an economic game [55].

All of these examples illustrate that having decided on your construct does not automatically determine your measures, and how to operationalize a construct is a separate decision. Further, just as you should thoughtfully operationalize an outcome construct (in this case, "trust") into a measure, you should similarly operationalize your predictor into a specific *manipulation*. Once you do that, you then must evaluate the quality of your manipulation using a *manipulation check* (Section 5.3).

The operationalization you choose will critically affect the soundness of your study design. When the manipulations and measures in your study design do not adequately reflect the theoretical constructs that gave rise to the research question in the first place, your peers might rightfully argue that your study has a problem with *construct validity* [32]. Table 1 shows a comparison between the four different kinds of validity presented in this paper: external, ecological, internal, and construct validity, along with factors that may negatively affect them. See also Wilson *et al.* [87] or Bordens and Abbott [10] for an overview with more examples.

There is no easy way to choose the right operationalization, but you should expect to spend a significant amount of time at this stage and brainstorm different possibilities before settling on your final decisions.

4.3.1 A Terminological Note. Operationalized variables involved in causal relationships between constructs are often called *independent* and *dependent* variables, mapping roughly onto the predictor-outcome relationship between constructs, but with a special emphasis on experimental manipulation: The independent variables are the ones that you directly manipulate as part of an experiment, and the dependent variables are the ones that you measure as a result of (they are "dependent on") the manipulation.

4.3.2 Subjective vs. Objective Measures. Generally speaking, when operationalizing a construct into a variable, you can choose between *subjective* and *objective* measures. Subjective measures are self-reported attitudes, thoughts, emotions, and moods of participants, collected through participants' verbal responses. Objective measures are behavioral indicators you can measure independently of people's stated opinion. The same construct can often be measured by subjective or objective means, and it is usually good practice to use a combination of both. For example, people's trust in the robot

can be measured using a trust questionnaire (subjective) or by measuring how many times people look at the robot to ensure it is following them (objective).

The two main ways to implement subjective measures are through questionnaires and interviews. Objective measures include a broader range of possibilities; for example, you can measure people's decisions, reaction speed, physiological reactions, and so on. Such data are usually obtained through observation, either during the study or later by reviewing video footage of the study, or via data logs from the system the person is interacting with. This system can be a physical robot in field or lab studies or a computer that administers study materials.

Deciding which measure to use requires careful evaluation of the advantages and disadvantages of each. Self-report is easier to administer and allows you to take advantage of pre-validated and widely adopted questionnaires from social psychology, personality psychology, HRI, and other fields.[2] Whether you use an established questionnaire or need to create your own due to a lack of established instruments, you must conduct a reliability analysis to make sure that your survey items have good internal consistency and report this analysis. Internal consistency in this context means that the survey items intended to measure the same psychological construct (e.g., "perceived trustworthiness") are indeed closely related as a group. The most widely used measure of reliability is Cronbach's alpha, and many consider a Cronbach's alpha of .70 or higher necessary to justify calculating a composite score from multiple items. However, some advocate for more nuances in deciding what a minimally acceptable Cronbach's alpha should be [8].

That said, self-report has its own set of problems. For instance, people's responses can be strongly influenced by wording, format, context, and their mood [69]; they might be inclined to provide socially desirable responses in order to portray themselves in a more positive light [82]; and perhaps most importantly, in the real world, people do not go about their everyday life by filling out questionnaires—it is people's actual behavior that ultimately matters [87]. Therefore, an objective measure is often more convincing, interesting, and generalizable to the real world.

But behavioral measures also have important limitations. For instance, some behaviors may be performed in private or specific contexts and cannot be easily observed by researchers. Moreover, the same behaviors may arise due to different reasons, so it is not always clear what a behavior actually reveals—for instance, one may stand closer to a robot either because they feel safe or because they are curious and want to inspect it more closely to form an opinion. Finally, when one is interested in understanding the underlying cognitive processes and mechanisms, such as what

[2]It is almost always preferred to use a previously validated questionnaire than to make one up. A previously validated questionnaire can help ensure the construct validity of your research and promote reproducibility.

	Claim	**Threatened by**
External Validity	The findings generalize beyond the specific setup of this study.	Idiosyncratic setups, contrived activities, a biased sample
Ecological Validity	The findings generalize to real-world situations.	Unnatural contexts such as laboratories, unrealistic tasks
Internal Validity	The findings are only related to the constructs we are interested in.	Weak experimental control, such as confounding variables, ineffective randomization, inconsistent procedures
Construct Validity	The measured outcomes tell us about the theoretical constructs we care about.	Bad operationalization, weak connection between theoretical constructs and manipulations/measures

Table 1. The four types of validity and factors that can threaten them.

psychological processes mediate the effect, subjective measures provide a window—albeit an imperfect one [60]—to people's underlying mental states.

Given the strengths and weaknesses of subjective and objective measures, a general recommendation is to include both types of measures to corroborate your findings whenever possible [7]. However, be prepared that these two types of measures may sometimes lead to weakly correlated and even inconsistent results, which can seem puzzling at first glance but can be very informative and revealing in many cases (see Dang et al. [21] for a discussion).

As with all of the study design decisions, make sure to explicitly consider and discuss different measurement options during your study design phase, and to describe the resulting decision and reasons for your choice in the research paper.

4.3.3 Quantitative vs. Qualitative Methods. Note that the above distinction between subjective and objective measures is *not* the same as that of *qualitative* vs. *quantitative* measures, although they are often confused. Quantitative measures require "the reduction of phenomena to numerical values in order to carry out statistical analyses"; by contrast, qualitative research often involves collecting data in the form of naturalistic verbal reports, and the analysis is textual [74]. Therefore, collecting numerical trust ratings from a questionnaire is a *quantitative* method, whereas coding trust-related themes from a participant's interview transcript is a *qualitative* method.

For illustration purposes, let us operationalize the construct of "psychological safety" into each of the four classes of variables (subjective/objective x qualitative/quantitative). Giving participants a scale questionnaire asking about their sense of safety would be a subjective quantitative measure. Interviewing them about their sense of safety and bringing up quotations and themes uncovered in the interview is a subjective qualitative measure. Counting how many times they look at the robot using data from a motion-tracking system or taking the median distance during the walk would be objective quantitative measures. Finally, describing participants' body language verbally, in a phrase such as "when first entering the room, people tended to avoid getting near the robot", is an objective qualitative measure.

With the development of automated text analysis tools, researchers now also have the option to use quantitative methods on textual data, somewhat bridging the distinction between quantitative and qualitative methods. For more details on these automated content analysis methods, see Section 10.

5 PLANNING THE STUDY PROCEDURE

Once you have a set of hypotheses operationalized into variables, the next step is to plan the particulars of the study. The outcome of this process is usually summarized in a "Procedure" section of your research paper.

5.1 Writing a "Stage Script"

When planing the study procedure, it is recommended to be explicit about every aspect of your procedure. Consider preparing your study protocol as writing a "stage script" for a play: What "props" do you need? How will you set up your "stage"? What should happen, step-by-step, from the moment your participants enter your space (be it a laboratory, a field site, or a website) to the moment they exit? How will you securely store participants' data after they complete your study? When situations deviate from your plan—which may range from an uncooperative or inattentive participant to a malfunctioning device—what is your contingency plan? Your protocol needs to cover all of these issues.

Obtaining consent is another aspect of this script that requires some planning. While the informed consent procedure is similar in many laboratory studies involving adults, certain studies, such as those involving deception, recruiting participants from a public space, or working with vulnerable populations (e.g., children) may necessitate alternative consent procedures. These procedures need also to be accounted for during the planning stage.

In addition to specifying how the experimenter should verbally introduce the task ("Tell the participant: *This is the robot you will be working with for the next hour*"), your protocol also needs to include "stage instructions" for the experimenter, such as where to meet participants, how to orient them in the lab space, when to prompt participants for questions ("At this time ask the participant if they have additional questions"), where to direct participants to fill out a questionnaire, and so on. We have seen well-written study scripts that include instructions like "pause here and look at the participant to make sure they are following the instruction," and "point to the robot at this moment." Including such information in your study script will allow your experimenters, or "actors," to take the procedure document and know exactly how to run the experiment without adding personal interpretation or subjective deliberation. If two or more experimenters are going to run the same procedure with different participants, as will often happen in lab or field studies, you also need to provide standard training to ensure that the way they each handle your study procedure, both verbally and non-verbally, is comparable.

When specifying the procedure, be aware that researchers can unintentionally transmit their expectations to participants through inadvertent cues such as nonverbal behaviors. These cues may skew participants' responses toward a particular result that the experimenter desires. Therefore, researchers need to put various measures in place to prevent such *experimenter expectancy effects*. One solution is to keep the experimenter from knowing the experimental condition they are administering. When it is impossible to keep the experimenter blind to the condition—likely because they must know a participant's condition in order to administer it—an alternative solution is to have one experimenter deliver the manipulation and then be replaced by another experimenter, who is blind to the condition, to finish the study procedure. If that is not viable, orienting participants to face away from the researcher while participating in the experiment can also help. An additional way to mitigate expectancy effects is for experimenters to tell participants that they did not themselves develop the robot technology but are truly interested in how well it works.

Writing a good study protocol takes considerable effort. It is almost guaranteed that you will overlook some important details in your first draft. Therefore, you should ask all collaborators and experimenters to read your draft and try to find overlooked issues in the proposed procedure. You also need to pilot your procedure on a few participants to make sure that your protocol makes sense to participants (see the next section). The more meticulous you are at this stage, the less likely it is that you will have to restart your study mid-way. In many cases, the institutional review board—an administrative office that reviews your study to protect the rights and welfare of human research subjects, or the "IRB" for short—also wants to see a detailed procedure before issuing their approval.

5.2 Piloting

As you are developing the procedure, you will start to pilot it. Piloting is so important that it deserves its own subsection. Even if you have spent weeks on study design and procedure debugging, there may be things you are not taking into account or simply cannot predict when developing your study without feedback from real participants. For instance, participants might find your study setup unbelievable, your instructions confusing, or your study design ineffective in manipulating the construct you actually aim to study (see also:"Manipulation Check" below). Once you pilot your procedure with a handful of volunteers as participants, you may uncover unexpected glitches that often require you to modify your script. Piloting is an illuminating and crucial step in study development. Researchers who do not pilot their procedures often end up aborting their experiments mid-way and wasting valuable participant time and payment. If you are working on a deadline, this becomes doubly disappointing.

There is an additional source of frustration that can be alleviated with piloting. Too often, researchers spend months on a costly experiment only to discover that none of their hypotheses were supported. To protect against this, there is

a temptation to measure a large number of constructs and later engage in questionable practices such as HARKing or cherry-picking that inflate the study's false positive rate [58]. In some cases, this lack of evidence happens due to oversights in the procedure or measures that could have been identified and redressed had they run a pilot study. In other cases, a pilot study could have suggested that the hypotheses are not likely to be supported. In both cases, a pilot or exploratory study can provide a more solid foundation for strong hypotheses.

When you write up your paper, you should never report pilot study findings as if these are your real experimental data or mix the data together into one pool. While this is especially true if you modified your procedure after the pilot, pooling the pilot and subsequent findings together is bad practice in any case, as it introduces vagueness into the literature. You may, however, report a pilot study in a separate section before reporting your actual study, especially when the pilot critically informed your final study design. In this case, the description of the pilot study and findings are helpful in justifying your design decisions.

5.3 Manipulation Check

The act of setting the level or value of a predictor variable in experimental studies is often referred to as *the manipulation*. To confirm the effectiveness of a manipulation, including a *manipulation check* is another important, but often overlooked, aspect of the experimental design.

When your predictor construct is a simple fact, like the height of a robot, it is easy to claim that you have successfully manipulated your predictor construct if you use a tall robot versus a short robot. However, if you hope to manipulate people's psychological states such as fear of robots, it is important to provide direct evidence that your study design is effective in manipulating the predictor construct that you aim to study.

You usually report the result of your manipulation check at the beginning of your results section before reporting your tests of any actual hypotheses. Manipulation checks are informative regardless of whether or not you find your hypothesized effect. When you find a difference in the outcome variable across different conditions, your peers want to see evidence that such a finding is caused by the successful manipulation of your predictor construct (e.g., fear of robots). When you do not find the effect you hypothesized, you need to diagnose whether it means your hypothesis might be wrong, or because you failed to manipulate the predictor construct that you want to study. In the latter case, you cannot draw any conclusions about your hypothesis, and you need to restart your study design.

6 SAMPLING

A final important decision to make between completing your procedure and running the study is how many data points you would like to collect, and what population will participate in your study.

6.1 Sample Size and Statistical Power

In the past, it was common for researchers to decide their sample size based on intuitions, prior practice, rules-of-thumb, or practical constraints [2, 4]. In the wake of the replication crisis in the social and life sciences, more and more publication venues expect their authors to justify their sample size with a more statistically sound method—usually a *power analysis*. To understand this concept, we need to start with the fundamental idea of the sampling process and the types of errors that it introduces.

6.1.1 Inferring from a Sample about the General Population. To know for sure whether any hypothesis is true or false on a general population (e.g., all humans, or people working with robots), you would need to measure every single

person of that entire population. However, in most cases, this is infeasible. Instead, researchers *infer* a probable answer to this question by sampling a subset of the population, measuring how the sampled participants behave in different experimental conditions, and then using statistical tools to determine whether the effect measured in the sample is sufficiently large to support their hypothesis. To determine what constitutes a "sufficiently large" effect, one needs to set a threshold in advance, which is also called the *critical value* of the test, and they can only declare the hypothesis supported by the sample when the difference is above the threshold.

Why is having a predetermined threshold necessary? Why can someone not just declare that their hypothesis is supported when they have observed any difference between the means of different conditions? The reason is that, when making inferences about a general population from a sample, two types of errors may occur.

6.1.2 Type I and Type II Errors. Let us go back to the navigation algorithm example. You want to know whether people trust your algorithm more than the state-of-the-art algorithm. You operationalized trust as a subjective quantitative measure obtained through a validated trust questionnaire. Then you recruited 40 participants and randomly assigned them to walk alongside a robot with either your algorithm or the baseline algorithm before completing the trust questionnaire. As you have suspected, people's average trust levels for the two algorithms are indeed different—specifically, people rate their trust in your algorithm, on average, by one scale-point higher than in the baseline algorithm. Is it time to conclude that you have obtained "empirical evidence" supporting your hypothesis about the general population? Not necessarily.

Since your sample constitutes only a subset of the population, such an inference might either be correct or incorrect. It might be that there is actually no difference between the general population's trust level elicited by the two algorithms—yet you have detected a difference merely due to variation introduced through sampling. This is called a *Type I error*, or a "false positive." To avoid publishing findings that are untrue, a threshold (critical value) for the effect is needed. The higher the threshold, the lower the rate of Type I errors–known as the alpha level (α). The conventional, yet somewhat arbitrary alpha level in many fields is .05, although recently there have been calls for lower alpha levels [6].

Why not set your critical value so high as to achieve a very low alpha level and minimize the chance of producing false positives? The reason is that, as you increase your critical value (and thereby lower your α level), you may inadvertently commit a *Type II error*, producing a "false negative": Your algorithm may actually have an effect on trust in the general population, yet you mistakenly declare that there is no effect because your threshold is too high. The rate at which you commit a Type II error is commonly labeled β. You can easily see this trade-off between Type I error (false positive) and Type II error (false negative): The lower you set your alpha level in order to avoid a Type I error, the higher the rate of a Type II error becomes (sample size being equal).

6.1.3 Statistical Power. Understanding sampling and sampling errors, we can now turn to the concept of statistical power. The *power* of a test is the probability that, if there is an effect in the general population, your test will be able to identify that effect in the sample. It is therefore simply $1 - \beta$, the complement of the Type II error rate. Intuitively, you want more power in your study, so that you do not miss your discovery and abandon a fruitful project after working on the study design for months.

Several factors influence the power of a study: the alpha level, the effect size, and the sample size. For a given level of alpha and a given sample size, if your manipulation has a large effect size, the power of your test is higher than if it had a small effect size. For a given effect size and a given alpha level, the only solution to gain sufficient power is to increase your sample size.

	Your test statistic **passes** the critical value	Your test statistic **does not pass** the critical value
There is **no effect** in the population	**Type I Error** (false positive) happens at a rate of α	**No Error** (true negative)
The **effect exists** in the population	**No Error** (true positive)	**Type II Error** (false negative) happens at a rate of β

Table 2. Type I and Type II errors, α and β rates.

Traditionally, HRI studies have used relatively small sample sizes. Many researchers did not realize that a small sample size, such as 15 participants per condition for a between-participants design with two conditions, would result in a low power of .26 even with an effect size of $d = 0.50$.[3] That is, if your algorithm actually has a meaningful impact on people's trust toward the robot, with a total of 30 participants, three out of four times you run the experiment you will not obtain a p value smaller than .05 to support your hypothesis. Beyond lowering your chance of making a discovery, even when you do report a statistically significant result, a small sample size will also raise questions as to whether your effect truly exists, because an observed significant result is more likely to be a false positive in an under-powered study than in a sufficiently powered study.

6.1.4 Power Analysis. To determine the target sample size, you need to conduct a *power analysis*. This method requires you to specify two numbers for your study: the desired statistical power and the expected effect size. It is conventional to consider .80 a reasonable power for an experiment [15, 16]. To estimate your effect size, you may look up previously published research or conduct an exploratory study. Alternatively, you may have to assume an effect size when performing your power analysis, because even in this case, power analysis can be informative. For instance, suppose that you expect an effect size of $d = 0.50$. Using a between-participants design with two conditions, your power analysis shows that you will need 64 participants per condition, or 128 in total, to reach an 80% chance to detect an effect at $\alpha = .05$. You may consider using other study designs, such as a within-participants design with repeated measures, to bring the required sample size down. You can find many toolkits that can help you perform power analysis and determine your target sample size.

Typically, you report your power analysis under the "Participants" heading in your "Method" section. Here is a template that you can follow: "We recruited 128 participants in total, which would allow us to detect an effect size of $d = 0.50$ with .80 power at an alpha level of .05 (calculated using the `G*Power` software [29].)"

Finally, even if you decide not to run a power analysis and just use a rule-of-thumb sample size, you still need to determine and declare your sample size *before* you run your study. It is unacceptable to first run some participants,

[3]Cohen's d has been a popular measure of effect size when comparing the means of two samples using a t-test. Traditionally, $d = 0.50$ is often considered as a "medium" effect size upon Cohen's recommendation. However, Cohen's recommendations are inconsistent across different statistical tests. See a summary on the number of participants needed to reach a "medium" effect size, given different statistical tests, in Correll *et al.* [19].

look at the results, and then decide whether you will add more participants in order to obtain significant results. This practice has been shown to dramatically inflate your Type I error rate from the claimed alpha level [71].

6.2 Sample Population

You should also consider what population you are drawing your samples from and justify it in your research paper. If you intend to study what you believe is generally true for every human, then ideally you want to draw a random sample from the whole population of humanity. However, researchers rarely, if ever, meet this ideal. Most studies rely on so-called *convenience samples*, which may be undergraduate students, as there are many of them in the vicinity of researchers. Recruiting online participants on crowdsourcing platforms can increase the diversity of your participants beyond college campuses, but workers who self-selected to use these platforms are not representative of the population, either [14, 64]. For in-person studies, you may be able to collect a more diverse sample than only college students by conducting a study in the field or by putting more effort into recruiting community members outside the university. Even then, you are still constrained by specific cultural or societal contexts that may not be representative of "humans" in general. As a result of these limitations, Henrich *et al.* [39] noted that most research studies use samples that are "drawn entirely from Western, Educated, Industrialized, Rich and Democratic (WEIRD) societies", who are "particularly unusual compared with the rest of the species". This negatively impacts the external validity, or generalizing capacity, of findings from these studies to other cultures and societies.

In some cases you want to exclude participant groups from your sample. This should be principled and explicit. You may have additional selection criteria, such as language skills or prior experience with robots. Decide on these criteria before you sample and list the exclusion considerations in your paper. Make sure that every experimenter running the study is aware of the previously determined exclusion criteria and adheres to them. If you decide to exclude a participant post hoc, this should also be clearly justified in your writing. Discarding participants post hoc without explanation should be avoided.

7 RUNNING THE STUDY

Running the study boils down to measuring the variables you have defined using the procedure you have ended up with after many iterations of piloting and revising. This section is short—perhaps surprisingly so. The reason is that, once you have a strong basis behind your study design and a carefully refined procedure, running the study should be an almost robotic endeavor. Everything should be perfectly laid out for experimenters to run the study, no matter if there is a single experimenter, possibly you, or a team of experimenters taking shifts running the study instead of you. The bulk of this section emphasizes some aspects of running a study that were not covered above.

7.1 Pre-registration

In an effort to combat problematic research practices that can inflate false positives, such as cherry-picking results, HARKing, and arbitrary decisions on sample size and statistical analyses, there is an increasing expectation for researchers to commit to their research questions and analysis plans before collecting data. This practice is called pre-registration [62, 83].

Pre-registration usually includes the full set of hypotheses, research materials, procedure, sample size and justification, exclusion criteria, and choice of statistical tests. There are many online services that support pre-registration, such as the Open Science Framework (OSF), the AEA Registry, EGAP, AsPredicted, and trial registries in the WHO Registry

Network. Some pre-registration websites provide a template, but researchers need not adhere to one; the key thing is to make clear which aspects of the study were specified in advance.

7.2 Informed Consent

Most publication venues and research institutions require an informed consent procedure, which, at a minimum, includes a document that participants sign, and steps ensuring the voluntariness of their participation, the privacy and confidentiality of their data, and other protections. Please consult with your institution's ethics review board (IRB) and your target publication venue for their requirements.

Ethics approval can take a long time, spanning several weeks or even months. Submitting a request for approval also generally requires that all of the documents discussed in previous sections be completed. Take this into account as you plan your study timeline.

7.3 Random Assignment

If you are conducting an experiment, you have to randomly assign participants to conditions. This applies to both between-participants designs, where you assign participants to one of the experimental conditions, and within-participants designs, where you counterbalance the order of the experimental conditions. The importance of random assignment cannot be overstated in empirical research as it provides the logical foundation of any claim to causality.

Random assignment precludes the possibility of participants self-selecting into a condition. In addition, the experimenter should not use a systematic, non-random process to generate the assignment, such as assigning participants in the morning to one condition and those in the afternoon to another, or systematically alternating between two conditions. Such practices introduce confounding variables (like time of day, participant preference, or experimenter bias) and undermine the internal validity of your research.

There are many valid ways to perform random assignment.[4] When using a digital survey, you can use the randomizer provided by common survey platforms. Alternatively, you may rely on random number generators found online to create a spreadsheet specifying the order of your conditions (in a within-participants design) or what condition each participant should be assigned to (in a between-participants design). Given the critical role of random assignment in empirical research, we recommend our readers to familiarize themselves with common techniques in order to choose one that works for their study design. For a practical guide, see the "Random Assignment" chapter in Coleman [17].

7.4 Debriefing

After the study is finished, the experimenter should inform participants about the nature of the research. This is especially important when the study involves deception or incomplete disclosure of information, because debriefing serves important ethical functions such as to identify any unforeseen harm, discomfort, or misconceptions, and arrange for assistance as needed. Moreover, by probing participants' reactions and responses, researchers can identify procedural problems, effectiveness of manipulations, participant suspiciousness, and so forth [75]. And finally, because many participants are college students, debriefing also provides educational benefits, such as insight into HRI research.

[4]Strictly speaking, what most researchers refer to as random assignment is actually pseudo-randomization that meets several additional constraints—such as ensuring equal number of trials in each condition and avoiding long runs of trials from the same condition [81].

7.5 Logging

Finally, you should keep a study log, in which you track an identification number for each participant along with the condition they are randomly assigned to. After completing the study procedure with each participant, you will update this log regarding their study date, time, and experimenter (in the case that there are several experimenters), as well as document any unusual events for each participant. Similarly, keep as many written records and computer logs from each run as possible for later review, in case any future concerns arise.

8 STATISTICAL TESTS

If you collected quantitative data, statistical tests can help you make claims about your research questions and hypotheses. There are large volumes written about inferential statistics, and there continue to be active debates over their use. A scholar usually requires many years to achieve a full appreciation of this topic, and we cannot do justice to the complexity of this issue in a single section.

With that disclaimer in mind, we will try to present here some of the more popular and useful statistical methods and when they are appropriate in an HRI study scenario. We will also try to address common misunderstandings and misuses of these tests. You are highly encouraged to read further about any particular statistical test, and its assumptions that you need to satisfy, before using it for your research.

8.1 Descriptive Statistics

The following sections focus on *inferential statistics*, which is concerned with using statistical tools to make claims about hypotheses. Before jumping in, it is worthwhile to mention that a study report should also include *descriptive statistics*, which do not directly test your hypotheses, but describe the collected data, usually grouped by condition.

The most commonly reported descriptive statistics relate to the *central tendency* of the data, most often presented by the *mean* (average), abbreviated with the letter M, but sometimes also by the *median* (a number relative to which 50% of the data are lower or equal) or the *mode* (the most commonly measured value). In addition, the *variance* of the data is usually presented, often in the form of the square root of the variance, called *standard deviation* (abbreviated as *SD*). Furthermore, the *range* of the data can be described, in terms of the *minimum* and *maximum* values detected. More detail can be given by specifying the *quartiles* of the data points, which represent a four-way division of your data, similar to the median but more precise. More advanced features, such as *skewness* and *kurtosis* of the data are sometimes also reported. The latter two are often estimated to understand how much the distribution deviates from a normal distribution, because most of the techniques described below rest on the assumption that key aspects of the data are normally distributed.

8.2 Student's T-Tests

When it comes to inferential statistics, one of the most popular tests in HRI studies is "Student's t-test" (or just "t-test"), which deals with the comparison between the means of two samples. A t-test compares two data sets and tries to answer the question of whether they came from the same population.[5] Usually your hypothesis is that your data sets did not come from the same population. The reason for this is that, if your experimental manipulation has an effect on a variable, it should create a distinct population from that of the control condition in terms of this variable; conversely,

[5]Formally, a t-test can also compare a single sample data set with a value, but we limit our discussion here to two-condition t-tests.

if it has no systematic effect on participants, then the observations in two conditions should behave as if they came from the same population.

T-tests are appropriate to use when you have exactly two groups to compare and your measure is a continuous variable that conforms to a normal distribution.[6] In our example, we could collect two data sets of how often people looked at the robot: one from people walking with a robot running our algorithm, and another of people walking with a baseline algorithm. Then we could use a t-test to compare the means of these data sets and see whether they are sufficiently different.

There are two different types of t-test commonly used in HRI research: An *independent sample t-test* applies to a situation where the two data sets are unpaired, which is typically the case in a between-participants design. A *pairwise t-test* applies to cases where there are pairs between two data sets that are related in some manner, such as two measurements coming from the same participant, with each measurement placed in one of the two data sets. An independent sample t-test evaluates the *difference between the means* of both groups and compares that difference to zero, i.e., $(\frac{1}{n_a}\sum a_i - \frac{1}{n_b}\sum b_j) \Leftrightarrow 0$, whereas a pairwise evaluates the mean of the differences between two observations in each pair and compares that mean to 0, i.e., $\frac{1}{n}\sum(a_i - b_i) \Leftrightarrow 0$. In the walking robot example, we would use an independent sample t-test if every participant only walked with one type of algorithm (i.e., a between-participants design), but a paired t-test if each participant experienced both algorithms (i.e., a within-participants design).

You may also decide whether you use a one-tailed or two-tailed test. Unless you have strong reasons to hypothesize a directional difference (e.g., the number of gazes with our new algorithm is *greater than* the number of gazes with the baseline algorithm), you should use a two-tailed test (e.g., gaze count with our algorithm is *different from* the that of the baseline algorithm), and make sure to include this decision in your pre-registration and research paper. Do not use t-tests when you have more than two conditions. This requires a test called "Analysis of Variance" (ANOVA).

The test itself involves calculating a so-called *t-value* or *t-ratio*, which is the ratio between the actual difference between the two means found in your sample and the difference you would expect just from random variation. The higher the t-ratio, the more confident you can be that your means are different due to something inherent in the conditions. The p value associated with your t-ratio is the probability that you would find a t-ratio as large or larger than the one you found only due to unrelated sample variation. The larger the t-ratio, the smaller the resulting p-value.

You generally report t-ratios and p-values along with descriptive statistics. Many publication venues expect researchers to report effect sizes in addition to precise p-values. For t-tests, one common effect size metric is Cohen's d, which is the difference between two means divided by a standard deviation measure for the data [15].

8.3 One-Way ANOVA

The one-way Analysis of Variance (ANOVA) test can be used to compare the means of more than two groups. For instance, if we were to compare across three conditions (one group of participants would walk with our algorithm, one with the state-of-the-art baseline, and one with a teleoperated robot), we would use a one-way ANOVA. Just like the t-test has a test statistic called the t-ratio, the statistic of a one-way ANOVA is the *F-ratio*. This is (roughly speaking) the ratio between the mean variance-between-groups and the mean variance-within-each-group. Again, the higher the F-ratio, the more different your study groups are. Therefore a higher F-ratio, for a given number of degrees of freedom, results in a lower p-value and suggests that the groups are more likely to be different. For F-tests, effect sizes are usually

[6] If you are dealing with non-normal distributions, a large enough sample size could allow you to use a t-test; otherwise, use non-parametric tests.

reported as partial eta-squared (η_p^2), which is the proportion of the variance in the outcome variable accounted for by the predictors, although some argue that omega-squared (ω^2) is a less biased, and thus preferred, effect size measure.[7]

In practice, one-way ANOVA is the multi-group equivalent of an independent t-test. It is not the correct test to use when data from different conditions are related, for example in a within-participants design. In this case you would use a repeated-measures ANOVA.

The F-ratio of a one-way ANOVA does not tell you which of the three or more groups are different from each other, only that they are not all the same with respect to the measured variable. To further understand the relationship between groups, you need to use either *planned contrasts* or ad-hoc *pairwise comparisons* between the three or more conditions. Planned contrasts are appropriate when you have specific hypotheses about the differences between certain conditions, so you plan these contrasts ahead of time and pre-register your plan before running your study. For example, you may hypothesize that there is a difference between the group who experienced your algorithm and both of the baseline groups (i.e., the group walking with a teleoperated robot and the group walking alongside the state-of-the-art algorithm), but you do not care whether there is a difference between these two baseline groups.

On the other hand, ad-hoc pairwise comparisons simply compare each of the conditions to every other group after running the ANOVA. When you run pairwise comparisons, you need to be careful to avoid the "multiple comparisons problem," because the more comparisons you make, the more likely that at least one comparison will reach your predetermined alpha level (say, .05) merely due to random sampling error. To combat this problem, you should use corrections, like the Bonferroni correction, or "honest" tests, such as Tukey's HSD (honestly significant difference) test. The details of these are outside of the scope of this primer; for more information, see Coolican [18].

A final note on the one-way ANOVA is that it is not appropriate when there are two or more predictor variables. This requires a factorial ANOVA.

8.4 Factorial ANOVA

A study designed to investigate the effect of two or more predictors simultaneously is also called a factorial design. In this case you need to use a factorial ANOVA. The factorial ANOVA methodically evaluates the effects of each predictor construct on your measured outcomes, as well as interactions among predictors.

One example of a factorial ANOVA is the two-way ANOVA. It has two factors, which results in three evaluations: the main effect of each factor and their interaction. In our example, you could ask what the main effect of the adaptive algorithm on trust is, what the main effect of dog ownership is, and what the interaction effect is.

When an interaction effect is detected, you should be careful in describing your main effects. In many cases, you may wish to know whether a factor has a significant effect on the outcome variable when the second factor is held at a specific level. For instance, you may wonder whether your new algorithm works better than a control algorithm for dog owners and for non-dog owners, respectively. To answer these questions, you need to test for "simple effects" by comparing your algorithm against the baseline algorithm at each given level of the dog ownership dog factor (see Section 14.3 in Gravetter *et al.* [35] for more information). Importantly, you should not describe your significant interaction as if you are describing two simple effects—you need to actually perform simple effects tests order to make such claims.

Sometimes, you may need to include more than two factors in your factorial design. For instance, a three-way ANOVA has three factors and thus seven evaluations: the main effect of each factor, each of the three two-way interactions, and

[7] https://daniellakens.blogspot.com/2015/06/why-you-should-use-omega-squared.html

one three-way interaction. The more factors you include, the more complicated it becomes to interpret interaction effects. Therefore, in practice, researchers rarely manipulate more than three factors in their studies.

8.5 Linear Regression

Constructing linear regression models is a useful alternative to the above-mentioned tests, and while it is increasingly common in some behavioral research communities, it is currently underutilized in the HRI literature. This approach is founded on the insight that t-tests, one-way ANOVAs, and factorial ANOVAs are all special cases of a general linear regression model. Using regression analysis can help ensure consistency in analyzing and comparing results across studies. It also offers a straightforward way to control for confounding variables.

The regression model is stated as a linear relationship between your predictors and your outcome with some exogenous "noise" ϵ:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_n x_{ni} + \epsilon_i \tag{1}$$

Here, x_{1i} is the value of the first predictor variable in the ith data point, x_{2i} the value of the second predictor variable in the ith data point, and so on; β_1 is the *coefficient* of the first factor, and so on; ϵ_i is the unmodeled "error" or "noise" for that data point; and β_0 is the so-called "intercept," representing a baseline value. The predictors x_{ki} can be either continuous or categorical. In the case where a predictor has two conditions, such as when one would usually use a t-test, the values of x are either 0 (for Group A) or 1 (for Group B). The coefficients then turn out to simply be the mean for Group A ($\beta_0 = \mu_A$) and difference between the group means ($\beta_1 = \mu_B - \mu_A$):

$$y_i = \mu_A + (\mu_B - \mu_A)x_i + \epsilon_i \ ; \ x_i \in \{0, 1\} \tag{2}$$

Two-way ANOVAs can be represented by two categorical variables, x_1 and x_2. To model interactions, you would multiply two factor variables and assign them a third interaction coefficient:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{int} x_{1i} x_{2i} + \epsilon_i \tag{3}$$

Here, β_1 represents the magnitude of the effect of our algorithm, and β_2 represents the magnitude of the effect of dog ownership, while β_{int} represents the interaction between these two factors.

To control for confounding variables such as, for example, the time of day a person participated in the study, you can add that time variable as an additional additive term in your linear model (for instance, coding morning = 0 and afternoon = 1), and mention that you "controlled for the effect of time" when you report your results. You generally do not report the coefficient related to the controlled variable. In the case of within-participants design, you can also construct a multilevel linear model to control for the interdependence among repeated measures within the same participant, which is generally recommended over repeated measures ANOVA given that it can accommodate more complex designs.

With this very brief introduction, we highly encourage the reader to further study this powerful and general tool, as it could make for more consistency in the statistical analysis used in HRI experiments. For information on how to run linear regression models, please see James *et al.* [43]. When repeated measures are employed, tutorials such as Judd *et al.* [45] and Singmann and Kellen [72] provide information and software listings for multilevel linear models.

8.6 Chi-Square Tests

If your outcome variable is categorical rather than continuous, you can use a chi-square (χ^2) test. In our example, you may ask all participants to choose which of the two algorithms they trust more. To find out whether more people prefer your algorithm over the other one, you would need a χ^2 goodness-of-fit test. This test examines how closely the distribution of your data matches the expected distribution of a population under an adversarial null hypothesis—for instance, the two algorithms are preferred by an equal number of people. In this case, the χ^2 test asks: how different is your data from an even distribution (a 50-50 split in the case of two levels)? If the actual frequency in each cell considerably deviates from what you would expect from the null hypothesis, you will get a large enough χ^2 value, which in turn would lead you to reject the null hypothesis (e.g., 50-50).

When you have a second variable—for instance, dog ownership—and you want to examine whether dog ownership influences people's algorithm preference, you will need a χ^2 test for independence to test the relationship between these two variables. The spirit of a χ^2 test for independence is the same as the goodness-of-fit test, except that now you compare the distribution of people's algorithm preference across levels of the second variable (i.e., dog owner vs. non-dog owner), and the χ^2 score describes the interaction between two variables. It is worth noting that you can also analyze binary outcome variables with logistic regression, but this is beyond the scope of this paper (see James *et al.* [43] for more information).

8.7 Rank Tests

In some cases, you cannot in good faith make the required assumptions to run the above-mentioned tests. For example, sometimes the distribution of your data severely violates the assumption of being normally distributed, or you used a non-uniform ordinal measure (one where the difference between adjacent values is likely to be unequal). In these cases, you need to use non-parametric tests.

One common example of non-parametric tests are rank tests, where you only compare the rank of outcomes rather than their numeric values. When you only have one predictor with two levels, you may use the Mann-Whitney U-test (a non-parametric equivalent of the independent sample t-test) or the Wilcoxon Signed Rank Test (a non-parametric equivalent of the paired t-test), depending on whether your two samples are independent or paired. The spirit of rank tests is very similar to that of the t-tests, except that instead of directly comparing the means of your outcome data in each group, you rank the outcome data and then compare the means of the rank order between two groups to determine if they come from the same population. Therefore, rank tests are more lenient.

Similarly, the Kruskall-Wallis H-test is the rank version of the one-way ANOVA test discussed above. It also loosens the requirement for normal distributions, and is useful in cases where you are interested in the comparative rank of your measured variables.

8.8 Assumptions of Statistical Tests

All of the above-mentioned statistical tests are only valid given certain assumptions. For example, ANOVA assumes independence of observations, a normal distribution of residuals, and homoscedasticity of variances. There exist statistical methods that check for violations of these assumptions. While it is beyond the scope of this paper to describe all of the assumption tests in detail, researchers should be aware about these assumptions and check them before using a given statistical test.

9 REPORTING RESULTS

Reporting the results in a complete and consistent fashion ensures that your readers can accurately evaluate your work and findings. Too often authors use unconventional or confusing presentation methods when reporting test results, or omit important details about their results or the way the results were calculated.

Just as completeness is important when reporting your hypotheses, constructs, and procedure, you need to report all statistical tests regardless of statistical significance. This, in combination with the pre-registration principle mentioned above, can help reduce "researcher degrees-of-freedom" such as cherry-picking and p-hacking [59].

9.1 Reporting Templates

To support consistency, it may be useful to provide some reporting templates for the common tests presented above. These are adapted from Field *et al.* [31]. Please refer to this source or other statistics references for the reporting of descriptive statistics such as the mean, standard error, standard deviation, effect sizes, and the determination of degrees-of-freedom.

Template for t-test: "To test Hypothesis **H1**, we ran an independent samples t-test. Consistent with our hypothesis, participants rated their trust in robots higher when the robot was running our adaptive algorithm ($M = 5.64, SD = 1.47$) compared to the baseline algorithm ($M = 4.86, SD = 1.62$), $t(102) = 2.54$, $p = .013$, $d = 0.50$" [8]

Template for one-way ANOVA: "To test Hypothesis **H5**, we ran a one-way ANOVA and found a main effect of robot algorithm on mean distance from the robot: $F(2, 147) = 7.27,\ p < .001,\ \omega^2 = .077$. Post-hoc multiple comparisons using Tukey HSD further showed that our algorithm ($M = 56.12cm,\ SD = 1.19$) resulted in higher distance than Baseline 1 ($M = 46.12cm,\ SD = 1.73$), $t = 3.23,\ SE = 0.31,\ p = .004$, and Baseline 2 ($M = 45.77cm,\ SD = 1.62$), $t = 3.39,\ SE = 0.30,\ p = .003$. The two baselines did not differ, $t = 0.17,\ SE = 0.30,\ p = .99$".

Template for linear regression: "To test Hypothesis **H2**, we ran a regression with the algorithm type and robot speed as predictors, controlling for participant fear of robots. Algorithm type positively predicted people's sense of safety score ($\beta = .39,\ t(178) = 2.71,\ p = .004$), while robot speed did not ($\beta = .002,\ t(178) = 1.07,\ p = .14$). We did not detect an interaction between the two factors ($\beta = .013,\ t(178) = 0.32,\ p = .42$)".

Template for χ^2 goodness-of-fit test: "To test Hypothesis **H3**, we ran a χ^2 goodness-of-fit test. Overall, we found that more people preferred the robot with our new algorithm over one with the baseline algorithm (65 vs. 39). $\chi^2(1) = 6.50, p = .011, r = .25$".

Template for Mann-Whitney U-test: "To test Hypothesis **H11**, we ran a Mann-Whitney U-test. Overall, we found that people brought the robot to more trips when it was installed with our new algorithm ($M = 5.64,\ SD = 2.82$) compared to the baseline algorithm ($M = 4.77,\ SD = 2.51$), $W = 1496,\ p = .04, d = 0.33$".

In many cases, presenting the descriptive statistics as a graph can enhance understanding of your results. Some researchers also opt to present inferential statistics results in a table rather than sprinkled throughout the paper narrative. That said, it is confusing and thus ill-advised to report the same results in two places in the same article.

9.2 P-Values, Significance, and Effect Sizes

Conventionally, hypothesis-testing studies use a p-value threshold of .05 to determine whether a hypothesis should be accepted or rejected. If the p-value was below the threshold, results were reported as "significant." For many years,

[8] Effect sizes can be reported in one of several ways. For t-tests, Cohen's d is the most common measure for effect size; others support using the correlation coefficient r or Rosenthal's *r-equivalent* as a more consistent measure, arguing that it can be compared across different tests, such as correlations, t-tests, ANOVAs, and regressions [31].

however, and more so recently, the use of the term "significant" for findings that cross this somewhat arbitrary p-threshold is pointed out as detrimental to research and reliable findings [47, 59]. Some authors have suggested a more stringent alpha level—for example, .005 instead of the traditional .05 (e.g., [6, 44])—yet this suggestion leads others to be concerned about inflated Type II errors [33]. At any rate, using the word "significant" can be misleading, and some scholars recommend to refrain from using this term altogether [41]. At the minimum, you should always report the specific p-values associated with your results.

Hand-in-hand with the reporting of a test statistic and a p-value should be the reporting of the *effect size* of your test. Keep in mind that p-values are designed to tell you if your result could be explained by random variation, not whether it is meaningful. With a big enough sample size, any difference in means can reach a p-value below any arbitrary threshold. Therefore, effect sizes are used to communicate how large of an effect you have found. Intuitively, the most straightforward effect size seems to be a comparison between the means. However, often times it is hard to evaluate whether a difference of 20 cm is large or small. Therefore, standardized effect size such as Cohen's d, partial eta-squared η_p^2 and omega-squared ω^2, r-values, and odds ratios are designed to remove the units and allow a more consistent evaluation of your effect size regardless of the scaling of the variables. For more information on how to calculate and report effect sizes, see Lakens [50].

You should think about the p-value of your test and your effect size as two separate markers; the first indicates whether you can say with some confidence that there is any effect of your predictor on the outcome, and the second is whether it is meaningful. For example, you may find that by using the new algorithm, people report an increase of 0.1 points on a 7-point trust scale. It can be argued that even if that effect is supported with $p = .001$, the practical value of this improvement is questionable.

Once you have measured your effect sizes, how should you interpret them? For some metrics, you can find recommendations on what is typically considered as a small, medium, or large effect (e.g., [15]). Although these recommendations are widely circulated, you should also take them with a grain of salt, because such recommendations can be highly dependent on specific research fields and specific statistical tests [19]. For instance, in psychology, a correlation coefficient of .50 is usually considered large, but in some physical sciences, a correlation coefficient of .99 is expected. Therefore, depending on your field, even a statistically small effect might be theoretically meaningful and important. Ultimately, the interpretation of your effect size is up to you and your peers as readers. Be that as it may, you should report effect sizes and address them in the discussion of your findings.

Finally, some researchers argue that given the limitations of *Null Hypothesis Significance Testing* (NHST, which covers most of the above tutorial) and the frequent misconceptions of p-values [36], we should abandon p-values entirely and switch to descriptive statistics, effect sizes, confidence intervals, and Bayesian statistics entirely [20, 26, 47, 79, 85]. Bayesian methods have several advantages over NHST, such as using prior probability distributions to incorporate information from previous studies, and the interpretability of posterior probabilities [49, 84]. In light of the ongoing debate and development in empirical disciplines, we concur that p-value "is just one of several heuristic cues available to the data analyst" [48], and we encourage our readers to explore Bayesian methods as the need arises.

9.3 Post-Hoc Exploratory Analysis

In many cases, there are ways to analyze the data that were not evident during the planning of the study. Sometimes the data suggest new insights and new ways to consider your constructs. In these cases, it is fine to add new analyses and tests. Because you have spent so much time thinking about the study and the data, you might have gotten new

ideas halfway through the project. Therefore, this kind of analysis is often interesting as it could suggest new directions and possibly pave the way for a follow-up empirical study (see: "Where do Hypotheses Come From?" in Section 3).

However, in each of these cases, these analyses should be reported as "post-hoc" or "exploratory" and come at the end of your research paper. Post-hoc findings should never be presented as confirming additional hypotheses.

10 LIMITATIONS

HRI is an interdisciplinary field, and its development has benefited from the wide diversity of academic interests and research practices. This primer focuses on a specific aspect of HRI research, namely that of hypothesis-driven experimental studies. That said, we recognize that the experimental approach only represents one research paradigm among many possible methodological approaches in HRI research.

For example, we do not provide an in-depth discussion of observational and qualitative research, research-through-design, and other practices that understand interaction as more dynamic and context-sensitive than experimental research usually allows for. These non-experimental approaches can often provide deeper, more ecologically valid, and more contextually relevant knowledge than experimental studies. As each type of research comes with a rich scholarly history and consequently a long list of practical recommendations, we encourage the reader to explore existing literature on these methods to gain a broader set of tools which they can combine to conduct the most effective research.[9]

Furthermore, new research tools are constantly developed and sometimes blur the line between qualitative and quantitative methods as they are traditionally defined. Such tools may enable researchers to apply quantitative analysis on rich, context-sensitive data. Consider textual analysis as an example, a technique widely used in qualitative research in order to provide valuable insights into people's subjective experiences [23, 34, 40]. Recent years have seen a proliferation of computational tools that enable researchers to conduct quantitative analyses on massive amounts of text. These tools vary from the well-established Linguistic Inquiry and Word Count (LIWC) program [78] to evolving natural language processing (NLP) and machine learning techniques such as topic modeling and word embedding (see Evans and Aceves [28]; also see Lee and Kolodge [52] for an example of using topic modeling to study trust in a self-driving car context). Similarly, new computer vision-based algorithms can be used to analyze people's nonverbal behaviors, such as facial expressions and body movements. Such tools allow researchers to harvest and analyze large amounts of data in the field, essentially avoiding many of the problems associated with small sample sizes in lab experiments. Although not elaborated in this tutorial, these quantitative-qualitative hybrids can help HRI researchers study how people perceive and interact with robots beyond laboratory contexts.

Therefore, our focus on the experimental approach should not be read as a lack of endorsement for non-experimental methods. As a field, we need to make sure that we do not weed out methodological diversity in the pursuit of methodological rigor, and take care not to impose the standards, research agendas, and methods from one empirical discipline onto the whole field.

A separate concern with the recommendations suggested in this paper is whether they can be pragmatically applied to HRI research or whether they are unrealistic ideals when it comes to human-participant studies with robots. In some social science fields, the only limitation on an experimental study design is the inventiveness of its authors and the cost of recruiting participants. In HRI, researchers have to deal with the additional hurdles of prototype technologies that do not always work, a lack of ecologically valid use-cases across diverse demographic populations due to the fact that robots are not yet widely deployed, a lack of established systems to recruit a large number of participants,

[9]For example, you may find more information on how to conduct interviews in Smith [74], and conduct observational studies in Adler and Adler [1] as well as in Coolican [18].

people's unfamiliarity with robots or their inaccurate presumptions about robotics, and sometimes a pressing conference submission deadline. Moreover, in many cases an evaluation study is only one of many components that a research group has to develop.

The position put forth in this paper is that, even given the above-mentioned constraints, researchers should be cognizant about best practices. Anyone running an experimental study should be able to recognize potential threats to the validity of their research design and try their best to remove those threats. When such issues cannot be realistically addressed in every HRI project, researchers should be forthcoming in the report about their decisions and discuss the potential limitations. In concert, reviewers and editors should respect that empirical research can be imperfect and that $p < .05$ is not the only criterion—and perhaps not even the most important one—that should decide whether a paper is valuable or should be published.

In addition, the methodological imperatives in this paper are not meant to discount the importance of past research. The HRI community has made great strides in the last few decades in a ground-breaking field. To continue on a solid foundation, however, it is time to self-reflect and improve the validity of our future findings.

11 CONCLUDING COMMENTS

This primer covered many topics, and the list of requirements and considerations may seem daunting at first, especially to someone just learning about experimental methods. We tried to provide a high-level overview of the various stages of empirical research, and hope we have not scared readers unfamiliar with these methods or deterred them from conducting experiments.

On the converse side, for readers experienced in empirical research, we are aware that we gave only a cursory view of many important concepts. In some cases we traded accuracy and completeness for readability. We have pointed to supplementary readings for those interested in familiarizing themselves with the details of the methods discussed here. Additional pointers for more information include the Open Science Teaching and Training Resources[10], the resources provided by the American Psychological Association's *Division 5: Quantitative and Qualitative Methods*[11], the statistics resources from the Society of Personality and Social Psychology,[12] and a useful R tutorial by Navarro.[13]

One of the motivations for this primer was to improve the reliability and credibility of HRI research going forward. In addition to providing more systematic and rigorous methodological training to the next generation of experimental researchers, we believe there are other important avenues to support this goal.

For one, we need more replication studies (see: "How can replications become more common in psychology?" in Shrout and Rodgers [70]). A replication crisis creates a fundamental problem of trust. When findings that have been widely taught and cited are shown to rest on flaky ground, it can cast serious doubt on previous research and undermine the public's trust in research studies in general.

So far, there has been relatively little systematic effort to examine whether the HRI field is subject to the same replication issues that have shaken up many other empirical disciplines, including psychology, medicine, neuroscience, life sciences, and economics [42]. Given the lack of systematic methodological training, the prevalence of small sample sizes, the common practice of measuring a large number of variables and selectively reporting a few, it seems likely that the question of replication is one our community must tackle.

[10] https://docs.google.com/spreadsheets/d/1kzJDrj3dtL9WOz_zRMEhgR7xxo9p3pGQOLJMUVkO1A0/edit
[11] https://www.apadivisions.org/division-5/resources
[12] http://spsp.org/resources/statistics
[13] https://learningstatisticswithr.com/book/

A promising avenue to address this question could be a large-scale research collaboration to replicate highly-cited studies in HRI, similar to the 2015 *Science* paper "Estimating the reproducibility of psychological science," led by the Open Science Collaboration [63]. Another thrust in this direction is already underway, with the recently added "reproducability" track in a leading HRI technical conference. Yet another avenue would be to create our own research tools to facilitate rigorous research practices in the HRI field [42], which may range from creating validated psychological instruments (e.g., the Robotic Social Attributes Scale (RoSAS), [12]) to creating community resources that enables systematic comparisons across different robots (e.g., the Anthropomorphic roBOT (ABOT) Database, [65]). Finally, we should also conduct more meta-analysis studies on key topics in HRI.

Learning from the discourse in other disciplines, we know that increasing the replicability of our findings requires researchers to spend more time and effort learning and adhering to updated methodological practices. In HRI, we have to increase sample sizes, separate confirmatory and exploratory analysis in our papers, and be more vigilant about common malpractices that might inflate the rate of false positives. These changes require the entire field to act collectively, as these practices have to be emphasized both in our own research and in the peer-review process.

As our field gains more interest, researchers in other areas and the general public look to us to provide accurate findings that shed light on how people interact with robots. With greater prominence comes greater responsibility to employ trustworthy empirical methods. We hope that this paper will inspire the next generation of HRI researchers to embrace empirical research with the goal of pushing forward the edge of our knowledge on human-robot interaction.

ACKNOWLEDGMENTS

We thank Alden McCollum, Jeremiah Zhe Liu, and Mark Brow for helpful feedback on an earlier version of the manuscript and Alap Kshirsagar for assistance on formatting the document. We also thank the anonymous reviewers for their insightful comments and suggestions.

REFERENCES

[1] Patricia A Adler and Peter Adler. 1994. *Observational techniques.* Sage publications, Thousand Oaks, CA.

[2] Samantha F Anderson, Ken Kelley, and Scott E Maxwell. 2017. Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological science* 28, 11 (2017), 1547–1562.

[3] Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature News* 533, 7604 (2016), 452.

[4] Marjan Bakker, Chris HJ Hartgerink, Jelte M Wicherts, and Han LJ van der Maas. 2016. Researchers' intuitions about power in psychological research. *Psychological science* 27, 8 (2016), 1069–1077.

[5] Marjan Bakker, Annette van Dijk, and Jelte M Wicherts. 2012. The rules of the game called psychological science. *Perspectives on Psychological Science* 7, 6 (2012), 543–554.

[6] Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. 2018. Redefine statistical significance. *Nature Human Behaviour* 2, 1 (2018), 6.

[7] 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.

[8] Douglas G Bonett and Thomas A Wright. 2015. Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size planning. *Journal of Organizational Behavior* 36, 1 (2015), 3–15.

[9] Serena Booth, James Tompkin, Hanspeter Pfister, Jim Waldo, Krzysztof Gajos, and Radhika Nagpal. 2017. Piggybacking Robots: Human-Robot Overtrust in University Dormitory Security. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI'17)*. Association for Computing Machinery, New York, NY, USA, 426–434. https://doi.org/10.1145/2909824.3020211

[10] Kenneth S Bordens and Bruce B Abbott. 2018. *Research design and methods: A process approach (10th edition).* McGraw-Hill, New York, NY.

[11] Mason Bretan, Guy Hoffman, and Gil Weinberg. 2015. Emotionally expressive dynamic physical behaviors in robots. *International Journal of Human-Computer Studies* 78 (2015), 1–16.

[12] 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 (HRI'17)*. Association for Computing Machinery, New York, NY, USA, 254–262. https://doi.org/10.1145/2909824.3020208

[13] 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.

[14] Scott Clifford, Ryan M Jewell, and Philip D Waggoner. 2015. Are samples drawn from Mechanical Turk valid for research on political ideology? *Research & Politics* 2, 4 (2015), 2053168015622072.

[15] Jacob Cohen. 1988. *Statistical power analysis for the social sciences*. Erlbaum, Hillsdale, NJ.

[16] Jacob Cohen. 1992. A power primer. *Psychological bulletin* 112, 1 (1992), 155.

[17] Renita Coleman. 2018. *Designing Experiments for the Social Sciences: How to Plan, Create, and Execute Research Using Experiments*. Sage publications, Thousand Oaks, CA.

[18] Hugh Coolican. 2017. *Research methods and statistics in psychology*. Psychology Press, Road Hove, UK.

[19] Joshua Correll, Christopher Mellinger, Gary H. McClelland, and Charles M. Judd. 2020. Avoid Cohen's 'Small', 'Medium', and 'Large' for Power Analysis. *Trends in Cognitive Sciences* 24, 3 (March 2020), 200–207. https://doi.org/10.1016/j.tics.2019.12.009

[20] Geoff Cumming. 2013. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge, New York, NY.

[21] Junhua Dang, Kevin M. King, and Michael Inzlicht. 2020. Why Are Self-Report and Behavioral Measures Weakly Correlated? *Trends in Cognitive Sciences* 24, 4 (April 2020), 267–269. https://doi.org/10.1016/j.tics.2020.01.007

[22] Kerstin Dautenhahn. 2007. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical transactions of the royal society B: Biological sciences* 362, 1480 (2007), 679–704.

[23] Maartje de Graaf, Somaya Ben Allouch, and Jan van Dijk. 2017. Why Do They Refuse to Use My Robot? Reasons for Non-Use Derived from a Long-Term Home Study. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI'17)*. Association for Computing Machinery, New York, NY, USA, 224–233. https://doi.org/10.1145/2909824.3020236

[24] Ewart J De Visser, Samuel S Monfort, Ryan McKendrick, Melissa AB Smith, Patrick E McKnight, Frank Krueger, and Raja Parasuraman. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied* 22, 3 (2016), 331.

[25] Edward Diener and Robert Biswas-Diener. 2019. The Replication Crisis in Psychology.

[26] Zoltan Dienes and Neil Mclatchie. 2018. Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic bulletin & review* 25, 1 (2018), 207–218.

[27] Victor DiFate. 2007. *Evidence*. The Internet Encyclopedia of Philosophy. https://www.iep.utm.edu/evidence/

[28] James A. Evans and Pedro Aceves. 2016. Machine Translation: Mining Text for Social Theory. *Annual Review of Sociology* 42, 1 (July 2016), 21–50.

[29] 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.

[30] Christopher J Ferguson and Moritz Heene. 2012. A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science* 7, 6 (2012), 555–561.

[31] Andy Field, Jeremy Miles, and Zoë Field. 2012. *Discovering statistics using R*. Sage publications, Thousand Oaks, CA.

[32] Jessica K Flake, Jolynn Pek, and Eric Hehman. 2017. Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science* 8, 4 (2017), 370–378.

[33] Jean Gaudart, Laetitia Huiart, Paul J Milligan, Rodolphe Thiebaut, and Roch Giorgi. 2014. Reproducibility issues in science, is P value really the only answer? *Proceedings of the National Academy of Sciences* 111, 19 (2014), E1934–E1934.

[34] Barney G Glaser and Anselm L Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge, Abingdon-on-Thames, UK.

[35] Frederick J Gravetter and Larry B Wallnau. 2020. *Essentials of statistics for behavioral sciences*. Cengage Learning, Boston, MA.

[36] Sander Greenland, Stephen J Senn, Kenneth J Rothman, John B Carlin, Charles Poole, Steven N Goodman, and Douglas G Altman. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology* 31, 4 (2016), 337–350.

[37] Andrew F Hayes. 2017. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Publications, New York, NY.

[38] Carl G. Hempel. 1966. *Philosophy of Natural Science*. Prentice-Hall, Englewood Cliffs, N.J.

[39] Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. Most people are not WEIRD. *Nature* 466, 7302 (2010), 29.

[40] Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research* 15, 9 (2005), 1277–1288.

[41] Stuart H. Hurlbert, Richard A. Levine, and Jessica Utts. 2019. Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires. *The American Statistician* 73, sup1 (March 2019), 352–357. https://doi.org/10.1080/00031305.2018.1543616

[42] 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 (HRI'18)*. Association for Computing Machinery, New York, NY, USA, 13–20. https://doi.org/10.1145/3173386.3173389

[43] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An introduction to statistical learning*. Vol. 112. Springer, Heidelberg, Germany.

[44] Valen E Johnson. 2013. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences* 110, 48 (2013), 19313–19317.

[45] Charles M Judd, Jacob Westfall, and David A Kenny. 2017. Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual review of psychology* 68 (2017), 601–625.

[46] Norbert L Kerr. 1998. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review* 2, 3 (1998), 196–217.

[47] Rex B Kline. 2013. *Beyond significance testing: Statistics reform in the behavioral sciences*. American Psychological Association, Washington, DC.

[48] Joachim I Krueger and Patrick R Heck. 2017. The heuristic value of p in inductive statistical inference. *Frontiers in Psychology* 8 (2017), 908.

[49] John K Kruschke. 2010. What to believe: Bayesian methods for data analysis. *Trends in cognitive sciences* 14, 7 (2010), 293–300.

[50] Daniël Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology* 4 (2013), 863.

[51] Vincent Larivière, Stefanie Haustein, and Katy Börner. 2015. Long-Distance Interdisciplinarity Leads to Higher Scientific Impact. *PLOS ONE* 10, 3 (March 2015), e0122565. https://doi.org/10.1371/journal.pone.0122565

[52] John D. Lee and Kristin Kolodge. 2019. Exploring Trust in Self-Driving Vehicles Through Text Analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 62, 2 (Sept. 2019), 260–277. https://doi.org/10.1177/0018720819872672

[53] D Stephen Lindsay. 2015. Replication in psychological science.

[54] David P MacKinnon, Amanda J Fairchild, and Matthew S Fritz. 2007. Mediation analysis. *Annu. Rev. Psychol.* 58 (2007), 593–614.

[55] Maya B Mathur and David B Reichling. 2016. Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition* 146 (2016), 22–32.

[56] Chad R Mortensen and Robert B Cialdini. 2010. Full-cycle social psychology for theory and application. *Social and Personality Psychology Compass* 4, 1 (2010), 53–63.

[57] Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie Du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John PA Ioannidis. 2017. A manifesto for reproducible science. *Nature human behaviour* 1, 1 (2017), 0021.

[58] Kevin R Murphy and Herman Aguinis. 2019. HARKing: how badly can cherry-picking and question trolling produce bias in published results? *Journal of business and psychology* 34, 1 (2019), 1–17.

[59] Leif D Nelson, Joseph Simmons, and Uri Simonsohn. 2018. Psychology's renaissance. *Annual review of psychology* 69 (2018), 511–534.

[60] Richard E Nisbett and Timothy D Wilson. 1977. Telling more than we can know: verbal reports on mental processes. *Psychological review* 84, 3 (1977), 231.

[61] Brian A Nosek, George Alter, George C Banks, Denny Borsboom, Sara D Bowman, Steven J Breckler, Stuart Buck, Christopher D Chambers, Gilbert Chin, Garret Christensen, et al. 2015. Promoting an open research culture. *Science* 348, 6242 (2015), 1422–1425.

[62] Brian A Nosek, Charles R Ebersole, Alexander C DeHaven, and David T Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences* 115, 11 (2018), 2600–2606.

[63] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), aac4716.

[64] Gabriele Paolacci and Jesse Chandler. 2014. Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science* 23, 3 (2014), 184–188.

[65] 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 (HRI'18)*. Association for Computing Machinery, New York, NY, USA, 105–113. https://doi.org/10.1145/3171221.3171268

[66] Russell A Poldrack, Chris I Baker, Joke Durnez, Krzysztof J Gorgolewski, Paul M Matthews, Marcus R Munafò, Thomas E Nichols, Jean-Baptiste Poline, Edward Vul, and Tal Yarkoni. 2017. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience* 18, 2 (2017), 115.

[67] 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.

[68] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. 2016. Overtrust of Robots in Emergency Evacuation Scenarios. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction (HRI'16)*. IEEE Press, New York, NY, USA, 101–108.

[69] Norbert Schwarz. 1999. Self-reports: how the questions shape the answers. *American psychologist* 54, 2 (1999), 93.

[70] Patrick E Shrout and Joseph L Rodgers. 2018. Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual review of psychology* 69 (2018), 487–510.

[71] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* 22, 11 (2011), 1359–1366.

[72] Henrik Singmann and David Kellen. 2020. An introduction to linear mixed modeling in experimental psychology. In *New Methods in Cognitive Psychology*. Psychology Press, Road Hove, UK. http://singmann.org/download/publications/singmann_kellen-introduction-mixed-models.pdf, preprint

[73] Paul Smaldino. 2019. Better methods can't make up for mediocre theory. *Nature* 575, 7781 (Nov. 2019), 9–9. https://doi.org/10.1038/d41586-019-03350-5

[74] Jonathan A Smith. 2015. *Qualitative psychology: A practical guide to research methods*. Sage Publications, Thousand Oaks, CA.

[75] Stevens S Smith and Deborah Richardson. 1983. Amelioration of deception and harm in psychological research: the important role of debriefing. *Journal of Personality and Social Psychology* 44, 5 (1983), 1075.

[76] Keith E Stanovich. 2013. *How to Think Straight About Psychology (10th edition)*. Pearson, Boston, MA.

[77] Neil Stewart, Jesse Chandler, and Gabriele Paolacci. 2017. Crowdsourcing samples in cognitive science. *Trends in cognitive sciences* 21, 10 (2017), 736–748.

[78] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[79] D Trafimow and M Marks. 2015. Editorial in Basic and Applied Social Pschology. *Basic and Applied Social Pschology* 37 (2015), 1–2.

[80] Daniel Ullman and Bertram F. Malle. 2019. Measuring Gains and Losses in Human-Robot Trust: Evidence for Differentiable Components of Trust. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI'19)*. IEEE Press, New York, NY, USA, 618–619.

[81] Maaarten van Casteren and Matthew H Davis. 2006. Mix, a program for pseudorandomization. *Behavior research methods* 38, 4 (2006), 584–589.

[82] Thea F Van de Mortel et al. 2008. Faking it: social desirability response bias in self-report research. *Australian Journal of Advanced Nursing, The* 25, 4 (2008), 40.

[83] Anna Elisabeth van't Veer and Roger Giner-Sorolla. 2016. Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology* 67 (2016), 2–12.

[84] Eric-Jan Wagenmakers, Maarten Marsman, Tahira Jamil, Alexander Ly, Josine Verhagen, Jonathon Love, Ravi Selker, Quentin F Gronau, Martin Šmíra, Sacha Epskamp, et al. 2018. Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic bulletin & review* 25, 1 (2018), 35–57.

[85] Eric-Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, and Han LJ Van Der Maas. 2011. Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011).

[86] Adam Waytz, Joy Heafner, and Nicholas Epley. 2014. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology* 52 (2014), 113–117.

[87] Timothy D Wilson, Elliot Aronson, and Kevin Carlsmith. 2010. The art of laboratory experimentation. *Handbook of social psychology* 1 (2010), 51–81.

[88] Haotian Zhou and Ayelet Fishbach. 2016. The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of personality and social psychology* 111, 4 (2016), 493.