Advice to new Human-Robot Interaction researchers

Tony Belpaeme

Abstract As the field of Human-Robot Interaction is relatively young and highly interdisciplinary, it often happens that we as researchers make mistakes which could have been avoided had we known more about good and bad research practices in fields other than our own. Bad practices such as convenience sampling, *p*-hacking, or the Hawthorne effect will be known to some, but are too often unfamiliar to others. HRI is lucky to mature during one of the biggest revolutions in experimental psychology: the "replication crisis" was one of the most seminal moments in psychology and its repercussions are felt far and wide, including in HRI. This chapter lists some of the important mistakes often made in HRI studies and suggest practical recommendations for better research studies.

1 Remember your first HRI study?

Imagine, your first day as a HRI researcher. You probably are a computer scientist, psychologist, engineer, sociologist, or designer. Or you might have a background in another discipline relevant to HRI work, HRI draws on many of the science and engineering disciplines. You will perhaps start by reading up on the vast and diverse literature in HRI, but will soon move on to getting your hands dirty. You might want to know what people's attitudes are towards robots. You might want to design a machine learning solution to read social cues from a range of sensors mounted on a robot. You might want to know how effective a social robot is in supporting children with a long-term medical condition. You could be working on a collaborative robot which applies glue to plastic parts that a human worker then fits onto a car. Or you might design the interaction of a companion robot for elderly users. Unless you are

Tony Belpaeme
Ghent University, IDLab – imec, Technologiepark 126, B-9052 Ghent, Belgium
University of Plymouth, Drake Circus, PL4 8AA Plymouth, United Kingdom
e-mail: tony.belpaeme@ugent.be

a philosopher, there will probably come a time when you want to know how well your robot works and when you want data to guide your efforts. In a nutshell: it is time for you to run a study.

You probably already have a good idea of what your colleagues would consider to be good work. If you are an engineer, you might value a live demonstration of your system. You demonstrate how the robot recognises you and responds to your spoken commands, and you use a video recording of your robot system to show the world what technical feats you've accomplished. If you are an experimental psychologist, you know your colleagues value carefully controlled experiments in which dozens of balanced participants interact with one of two or more conditions and where you carefully measure a number of outcomes. If you're a sociologist, you might have designed a survey which you send out online to capture data on how people's attitudes to robots differ across age groups. Each discipline and industry has its preferred ways of collecting and communicating research and design efforts, and during our education at university we have been trained to use these effectively within our discipline. Designers will be skilled at presenting, psychologists at measuring, sociologists at interviewing. However, it is unlikely that as an inexperienced HRI researcher without formal training in HRI –as there are few training programmes in HRI– you will be versed in a wide range of study methods bearing on the field. And just as the author of this chapter, you will probably make some beginner's mistakes. Some easy to avoid, but some serious enough to get your six-month research study rejected by your peers or to irreparably damage a roll-out of a commercial product. This chapter contains some of the most prevalent and fundamental errors committed in HRI studies, and suggests potential solutions.

2 Current practice in HRI studies

While many of the mistakes made in HRI studies are not unique to HRI (other disciplines are not perfect either), their prevalence is often higher due to the interdisciplinary nature of HRI [2]. This section will draw upon an analysis of three years of HRI studies published between 2013 and 2015 in the IEEE/ACM Conference on Human-Robot Interaction, the main conference in the field [2].

2.1 Lab-based or in the wild?

To help us understand the lay of the land in HRI research, it will be helpful to first define the different types of studies. A *lab study* requires participants to come to a specific location, a lab at a university or company, to take part in the study. A *non-lab study*, also known as an in-the-wild study, is one where the study is run on location, for example in a factory or a hospital. Almost three quarters of HRI studies are lab-based [2]. While there certainly are good reasons for conducting studies

in a carefully controlled lab environment, there are questions about their ecological validity [5]. Admittedly, experimental psychology often uses lab-based experiments to better understand human cognition, and the results obtained from lab experiments do indeed shed light on the functioning of cognition and the interplay of cognition with the external world. But in HRI the purpose is often not to elucidate human cognition, but rather to assess whether one approach to HRI differs from another. Is a robot that offers breaks to young learners more effective as a tutor [23]? Do robots have a role to play in Autism Spectrum Disorder therapy [25]? These questions are often better answered in ecologically valid settings, as what might appear to be effective in a controlled lab environment, is likely to be washed out by the buzz and noise of the real world. Still, there are good reasons for running lab-based studies. Sometimes the technical setup just does not travel well. It might contain an expensive robot which cannot easily be moved out of the lab [6] or the sensory rig is too cumbersome to dismantle, build and calibrate [11]. An attractive middle ground is the living lab, a semi-naturalistic environment in which conditions of natural environments are replicated. This can range from a single room, such as a living room, to an entire house. These environments allow for complex technological setups, while offering a perhaps more relaxed environment in which the behaviour of the user can be more natural. At any rate, the decision between a lab, living lab or non-lab evaluation environment should be carefully considered, and when resources and technology allow, preference should be given to the latter.

2.2 Wizard of Oz or full autonomy?

A second dimension along which studies differ is the level of autonomy that the robot has. Sometimes the interaction with the robot will run autonomously, apart from perhaps the robot being started or stopped by the experimenter. Sometimes, the robot will to some extent be controlled by another human, a method known as Wizard of Oz (WoZ) [24]. Key here is that the participant is unaware of this: to the user it appears as if the robot is fully autonomous, while in reality a remote operator, called the *wizard*, takes over some aspects of the robot's functionality. The wizard can take over perceptual and cognitive aspects of the control. When taking over perceptual aspects, the wizard fills in for the lack in perceptual abilities of the robot, such as speech or vision, due to the technology not being sufficiently robust or due to time and resource constraints in implementing autonomous perception on the robot. If the wizard handles cognitive aspects, it means they make decisions, for example on how the robot should respond to the user. Wizarding can even just serve as a "stub" for functionality that is not yet sufficiently mature, or for aspects of the robots functionality that you which to trial before moving an expensive implemention. About 40% of studies where people interact with a robot use autonomous robots, all other to some extent require the assistance of a human operator. When reporting your research, it is important to mention how the robot was controlled. While it might of course necessary to use a smoke and mirrors approach during the

experiment, the eventual report should disclose fully to what extent the robot was wizarded. Further reporting guidelines for WoZ experiments are available [24].

2.3 On-screen or real robot?

The type of exposure to the robot is also key in a study. In some HRI studies people will interact with a real robot, while in other studies they will see a robot on-screen, either as a still or as a video. In some of these the robot will be shown on its own, in others people will see an interaction unfold between one or more users and the robot. The temptation of using on-screen presentations of robots are many. A photo or video does not crash midway the experiment. You often don't need to program or even own a robot, but instead can use a photo [22] or an illustration [14] of a robot instead. And finally, and perhaps most importantly, using an on-screen presentation of a robot allows for setting up the study online, thereby giving you access to a potential participant pool of hundreds of thousands of geographically spread respondents. There are cases where the reasons for using an on-screen presentation outweigh the effort of using an actual robot, but as interacting with a real embodied robot is a more real and more visceral experience, in which the user is more invested in the experience [18], it is reasonable to expect that a study involving a real robot will in most cases result in more useful results, perhaps even different results compared to seeing pictures of robots.

2.4 Convenience sampling or representative sampling?

When running HRI studies at a university there is a steady supply of willing (or readily coerced) participants at hand: the students. Participants that are easily recruited -for example students or visitors to an exhibition– make up a *convenience sample*. Convenience samples are rife in HRI, at between 2012 and 2015 when research was not conducted with children (aged less than 18) or the elderly (aged over 65), 87% of studies reported at the HRI conference used samples which drew from university populations, see fig. 1. It might be that some intended to study how well-educated, technology-savvy people in their late teens and early twenties respond to robots, but it is unlikely that all wanted to do so. Collecting data using a convenience sample introduces a sample bias, and the results obtained in this way are unlikely to translate to the real world. Convenience samples do have a place in experimental work, but given that HRI studies often require a subjective response to the robot, it is unlikely that a biased sample of students or colleagues will provide good data on which to build your Human-Robot Interaction. Therefore, efforts should be taken to go out and collect data from the range of users who eventually are expected to interact with your robot. Creating an unbiased and balanced participant pool is difficult,

and is further compounded by the fact that robots are new to most people and this first time exposure to a robot might influence the results (see *novelty effect*).

Another issue to take into consideration is the size of the sample. Collecting data from real participants can be resource intensive and time consuming, so there is a temptation to go for a low number of participants. The exception to this is when participants are plentiful, such as with crowd sourced studies. In an analysis of three years of HRI studies shows that sample sizes in HRI usually are very small when judged by standards of other fields, see fig. 2. Small sample sizes lead to a lack of statistical power, which in turn lead to incoherent results, a concern often voiced in psychology [19]. The recommendation would be to go for larger and more representative samples, which unfortunately is easier said than done, as running studies requires considerable technical, logistic and staffing effort. In some cases, larger samples might not even be available, for example when working in a clinical context. The lack of quantity when it comes to data is not necessarily a problem, and single case studies can be informative, especially in early stages of research. However, the ambition should be to use large sample sizes in ecologically valid studies. HRI evaluations are not for the lazy.

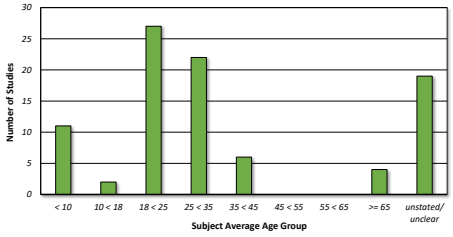

Fig. 1 The age of participants (if at all reported) in three years of studies presented at the HRI conference. A convenience sample of students is over-represented.

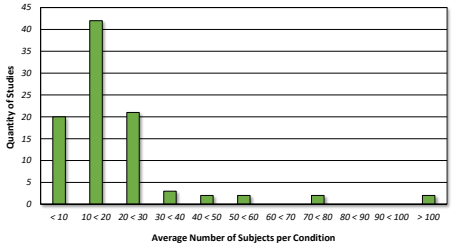

Fig. 2 The sample size in three years of studies presented at the HRI conference. Most studies use a very low number of participants. Data from [2].

2.5 Single or long-term interaction?

In most HRI studies participants only interact once with the robot (of 96 studies reported in [2] only 5 consisted of more than one interaction). This of course has rather profound implications for the field. As participants often have never interacted with a robot before and are typically naïve to the interaction under study, there is a *novelty effect*. The robot, the study setting or the interaction is new and unfamiliar to the participants, and it is likely that this will colour their interaction with the robot. Some outcomes might be stronger due to the novelty effect, while in some case the novelty effect might have an adverse effect on the outcomes. For example, young users might be intimidated by the robot, an effect which might wear off with repeated exposure. Long-term studies, in which the interaction last longer than a brief session or in which people interact with the robot over repeated sessions, are still relatively rare in HRI. Still, the value of doing long-term interaction research cannot be understated. While sometimes the novelty effect is actively sought (for example in entertainment applications), HRI is generally concerned with how people will interact with robots in day-to-day life and what the technical, societal and psychological consequences and applications are. As such there is a strong desire to know how the user's behaviour will evolve over time once the novelty effect wears off. It is difficult to say what a long-term interaction exactly is, but suffice to say that the novelty effect with regards to the robot, the interaction and the environment should have disappeared [17].

3 Setting up a study

Translating an appealing research idea into a well executed study almost always involves more than you think. Luckily many have gone before you and excellent introductions exist on how to set up experimental studies, with perhaps the best and accessible advice coming from experimental and social psychology [10, 12]. However, even when following these introductory guides, it is still very much possible to slip up, especially if you did not receive any formal training in experimental methods. In this section we look at new trends and a selection of problems and questionable practices that occur from time to time in HRI research. These are not necessarily unique to HRI research, and there are many exemplary HRI studies which manage to avoid all of these issues. But often, either through a lack of knowledge or experience, or through the context in which the study is set, the study inadvertently falls foul of one or several of these problems.

3.1 Null-Hypothesis Significance Testing

If quantitative data is collected during a study, there will be a need to compare data. Most HRI studies, just as in other experimental sciences, adopt the practice of using Null-Hypothesis Significance Testing (NHST). NHST needs a minimum of two conditions (the intervention and control, in analogy to the medical sciences where patients are given a treatment which is compared against no treatment or an established treatment, called the control). In each condition, one or more outcomes are measured. NHST then checks the hypothesis that the data distribution from the intervention conditions (consisting of sample size, mean and standard deviation when the data has a normal distribution) does not differ from the distribution of the control condition. This is called the null hypothesis. If the hypothesis can be "rejected", then the result is called "significant". Rejection is based on a statistical test, such as the Student's t-test, which return a p value. This value is the probability that the difference you see between the two conditions is due to chance. If the p-value is less or equal to some threshold, for example $p \leq 0.05$, it is safe to conclude that the difference is significant. The large majority of HRI studies will use NHST and will report p-values to support their conclusions.

However, in recent years NHST has come under fire [26]. A first problem is the term "significant", as it gives the uninitiated the impression that the results are important. Unfortunately, significant in a statistical context really does not mean that the results are important, substantial or major. It just means that it is highly unlikely that the difference we see between the two conditions is a coincidence. Even if the difference is very small, the results could still be called statistically significant if the comparative statistic measures say so. Imagine you are testing two versions of a robot –a polite robot and a direct robot– that delivers parcels in an office building. You do 100 runs of each condition. The polite robot has a mean time to finish its delivery round of 82 minutes (standard deviation = 8.0 minutes). The second robot has a mean of 82 minutes (standard deviation = 4.0 minutes). If you calculate the two-tailed t-test, the p-value is 0.0265, so the difference in mean time to complete a delivery round is statistically significant. But does it really matter? Honestly? Too often statistically significant results are presented as scientifically significant, while they are not. While this is a problem of how we communicate results, the other problems with NHST are more profound.

The first problem is the probability of rejecting the null-hypothesis being wrong. Typically, we assume that the p-value needs to be less than 0.05 before a result is called significant. This number is however arbitrary, and different scientific fields will use different threshold values (physics often uses 0.01 or 0.001). Determining whether a result is significant based on a random threshold seems at odds with the precision we usually expect from our scientific methods. The second problem is that p-values tend to fluctuate between repeated experiments. You would expect that when an experiment is repeated, the p-value would be similar. Unfortunately simulations have shown that p-values tend to be very unstable [9]. An experiment that first shows significant results is very likely to not be significant when run a second time!

4 The new statistics

As NHST and the reporting of p-values has been shown to be problematic [9, 26], what alternatives are there? A first recommendation is to report descriptive statistics of any data, a practice too often ignored in HRI studies. Descriptive statistics should at a minimum include the size of the sample, the mean, the standard deviation and details of the population (mean, range and standard deviation of age, and gender distribution) and details on how participants were recruited.

When comparing results, we're tempted to go for a t-test or for when comparing more than 2 groups, for an ANOVA and post-hoc test. But given the problems with p-values described above, it might be worth considering alternatives. Kaptein and Robertson [15] propose a number of solutions. One is to use Bayesian statistics, which computes the probability that a hypothesis are true. A Bayesian t-test computes the ratio of the likelihood of two competing hypothesis, for example the null hypothesis and an alternative hypothesis. Various tools exist, such as the BayesFactorPCL package for R or the free JASP tool, making the use of Bayesian statistics rather less of a pain than a few years ago. Still, despite the ease with which Bayesian statistics can now be computed, traditional approaches to statistics are here to stay.

When using traditional statistics, next to reporting p-values, it is very much worth reporting effect sizes as well. All too often there is a lot of enthusiasm for a p-value less than 0.05, without asking critical questions about the effect size of the result. If there is a difference between two conditions, how big is that difference and does it actually matter? Significance, as described above, can be a coincidence, but can equally be the result of large sample sizes. If, for example, responses are collected using crowd sourcing, where hundreds of responses can be collected in a matter of hours, it is likely that a significant effect will be found through the sheer size of N (i.e. the sample size). Therefore it is important to critically look at the effect sizes, and always mention these together with p-values.

Perhaps the most important advice, apart from using the most appropriate reporting of statistics, is to make use of the now available options to plot data. Gone are the days where the printer only could handle a simple black-and-white bar chart with error bars and an asterisk to mark significance. Modern tools, such as the Python seaborn library, can be used to create plots capturing all the richness of experimental data. Observations can be plotted alongside bar charts, violin plots show the density of the data, evolution of participants can be traced with hairlines. A plot can now capture the richness and complexity of data which is all to often lost through only using descriptive statistics. And with online papers now the norm, there is no reason to hold back on using colour to convey information (while keeping in mind that some of your readers might be visually impaired).

4.1 Selective publication of data

The temptation to only report the data that supports your agenda has always existed. Imagine, you set up a study to show how a robot can make a positive contribution when used as a tutor in an education context [3]. After building the setup and painstakingly programming the robot to teach children mathematics, you take the robot into a class room and when comparing the results of the pre test and post test, the children turn out to have learnt nothing. Now, the temptation might be to look for a silver lining in your study. If the children didn't learn, they might still have enjoyed the robot? So you reach for your questionnaire data in which children reported how much they liked the robot, next to answering a number of other questions on children's attitudes towards robot. And indeed, the children self-report to have enjoyed the robot very much. You decide to report this as your main results, and ignore the fact that the learning effect was insignificant.

Why is this selective use of evidence a problem? Surely, the fact that the children enjoy working with the robot is worth reporting? However, there are two problems in our example. One is that the results pertaining to the hypothesis on which the study was based are not reported, pretending that the study was intended to test something else. The second is that your colleagues might set up a similar experiment, and not knowing about your negative result, they cannot build upon your work. Luckily cherry picking and selective use of data are relatively innocent in HRI, and often occur through ignorance or a desire to put forward the most interesting results, rather than a conscious manipulation of data. Unlike other fields, such as climate studies or medicine, it is very unlikely that biased or selective reporting of data in HRI will cause ecological or personal harm. Still, in days of online sharing of data and results, the page limit of a paper really shouldn't be a reason to withhold data. Open data initiatives abound and a paper can be accompanied by an online data repository, sharing not only data, but extended information on methods and perhaps even code related to the experiment and data analysis.

4.2 The Hawthorne effect

The Hawthorne effect is named after the Hawthorne Works, a telephone equipment factory outside Chicago, where in 1932 the management brought in psychologists to see whether workers could be made more productive by making changes to the working environment [16]. Different aspects where varied, for example, the amount of light was varied with the assumption that good lighting should increase productivity. What the psychologists found was that no matter the changes to the environment, productivity increased. Whether lighting was increased or decreased, productivity improved. In the end, it turned out that it was the presence of the psychologists that increased productivity in the workforce, not the changes to the working environment. In its broadest sense, the Hawthorne effect refers to a change in the behaviour of people because they feel observed. Whether they are actually observed or not is

irrelevant, the mere belief of being observed is often sufficient to change people's behaviour. Although the original study has been criticised [1], the concept of the Hawthorne effect is still very relevant.

In the context of HRI research, the Hawthorne effect becomes acutely relevant. In research studies participants are often aware they are taking part in a study, for example through being recruited or through signing a consent form. Even if no experimenters or video equipment is visibly present during the study, the mere fact of taking part in an experiment will already change the participant's natural behaviour and responses. This often leads to unexpected results or the changed behaviour of the participants washes out small effects [13].

One way to manage the Hawthorne effect is to not let people know they are being observed. This is sometimes possible in natural settings. Nomura et al. [20] observed how children "bullied" a robot operating in a shopping mall. The bullying behaviour would likely not have occurred had the children known they were being watched. In a lab environment, one can try to use deception to let the participant believe the study is about something else, while the relevant aspect of the study is only revealed at the end. Care needs to be taken with deception or with not informing people they are part of a research study, and many ethics committees, often known as the Institutional Review Board (IRB) in US institutions, will frown upon the use of deception where it can at all be avoided. Still, the persistence of the Hawthorne effect does justify the judicious use of deception or naturalistic observation.

In HRI there is an additional complication when using social robots. It is not only the belief that one is being observed by an experimenter that impacts behaviour, but the social presence of the robot itself might also change people's behaviour. This is known as social facilitation, or audience effect, and states that people's performance will change in the presence of others as compared to their performance when alone. Typically, people tend to perform better at simple tasks or tasks they are skilled at, and worse at new or hard tasks. In HRI the social facilitation is not caused by other people being present, but could be caused by an artificial social agent, i.e. the robot, being present [27].

Overall, it is safe to assume that the presence of a social robot or the belief that one is being observed will change people's behaviour. This is not necessarily a bad thing, in some applications you intend for people to feel watched. If a robot is used to encourage people to choose healthy snacks over chocolate or use the stairs instead of the lift, then feeling watched is exactly what promotes healthy behaviour. An effect which might only be increased by feeling observed by the experimenters as well.

4.3 Crowdsourcing data

Collecting a large number of responses in HRI is both time intensive and expensive, and getting access to participants typically unavailable in or near your institution, such as participants from different cultures or geographical regions, is difficult. Crowdsourcing provides a cheap and convenient alternative to quickly and cheaply

collect responses. Once the study is set up an running, it often takes a matter of hours for hundreds of responses to come in at the price of just a few hundred US dollars. Some HRI studies will rely on online crowdsourcing platforms to collect experimental data, the best known of which is Amazon Mechanical Turk (often abbreviated to AMT or MTurk). Others exist as well, such as Figure Eight, Clickworker, or Microworker, but crowdsourcing platforms tend to pop up and disappear all the time.

Ever since crowdsourcing became available in 2005, the quality of crowdsourced research data has been discussed. Overall, the consensus seems to be that crowdsourcing allows you to reach more diverse participants, that the quality of the data is relatively good as long as a correct financial reward is given, and that data overall is as reliable as data collect using traditional methods [7, 4]. Results from experimental psychology obtained by using traditional methods are often replicated using crowdsourced data, convincingly showing the robustness of the method.

However, careful screening is needed of data and participants, and all too often in the excitement of setting up a study this aspect is glossed over. Several methods exist to identify participants and responses that do not meet quality criteria. Meta data of the participant's work, such as completion time and depth of responses, can be used to filter participants. Crowdsourcing platforms will also allow for the setting of a threshold on who can take part, giving you the option to only allow participants who have shown to provide high quality data in the past to take part in the study. Checker questions and gold questions are also an important tool in filtering out bad data: these are question to which there is an indisputably correct answer. What is the colour of the robot's shirt in the video? The robot mentioned a number, what was that number? If the worker does not answer correctly to these questions, his data should be discarded. Finally, the remuneration and incentives are important. If the task is enjoyable, the interface is easy to understand and the pay attractive, you can expect to get better data. Still, expect the have to throw away a significant amount of crowdsourced data. In some HRI studies, up to 50% of collected data had to be thrown out.

With HRI there is another issue which needs to be carefully considered. Crowdsourcing only allows for on-screen presentation of stimuli, and often the interactive character is severely limited. While a lab or field experiments usually involved a real, live encounter with a robot, this is very impossible to replicate on a crowdsourcing platform. Nevertheless HRI studies do frequently use crowdsourcing platforms to collect responses about interactions and design aspects of robots, but one should realise that an on-screen presentation of robots is a poor second to interacting or seeing an actual robot.

4.4 The replication crisis

The *replication crisis* refers to a key moment in science, starting in social and experimental psychology but spreading to other fields of science, including medicine,

where it was realised that a worryingly large number of positive research results were difficult to replicate or reproduce. Results which seemed solid enough to be part of student textbooks and popular science books, suddenly could not be reproduced when the original experiments were repeated by independent researchers or even by the original researchers. In 2011 a string of scandals tore through the psychology world, when fraudulent scientists and questionable results were exposed in quick succession [21]. The case of social psychologist Diederik Stapel was particularly alarming. Stapel's research had seemed solid and which often caused quite a sensation, such as the study that showed that meat eaters are more selfish than vegetarians, suddenly was revealed to be based on fabricated or manipulated data. 58 publications by Stapel and his co-workers were withdrawn as they were based on suspected or proven fraudulent data.

This sparked a critical appraisal of psychology research, and research in general, leading to establishment of the Open Science Foundation. The OSF set itself the task of promoting good research practice and to get a view on how widespread the issue of lack of reproducibility was it asked team across the world to replicate 100 high-profile psychology studies. Of these 100, only about half managed to reproduce the results of the original studies [8].

5 Move away from the low hanging fruit of short-term studies

In HRI, as in other scientific fields, we have a tendency to first go for the low hanging fruit. These are often studies that are short-term and that can be go from conception to completion with relatively minimal resources. Most studies use a single short exposure of the user to a robot. In some cases there is not even an actual interaction, instead participants are shown images or short videos of robots. The value of visual presentations of robots as a substitute for actual interaction is very much contested in the field. While many of these studies have exploratory value, it is unclear if these results will still hold if the interaction with robots moves from the short-term to the long-term.

A study requires at least a robot and a researcher, participants and time. Moving from short-term studies to long-term studies means that more of these resources will be needed. When running long-term studies additional goodwill will be needed from participants: it is relatively easy to convince people to come in for a quick 20 minute interaction study, but you will need to be very persuasive to get participants to return for 10 sessions or to have a robot in a real world environment for several weeks. Still, try we must, as the pursuit of HRI is to build the technical systems supporting HRI, and for the HRI to make it into the real world, where it will hopefully be used for longer than 20 minutes.

My team has experienced this first hand when we ran repeat interaction studies and had too many participants drop out before we reached the third interaction. The solution to getting everyone to return for repeat interactions turned out to be surprisingly simple: pay participants well, but only hand out the payment after completing

all sessions. We made the mistake of rewarding participants after each session by offering a hot drink, clearly not enough of an incentive to come back for more.

6 Conclusion

This chapters covered a selection of contemporary contentious issues in HRI and in experimental work at large. Many fields, including HRI, are realigning to a new reality. A reality where we desire more rigour from experimental science and more honesty in our reporting. While many solutions have been proposed to these problems, especially in experimental psychology, the consensus on what good practice is in HRI is still emerging. Our field would benefit from adopting agreed upon practices from psychology and medical sciences, including new practices such as preregistering studies, where the hypotheses and protocol are made public before the start of the actual study, taking away the opportunity to change the hypotheses to fit the data. HRI is fortunate to grow at a time where the global research culture is changing, moving towards a high-quality and transparent culture. HRI, as a field, can only benefit from these developments.

Acknowledgements The author would like to thank Paul Baxter, Bahar Irfan, James Kennedy, Fotios Papadopolous, Séverin Lemaignan, Emmanuel Senft for the discussion and insights used in this chapter.

References

1. Adair, J.G.: The hawthorne effect: a reconsideration of the methodological artifact. Journal of applied psychology **69**(2), 334 (1984)
2. Baxter, P., Kennedy, J., Senft, E., Lemaignan, S., Belpaeme, T.: From characterising three years of hri to methodology and reporting recommendations. In: The Eleventh ACM/IEEE International Conference on Human Robot Interaction, pp. 391–398. IEEE Press (2016)
3. Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., Tanaka, F.: Social robots for education: A review. Science Robotics **3**(21), eaat5954 (2018)
4. Berinsky, A.J., Huber, G.A., Lenz, G.S.: Evaluating online labor markets for experimental research: Amazon. com's mechanical turk. Political analysis **20**(3), 351–368 (2012)
5. Berkowitz, L., Donnerstein, E.: External validity is more than skin deep: Some answers to criticisms of laboratory experiments. American psychologist **37**(3), 245 (1982)
6. Boucher, J.D., Pattacini, U., Lelong, A., Bailly, G., Elisei, F., Fagel, S., Dominey, P.F., Ventre-Dominey, J.: I reach faster when i see you look: gaze effects in human–human and human–robot face-to-face cooperation. Frontiers in neurorobotics **6**, 3 (2012)
7. Buhrmester, M., Kwang, T., Gosling, S.D.: Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? Perspectives on psychological science **6**(1), 3–5 (2011)
8. Collaboration, O.S., et al.: Estimating the reproducibility of psychological science. Science **349**(6251), aac4716 (2015)
9. Cumming, G.: Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. Perspectives on Psychological Science **3**(4), 286–300 (2008)

10. Dawson, C.: Introduction to Research Methods 5th Edition: A Practical Guide for Anyone Undertaking a Research Project. Robinson (2019)
11. Esteban, P.G., Baxter, P., Belpaeme, T., Billing, E., Cai, H., Cao, H.L., Coeckelbergh, M., Costescu, C., David, D., De Beir, A., et al.: How to build a supervised autonomous system for robot-enhanced therapy for children with autism spectrum disorder. Paladyn, Journal of Behavioral Robotics **8**(1), 18–38 (2017)
12. Field, A., Hole, G.: How to design and report experiments. Sage (2002)
13. Irfan, B., Kennedy, J., Lemaignan, S., Papadopoulos, F., Senft, E., Belpaeme, T.: Social psychology and human-robot interaction: An uneasy marriage. In: Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, pp. 13–20. ACM (2018)
14. Kalegina, A., Schroeder, G., Allchin, A., Berlin, K., Cakmak, M.: Characterizing the design space of rendered robot faces. In: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, pp. 96–104. ACM (2018)
15. Kaptein, M., Robertson, J.: Rethinking statistical analysis methods for chi. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1105–1114. ACM (2012)
16. Landsberger, H.A.: Hawthorne revisited: Management and the worker, its critics, and developments in human relations in industry. (1958)
17. Leite, I., Martinho, C., Paiva, A.: Social robots for long-term interaction: a survey. International Journal of Social Robotics **5**(2), 291–308 (2013)
18. Li, J.: The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. International Journal of Human-Computer Studies **77**, 23–37 (2015)
19. Maxwell, S.E.: The persistence of underpowered studies in psychological research: causes, consequences, and remedies. Psychological methods **9**(2), 147 (2004)
20. Nomura, T., Kanda, T., Kidokoro, H., Suehiro, Y., Yamada, S.: Why do children abuse robots? Interaction Studies **17**(3), 347–369 (2017)
21. Pashler, H., Wagenmakers, E.J.: Editors introduction to the special section on replicability in psychological science: A crisis of confidence? Perspectives on Psychological Science **7**(6), 528–530 (2012)
22. Phillips, E., Zhao, X., Ullman, D., Malle, B.F.: 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, pp. 105–113. ACM (2018)
23. Ramachandran, A., Huang, C.M., Scassellati, B.: Give me a break!: Personalized timing strategies to promote learning in robot-child tutoring. In: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, pp. 146–155. ACM (2017)
24. Riek, L.D.: Wizard of oz studies in HRI: a systematic review and new reporting guidelines. Journal of Human-Robot Interaction **1**(1), 119–136 (2012)
25. Robins, B., Dautenhahn, K., Te Boekhorst, R., Billard, A.: Effects of repeated exposure to a humanoid robot on children with autism. In: Designing a more inclusive world, pp. 225–236. Springer (2004)
26. Vidgen, B., Yasseri, T.: P-values: misunderstood and misused. Frontiers in Physics **4**, 6 (2016)
27. Woods, S., Dautenhahn, K., Kaouri, C.: Is someone watching me?-consideration of social facilitation effects in human-robot interaction experiments. In: Computational Intelligence in Robotics and Automation, 2005. CIRA 2005. Proceedings. 2005 IEEE International Symposium on, pp. 53–60. IEEE (2005)