Effect confirmed, patient dead: A Commentary on Hoffman & Zhao's Primer for Conducting Experiments in HRI

KERSTIN FISCHER, University of Southern Denmark

This paper is a commentary on Hoffman & Zhao's 'A Primer for Conducting Experiments in Human-Robot Interaction.' I argue that a too narrow view of HRI methodology fails to address the dynamic systems properties of interaction. Furthermore, the focus on addressing the so-called 'replicability crisis' makes field studies next to impossible, inhibits interdisciplinarity and methodological pluralism, and draws our attention and resources away from the fact that contexts, people, cultures, expectations and interaction itself may influence how social signals are interpreted. Therefore, in spite of its great benefits, the 'Primer' may not be taken as an instruction on "how to carry out research in HRI" in general.

CCS Concepts: • **Human-centered computing** → Human computer interaction (HCI); *HCI theory, concepts and models*; *Empirical studies in HCI*

KEYWORDS

Methodology, replicability crisis, power analysis, interaction, interdisciplinarity

ACM Reference format:

Kerstin Fischer. 2020. Effect confirmed, patient dead: A commentary on Hoffman & Zhao's primer for conducting experiments in HRI. *ACM Transactions in Human-Robot Interaction*. X, X, Article X (December 2020), 4 pages. https://doi.org/10.1145/1234

In times of 'alternative facts,' when people feel that their opinions are not heard 'just' because they don't correspond to scientific findings, scientific rigor and methodological discussion are absolutely crucial, and the 'Primer' makes a laudable attempt at establishing a common ground among HRI researchers on what such a methodological rigor may consist in. At the same time, each methodology is based on certain premises about what constitutes the object of study and what knowledge is to be gained (Kuhn 2012). We thus need to be aware of what we buy into when promoting controlled methods like those described in the 'Primer,' and what we are missing out on.

What I would like to argue for in this commentary is that the methodology proposed in the 'Primer' restricts the object of study too much to account for the object of study of HRI, namely human-robot *interaction*. While being very useful to guide *some* HRI research, the methods suggested in the paper should not be expected to inform all or even most of the work in the field.

First, given the formulated expectations regarding statistical power and effect size (like 64 participants per condition in the example study), field studies are almost ruled out; many work places, for instance, don't even have this many employees, school classes are typically restricted to 20-30 kids, elderly care facilities have units with limited numbers of residents, and studies in the wild (say, in a shopping mall) will be practically impossible because real life comes with so much variation that identifying effects of interventions and of mediating and moderating factors would require the recording and analysis of hundreds of interactions.

In controlled lab studies, recruiting something like 64 participants per condition is more realistic, but introduces a selection bias since only people interested in novel technologies sign up for such experiments. The probability that participants are university students and staff, probably even with a technical background (cf. also Henrich et al. 2010), is quite high, and even if they stem from various backgrounds, they know that they are in a safe experimental setting where they will encounter a robot, and they have nothing better to do than to interact with that robot. This is very different from real-life scenarios in which people are not mentally prepared to encounter robots and have their own schedules, agendas and attitudes that may influence the interactions. Thus, even though Hoffman & Zhao, the authors of the 'Primer,' discuss field studies as one option, the methodological requirements proposed can only be achieved in controlled lab studies, which restricts the range of phenomena encountered and hence the object of study.

Second, the requirements proposed bind large amounts of resources - in terms of availability of a robust robot, student assistants, lab space etc. over the course of long periods of time in order to carry out experiments with that many participants per condition. I wonder whether the "replicability crisis" is really our biggest concern in the HRI community so that we should put all our efforts into addressing problems of statistical power. For instance, given what we know about the effects of participants' backgrounds (e.g. Strait et al. 2015), contexts (e.g. Butler et al. 2019), environments (e.g. Mutlu & Forlizzi 2008) and framing (e.g. Paepcke & Takayama 2010), many controlled HRI studies are unlikely to replicate with other populations, other collaborative tasks or other framings anyway - we just don't understand human-robot *interactions* well enough yet to make that kinds of predictions. So given the problems we face in HRI, I would not prioritize replicability (cf. also Gilbert et al. 2016). Furthermore, from the top of my head, I cannot think of any HRI study with that many participants per condition, but of lots of exciting work that furthered our understanding about how people interact with robots.

Third, the hypothesis-driven procedure proposed in the primer makes it very hard to study complex, dynamic interactional phenomena. Interactions are dynamic, mutually adaptive systems with emergent properties. For such systems, often no clear hypotheses can be formulated in a meaningful way. Hoffman & Zhao consider such experimental work 'exploratory' and thus not the real thing (to be addressed in pilot studies). In contrast, I fail to see the difference between reporting that a hypothesis was not confirmed and saying out front that one can't know what will happen - from the perspective of the knowledge gained, the latter seems even preferable; so I wonder whether we can afford to dismiss a broad range of potentially fruitful studies based on a standard adopted in psychology. Note that other disciplines take a hypothesis-driven approach to research to be misguided (e.g. ethnomethodological conversation analysis, cf. Sacks et al. 1974).

Another problem is that the kinds of research questions addressed in the hypothesis-driven research advocated in the 'Primer' focuses very much on the effects of single social cues or single social behaviors; however, while also in human interactions we can assign some values to human signaling systems, such as language, meanings are eventually interactionally achieved. For instance, we can say describe the meaning of the verb "to sit down" and of the imperative form in the utterance "sit down!" to some extent (for example, as defaults or prototypes), but whether the utterance is a command, suggestion or polite invitation has to be inferred from the context and ratified by the interactants; for example, by replying "thank you", the interaction partner displays a certain interpretation of the utterance that can then be rejected or accepted by the original speaker. Also human-robot interactions are joint actions (Clark 1996; Fischer 2016); however, the emergent aspects of interactions and people's joint sense-making processes over the course of an interaction are not addressed in the approach proposed.

The perspective taken in the 'Primer' is similar to the search for a protein that inhibits a virus' ability to connect to a human cell and thus to multiply. This is without doubt important research and should definitely be carried out. However, in order to develop a cure against the virus, researchers also have to ensure that the protein gets to the right place at the right time, under all circumstances, for all people alike, and that the

substance developed only kills the virus and not its host. Human bodies are extremely complex, dynamic systems, and thus identifying substances in petri dishes is a different challenge from understanding how the different subsystems of the body interact to defend themselves against a virus and keep the whole system running at the same time. Similarly, identifying with high certainty that under controlled conditions a social signal produced by a robot can have a particular effect is important - but it is only part of what makes up an interaction between a human and a robot, because in real life, signals are never isolated; instead, they occur in certain contexts (e.g. Andrist et al. 2014; Jung & Hinds 2018), in connection with other signals and robot behaviors (e.g. Jensen et al. 2017), with different people who may have very different expectations than those people who volunteer to be participants in our lab studies (e.g. Chang & Sabanovic 2015), and they may have different effects after repeated exposure (e.g. André 2018; Paetzel & Castellano 2019).

The 'Primer' is a great paper and will certainly be very helpful to generations of students who we are training in experimental techniques. However, it is also crucial that we as the HRI community do not take it as a measure for empirical work in HRI in general and understand the limitations of the approach advocated; that is, the 'Primer' may not be taken as an instruction on "how to carry out research in HRI". HRI is an interdisciplinary community out of necessity, but interdisciplinarity is hard because the kinds of questions addressed, methods used and results achieved may be radically different between different disciplines (again, see Kuhn 2012). My fear is that the 'Primer' will make interdisciplinarity in HRI even harder, because it states clearly evaluation criteria for *some* empirical studies - and it will be easy to forget that these criteria can and should not be used to measure *all* HRI work.

Eventually, I don't think we have a choice but to broaden the range of methodologies accepted within the HRI community (cf. also Jung & Hinds 2018): If we want to bring robots out of the labs and into people's real lives, we have to understand not only what effects certain robot behaviors may have *in principle*, i.e. in controlled scenarios in the lab, but also what effects they may have in different contexts, in relation to other behaviors, in interactions with different people and over time. Only if we can account for these effects as well will HRI survive as a discipline.

REFERENCES

André, E. Socially-Sensitive Technologies for HRI: Challenges and Perspectives. In *Proc. FAIM/ISCA Workshop on Artificial Intelligence for Multimodal Human Robot Interaction* (pp. 1-3).

Andrist, S., Tan, X. Z., Gleicher, M., & Mutlu, B. (2014). Conversational gaze aversion for humanlike robots. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 25-32). IEEE.

Butler, R., Pruitt, Z., & Wiese, E. (2019). The Effect of Social Context on the Mind Perception of Robots. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 63, No. 1, pp. 230-234). Sage CA: Los Angeles, CA: SAGE Publications.

Chang, W. L., & Sabanovic, S. (2015). Interaction expands function: Social shaping of the therapeutic robot PARO in a nursing home. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 343-350). IEEE.

Clark, H. H. (1996). Using Language. Cambridge: Cambridge University Press.

Fischer, K. (2016). *Designing Speech for a Recipient: The roles of partner modeling, alignment and feedback in so-called'simplified registers'* (Vol. 270). John Benjamins Publishing Company.

Gilbert, D., King, G., Pettigrew, S. & Wilson, T. (2016). Comment on ‘Estimating the reproducibility of psychological science’. Science, 351, 1037a-1037b.

Henrich, J., Heine, S., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), 61-83. doi:10.1017/S0140525X0999152X

Jensen, L. C., Fischer, K., Suvei, S. D., & Bodenhagen, L. (2017). Timing of multimodal robot behaviors during human-robot collaboration. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 1061-1066). IEEE.

Jung, M., & Hinds, P. (2018). Robots in the wild: A time for more robust theories of human-robot interaction. ACM *Transactions on Human-Robot Interaction* 7, 1, Article 2.

Kuhn, T. S. (2012). *The structure of scientific revolutions*. University of Chicago Press.

Mutlu, B. and J. Forlizzi. (2008): Robots in organizations: The role of workflow, social, and environmental factors in human-robot interaction. Proceedings of HRI '08.

Paepcke, S., & Takayama, L. (2010, March). Judging a bot by its cover: an experiment on expectation setting for personal robots. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 45-52). IEEE.

Paetzel, M., & Castellano, G. (2019). Let Me Get To Know You Better: Can Interactions Help to Overcome Uncanny Feelings?. In *Proceedings of the 7th International Conference on Human-Agent Interaction* (pp. 59-67).

Sacks, H., E. A. Schegloff, and G. Jefferson (1974). A simplest systematics for the organization of turn-taking for conversation. Language 50 (4), 696–735.

Strait, M., Briggs, P., & Scheutz, M. (2015). Gender, more so than age, modulates positive perceptions of language-based human-robot interactions. In *4th international symposium on new frontiers in human robot interaction* (pp. 21-22).