Cimiano P., Pietsch C., & Wiljes C. (Eds.) (2021). Studies in Analytical Reproducibility: the Conquaire Project. Bielefeld. https://doi.org/10.4119/unibi/2942780

# 10 | Reproducibility in Human-Robot Interaction Research: A Case Study

Florian Lier[1], Sebastian Meyer zu Borgsen[1], Sven Wachsmuth[1], Jasmin Bernotat[2], Friederike Eyssel[2], Robert Goldstone[3], Selma Šabanović[3]

1 – Faculty of Technology & Cognitive Interaction Technology Excellence Center (CITEC), Bielefeld University
2 – Faculty of Psychology and Sports Science, Department of Psychology, Bielefeld University
3 – Indiana University Bloomington

## Abstract

Studies in human-robot interaction (HRI) typically involve computational artifacts, i.e. the robotic system, as the subject of investigation. Thus, the reproducibility of any result in HRI studies directly relates to the reproducibility of this computational artifact in the first place. This has certain consequences for appropriate workflows that will be discussed in this chapter. We argue for a higher awareness, improved standards, and further automation of tool chains used to conduct robotic experiments. We identify this as a research topic in its own right, especially in cases where robotic systems are used in interdisciplinary research. This inherently includes that technically complex robotic experiments should also be reproducible by scientists with a *non-technical* background. We analyze and discuss a dedicated study by the CITEC Central Lab Facilities and an international team demonstrating that it is possible to replicate a relatively complex HRI experiment in two different laboratories across the globe by a research assistant with no experience in robotics at all.

## Keywords

Human-Robot Interaction, Reproducibility, Robotic experiments

## 10.1 Introduction

The Central Lab Facilities (CLF) group of the Excellence Cluster Cognitive Interaction Technology (CITEC) at Bielefeld University aims to develop and improve technology, workflows, and tool chains for building as well as experimenting with interactive intelligent systems [1, 2, 3, 4, 5]. An important application and research field is human-robot interaction (HRI), which requires sophisticated robotic research platforms that include many software and hardware challenges besides the core areas of perception, behavior generation, and interaction design. Thus, research in HRI is a highly interdisciplinary endeavor. It aims to model the physical as well as mental dynamics between a human and a robot in a communicative or cooperative situation. It builds upon concepts and ideas from the area of human-human interaction in order to make the human-robot interface as smooth and intuitive as possible. Dealing with physically embodied agents, this includes many engineering issues towards flexible and save movements, many issues from machine perception, e.g. recognizing the interaction partner, many issues from artificial intelligence towards an interpretable and goal-oriented behavior of the robot, as well as many issues explored by the social sciences (psychology, linguistics, cognitive science, etc.) in order to understand associations, attributions, and expectations that humans have when interacting with a robot. Last but not least, any experiment with an autonomous robot includes many system engineering challenges including significant complexity issues on the software side which are frequently underestimated. Although there has been considerable progress in robot technology including available robotic standard platforms (e.g. iCub, Softbank's Nao and Pepper, Toyota's HSR), software frameworks [6, 7, 8, 9], and benchmarking activities [10, 11, 12, 13], the theoretical and practical foundations for experimental replicability of experiments in robotics is still in its infancy [14]. In this regard, Bonsignorio et al., e.g., states that *'even determining the information required to enable replication of results has been subject of extensive discussion'* [14].

In this chapter, we argue for a higher awareness, improved standards, and further automation of tool chains used to conduct robotic experiments. We identify this as a research topic in its own right, especially in cases where robotic systems are used in interdisciplinary research. This inherently includes the fact that technically complex robotic experiments should also be reproducible by scientists with a *non-technical* background. While this goes beyond the goals of Conquaire to reproduce the analytical part of an experiment only, in human-robotic interaction studies the replication of the technical settings is essential to understand the experimental results. The other Conquaire studies mostly deal with computational workflows and tools that are applied to datasets *after* these have been recorded in an experiment. Because most studies in, e.g., the natural sciences deal with 'natural' phenomena – i.e. they are not produced by an artificial artifact – the dataset can be interpreted with regard to this

phenomenon at any place in the world. This is not the case for experiments including robots. The dataset can only be interpreted with regard to the specific artifacts used in the experiment. As a consequence, the reproducibility of an experiment and the validity of the data must include the possibility to reproduce also the robotic system and its behavior in the study.

In the following, we report our experiences and lessons learned in analysing a replication study conducted by the Central Lab Facilities involving a human-robot interaction (HRI) experiment in Bielefeld and at a partner site of the DFG Excellence Cluster CITEC within the DAAD Thematic Network Interactive Intelligent Systems. The study investigates an extended version of Stenzel et al.'s *'Joint Simon effects for non-human co-actors'* [15], in two labs in different institutions and continents. In psychology, the Joint Simon effect is used to investigate to what extent people mentally represent their own and other agent's actions in a joint task. This leads to delayed decision effects when a human is prompted with stimuli that are spatially incompatible with the roles in a team. The effect disappears when people think that they interact with a non-biological, technical artifact. Thus, it is an open question to which degree humanoid robots are perceived as social agents or team mates and if this can be shown using the Joint Simon effect (see Sec. 10.2.1 for more details).

To this end, the CLF researchers applied a novel software tool chain and methodology that implements state-of-the-art techniques with the objective of facilitating reproducibility in robotics research. The experiment was designed in cooperation between Bielefeld University and Indiana University Bloomington by a team of interdisciplinary scientists originating from psychology & brain sciences, informatics and robotics. The team initially conducted the study in Bielefeld before a replication attempt in Indiana was conducted. In this context, they specifically chose the following constraints in order to impose the same restrictions and obstacles encountered in 'regular' replication attempts:

1. The experiment must be replicated by a staff member who is *not* part of the research project.

2. The only starting point for replication is an online manual explaining our approach and the literature references therein.

3. Assistance from Bielefeld is only provided in otherwise irresolvable situations.

A replication of this experiment at different sites is an interesting case study from two different points of view. On the one hand, it is interesting to investigate whether there are cultural factors that affect the results. On the other hand, the setup includes a behavioral study with a robotic platform (the NAO robot), which is programmed to physically press a button where timing matters. Thus, from the perspective of reproducibility and the lessons learned from the Conquaire project, there are the following research questions: **(H1)** Is the tool

chain and methodology been suitable to represent all aspects required for successful replication? **(H2)** What can we learn about reproducibility in general with respect to unexpected technical obstacles or situations one did not anticipate? **(H3)** Can the second study cross-validate the results obtained in the original Bielefeld study?

## 10.2 Experimental Settings and Methods

The following part of this contribution will cover the replication approach and the lessons learned. Important parts of the study and tool chain have been published previously [16, 17]. A final evaluation of the second study is still ongoing work. First, we will shortly introduce the theoretical background of the experiment. Then, we present the procedure and methods, and finally discuss our findings.

### 10.2.1 The JSE Experiment

The study was designed out to reproduce a variant [18] of a well-documented psychological effect, the Joint Simon Effect (JSE) [15]. The JSE describes a difference in reaction time depending on identity (*compatibility*) or disparity (*incompatibility*) of a stimulus' and the co-actors' spatial position in relation to the participant during a shared go/no-go task. The team aimed at reproducing this effect with a robot as co-actor as described in [18] and adopted the stimuli and procedure attributes. The original experiment was extended with a *robot position* condition to additionally test the influence of the robot's spatial relation to the human subject. While more detailed information about the JSE experiment can be found in the paper by Dolk et al. [18], we will briefly describe the experiment setup variant used in the particular study described in this chapter. Due to its wide distribution and availability, the team used the humanoid robot NAO as the participant's co-actor (Figure 10.1). The robot kneels next to the test subject on a table or chair. The barycenter of the robot is approximately at elbow height of a sitting subject.

The participant and the robot each have their own keyboard of identical type. The keyboards are directly adjacent and on the same level. During the experiment, stimuli, e.g. a square and a diamond, are displayed on a screen at randomized positions and in randomized order. Based on the initial assignment, either the robot or the human have to press the space-bar key as soon as the assigned stimulus appears. The corresponding reaction times (RT) of the human co-actor are measured.

In Bielefeld, the team tested 47 subjects from the nearby campus ($M$ age = 24.61 years, $SD$ age = 4.01 years). Each run consisted of 512 trials with short breaks per 128 trials and took approximately 30 minutes. The findings were similar to those found by Stenzel et al., the experiment showed a significant

main effect of compatibility when analyzing the response times (RT), $F(1,48) = 11.639$, $p < 0.001$, partial $\eta^2 = .43$, indicating shorter RTs in compatible (423 ms) compared to incompatible trials (434 ms), which confirms the presence of an overall JSE. The team did not find a significant interaction between *compatibility* and *robot position*.

The data of the experiment were logged within the software tool jsPsych [19] that controlled the prompting, triggered the execution of robot movements, and recorded the reactions of the human participants (execution protocol of the experiment including timing events for prompting and robot, spatial configuration of prompts, etc.). The data is stored as comma-separated-value files which are preprocessed with documented shell commands and python scripts. Data analysis was conducted with SPSS[1] or R[2] tools.

### 10.2.2 Replication in Indiana

In order to reproduce the experiment in Indiana, under consideration of the demands and requirements in the current literature and the issues presented in section 10.1, there are two core issues to be solved:

1. A systemic solution for deployment, configuration, and integration of all necessary software artifacts.

2. A structured methodological 'how-to' for setup and execution considering user groups and tools from other disciplines, here, psychology.

This should not come as overhead for the replication of an experiment. It is essential that the replication tool chain is already in place and used when the first experiment is developed and conducted. Thus, the replicability of an experiment including software-intensive systems as core components has to be planned already when setting up the original experiment.

#### The replication tool chain

In order to address the above issues, the research team developed a software tool chain that has been explicitly designed to foster reproducibility of software intensive experiments in robotics — the *Cognitive Interaction Toolkit* (CITK) [3]. More detailed technical information is provided by Lier et al. [1, 2]. The requirement to support disciplinary tools to design and run experiments will be additionally covered by jsPsych [19].

At its core, the CITK provides a template-based "artifact-description" repository in order to pool and aggregate all required artifacts of a robotics experiment (cf. 10.2). There are basically two types of descriptions. The first is called

[1] `https://www.ibm.com/de-de/products/spss-statistics`
[2] `https://www.r-project.org/`
[3] `https://toolkit.cit-ec.uni-bielefeld.de`

recipe: it defines required system artifacts, e.g, software components, downloadable data sets, or system configuration files. Templates for new types of artifacts can be added on-the-fly by developers. With regard to pure software aspects, the existing set of templates contains macros for the most common build tools like autotools, maven, CMake, and ROS/catkin[4], enabling native builds of various kinds of software. These macros also help to remove redundancy and keep the recipes clean and well-structured. The second type is called distribution. A distribution is a composition of a number of arbitrary recipes and hence determines an entire system. Distributions, as well as recipes, mandatorily reference *versions*, e.g., tags, branches, or commit hashes of an artifact, such that a distribution reflects a *fixed* description of a system. Recipe and distribution files are publicly available in our Git-repository[5]. Another core-component is a prepackaged, i.e, download and run it, no configuration required, CI server. It is utilized to compile, deploy, and run entire software systems defined in distribution files. The server provides a web front-end that can be accessed via a browser for ease of use. In order to deploy and run a system, the CITK implements a generator-based approach. A so-called build-job-configurator tool automatically creates all required build-jobs (for every recipe in a distribution) on the server. A user merely selects the desired distribution file. Moreover, it is also possible to connect a physical robot to the machine that runs the CI server in order to control/actuate it. Lastly, our approach also provides a framework to automatically bring up (statefull execution), stop, and introspect a robotics software system. Executing a system merely requires to select and activate a designated build-job in the web front-end. Data that is acquired/logged during each system run is also stored on the server and accessible via web browser. By utilizing this part of our structured CITK approach, the team could ensure technical reproducibility of all required artifacts and also repeatable experiment execution regarding the software side of an experiment. An exemplary CITK tool chain demonstration video can be watched here: `https://vimeo.com/205541757` With respect to experiment design and orchestration, the study additionally made use of a framework called jsPsych. jsPsych is a JavaScript library for creating behavioral experiments in a web browser. It provides a description of the experiment structure in the form of a time line. It handles which trial to run next and storing the obtained data. jsPsych uses plugins to define what to execute at each point on the time line. The functionality of jsPsych was extended in order to i) trigger an experiment run on the CI server and ii) execute experiment-specific behaviors of the NAO/Pepper robots, e.g, based on the current state of the time line in jsPsych. Detailed information about jsPsych can be found in [19].

[4] `http://wiki.ros.org/catkin/conceptual_overview`
[5] `https://opensource.cit-ec.de/projects/citk`

**The replication experiment**

Due to the fact that the entire software system was already modeled using the CITK for the Bielefeld study [6], no additional work, besides the translation from German to English, e.g, in the jsPsych time-line slides was required. Hence, the software part including robot movement control interfaces, calibration procedures, and jsPsych experiment orchestration was already at hand. Since there was no prior knowledge about the (scientific) background of the staff member who would eventually replicate the experiment in Indiana, the team implemented a generic GUI-based application for all crucial technical steps with respect to the robot *hardware*, e.g, the calibration procedures. Finally, a detailed instruction was compiled on a public GitHub page (final version [7]). This online manual included the following steps:

1. Introduction
2. Hardware Requirements and Prerequisites
3. Software Requirements and Prerequisites
4. Physical Experiment Setup
5. Subjects
6. Executing the Experiment
7. Results
8. Literature

In summary, the manual included the following content:

- a brief introduction to the research topic and study goals, plus references to related literature,
- a specification of the required hardware, e.g, a NAO robot acquired within 2-3 years,
- a PC or laptop including CPU and RAM specifications including the size, resolution and refresh rate of the utilized screens,
- a specification of the operating system requirements, i.e., Ubuntu Xenial (16.04, 64 bit),
- an explanation of how to setup the physical experiment, such as height and position of the robot, position of the keyboards, monitors, etc., and

[6] `http://www.webcitation.org/6xlwomECk`
[7] `https://Git.io/vAxml`

- a brief explanation of the network setup.

Moreover, the document included detailed instructions about the installation and usage of the CITK in order to deploy the software system, calibrate the robot, and run the experiment. The instructions also provided information about the subjects, the welcome and actual experiment procedure. Lastly, it was explained how to obtain and inspect the gathered data. So far, the documentation included detailed information with respect to technical (soft- and hardware), as well as methodological/procedural aspects, to reproduce the study as it was conducted in Bielefeld. The team also established a communication channel via instant messaging using Slack. The channel was intended to provide 'emergency support', but only in case of an otherwise irresolvable situation. Hence, the chat history could also be exploited for post-experiment analysis, if required.

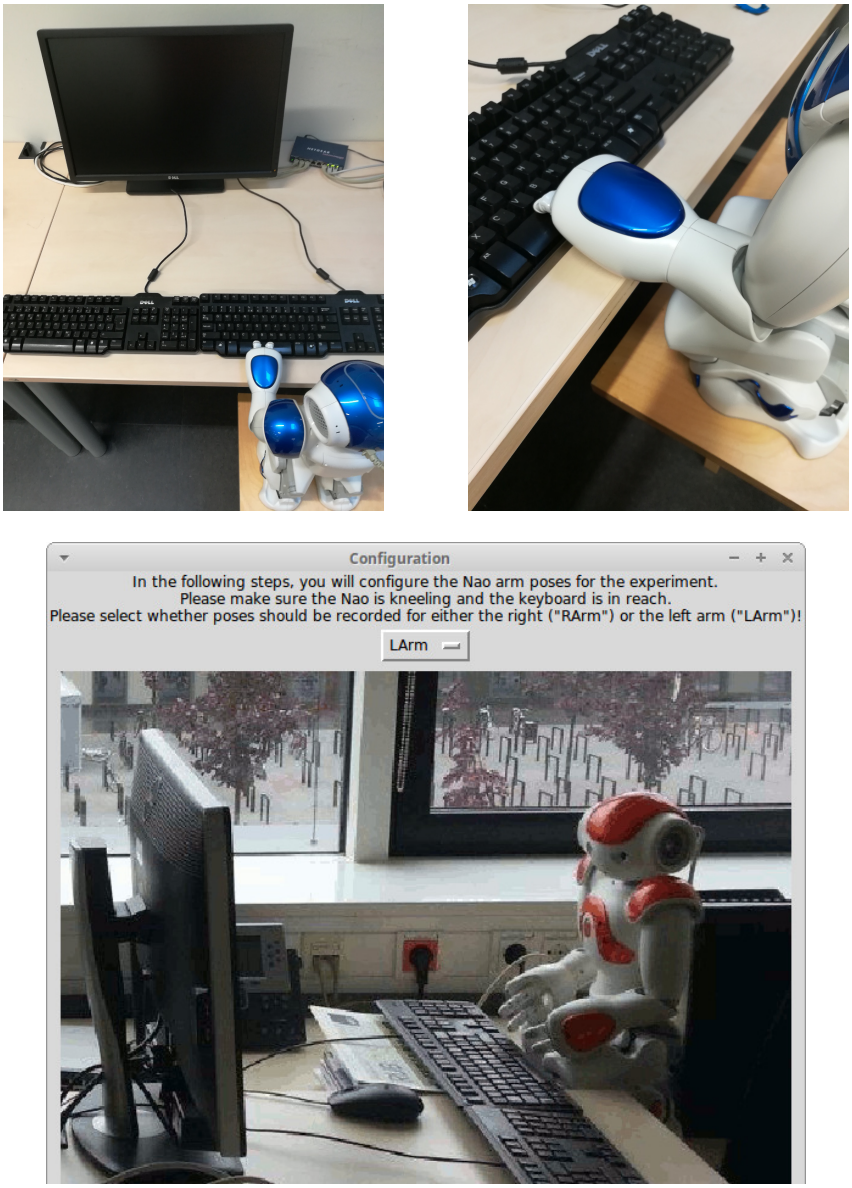


Figure 10.1: Top left: The NAO JSE setup used in one of our Bielefeld setups; Top right: NAO keypress pose; Bottom: Screenshot of the robot calibration GUI


Figure 10.2: Toolchain for the replication of robotics experiments: An experiment for which the associated computational artifacts are documented in a browsable online catalog which is linked to the corresponding repositories. From here a researcher is instructed to automatically roll out the system distribution using a continuous integration server which is linked to the robot platform as well as the computer controlling the behavioral experiment using JsPsych. Left: Overview of the toolchain, Right: Screenshot of the CI server web front-end. Each row corresponds to a recipe that has been translated into a build job. Build jobs can be activated by selecting the most right icon (stopwatch). Execution of an experiment can also be triggered using this front-end.

## 10.3 Analytical Reproducibility: Results & Lessons Learned

We report on the lessons learned in a time line based manner. Depending on the reader's background, either in computer science or the humanities for instance, some of the reported obstacles may appear 'trivial'. However, we claim that it is crucial to raise awareness for false assumptions made by domain experts, e.g., with regard to common knowledge about specific technological or methodological aspects of an experiment, which are by far not so obvious/common for others outside their domain. Furthermore, we would like to point out that the reported observations are based on a *practical interdisciplinary replication attempt*, which is especially valuable in order to learn about all *the different characteristics* and challenges concerning replicability of robotics systems.

The replication attempt of this JSE experiment was conducted by a research assistant (RA) with a background in psychology. With respect to interdisciplinary research this was, on the one hand, an almost ideal scenario, on the other hand however, a technically-challenging one as well.

### 10.3.1 Technical Obstacles & Procedural Issues

The following issues were reported during the replication study in Indiana. The research assistant started with a plain laptop. Thus, the first issue was reported shortly after the study officially started. Even though the deployment of the required software components (using the CITK) was successful on first attempt, the RA faced a couple of issues with the installation routine of Ubuntu. The team in Bielefeld could resolve these issues by pointing the RA to the correct Ubuntu documentation. The second technical issue was reported a few days later. The operating system as well as the robot software environment were already installed successfully. Nonetheless, during the required robot calibration procedure, a connection to the robot could not be established via local network. The team in Bielefeld resolved this issue by instantly updating the online manual for the network setup which is also hosted in the linked repositories. The third issue was reported after a first test run of the experiment. So far, the entire software system was deployed, the robot calibrated, and also the physical experiment setup was in place. However, during the run, the RA noticed that the translation of two single lines of text on a slide in the jsPsych time line was missing. The team in Bielefeld could resolve this issue by correcting the error in the code base and subsequently updating the Git repository. In Indiana, the RA just had to re-trigger the corresponding build job, thus automatically installing the updated version of the experiment. The fourth and last reported obstacle occurred in an early stage of the actual experiment. Since the tool chain allows to download and inspect already gathered data via web browser, the colleagues in Indiana soon took a first look at intermediate results. They noticed that the

distribution of the participants' position with respect to the robot indicated a strong preference towards only *one* side. The team in Bielefeld discovered that the instructions provided for the experimenter in jsPsych, addressing the procedure of subject positioning, were not as precise as they should have been. This could be corrected by updating the description in the repository.

### 10.3.2 Results of the Pilot Study on Reproducibility in HRI

At time of writing, the JSE experiment in Indiana has been finished; first results show a weaker but observable Joint Simon effect. However, besides the obstacles already discussed, there were several positive lessons learned. It is very difficult, if not impossible, to foresee all pitfalls faced by the researcher replicating the experiment. As a local solution or patch does not solve the issue in a consistent manner, a flexible tool chain is required that allows for almost instant patching and deployment of experiment artifacts. In this regard, the technical complexity of the deployed robotic system (hard- and software) was completely hidden to the research assistant (RA) in Indiana. The time required to setup the entire software system was limited to a few hours, including the installation of an operating system. Moreover, the acceptance threshold and usability of the CITK tool chain appeared to be positive, given the fact that it was easily usable by the RA. Also, the transition from design, implementation and execution of the experiment in Bielefeld to the deployment in Indiana merely required sending a link to an online manual. In a short post-experiment interview we asked the RA for a self-assessment regarding the experience with Linux-based systems, robotic hardware, robotic software, the Linux network stack, conducting HRI experiments, and conducting psychology experiments. In summary, the RA was reasonably experienced in conducting experiments in psychology and, having used Linux before, knew a few basic Linux commands. Regarding the remaining topics, the RA was an inexperienced user, i.e, had never operated a single robot before.

## 10.4 Analysis of reproducibility experiment

In this section, we discuss lessons learned from our cross-site replication study of a robotic study.

**Reproducibility is decided at development time:** We would like to stress that without having the tool chain in place at the development and preparation time of the study in Bielefeld, a replication study at Indiana would have been extremely time consuming if not impossible. Thus, any tool chain used for the replication of results should be established in the daily workflow of the researchers understanding it as a *development tool* instead of a replication tool.

**Experimental protocols:** Besides the technical requirements and issues involved in replicating studies and their scientific results, it is also important to neatly document the experimental protocol. Typically, this is solved by workflows, policies, and tools within the specific discipline without being integrated in the technical framework of a robotic experiment. In the study reported, a tool from psychology was integrated for experimental control. This is also a prerequisite for a systematic logging of all experimental data. However, we can observe in the study that the non-technical aspects of the experimental protocol were not sufficiently described, which raised several questions when intermediate results were analysed and discussed. Thus, there is still an open issue to more formally describe the experimental protocols.

**Scientists with a non-technical background:** An interdisciplinary field like HRI involves experts from different backgrounds. Reproducibility should not depend on having a robotic expert on-site. Even though the current approach demonstrated that *it is possible*, even by an inexperienced user, these first obstacles were not even close to what a robotics engineer would consider 'an obstacle'. On this account, we deem this lesson learned even more valuable. Furthermore, these kind of 'low-level' obstacles can be easily mitigated by providing detailed *beginner-level instructions*.

**Automated documentation roll-out:** It appears extremely useful to be able to quickly and dynamically alter instructions provided for replication attempts if errors are reported/discovered. SCM-based repositories, not only for source code, but also for this kind of manuals seem to be a well-applicable solution. Further, adding replication instructions to the corresponding source code of a publication is not labor intensive at all.

**Report intermediate results:** The issues and obstacles discussed before imply that it is important to automate the collection and evaluation of (also) intermediate results to prevent subsequent failures, especially if data acquisition is time-consuming. Thus, the requirement for an analytic reproducibility also applies to intermediate results. In the case of the experiment considered here, all preprocessing steps and scripts were precisely documented. Further, the 'R'-toolbox can be used as an open source alternative to SPSS for the statistical analysis of the data.

**Open issues:** Regarding the limitations of the approach presented, the toolbox currently does not incorporate any standardized benchmark procedures for more general HRI experiments. It does not provide any tool or template support for metrics (if existent/agreed-upon) with respect to comparability of HRI systems. We are open for discussion and welcome contributions concerning this topic.

**Final remarks:** In this contribution we discussed and analyzed a dedicated study on the replication of a reasonably complex HRI experiment in two different laboratories across the globe without a) flying experts in and b) making a single video/phone call — by a research assistant with *a non-technical background* and no experience in robotics at all. We reported on the lessons learned during this practical replication process.

## 10.5 Conclusion

This chapter has shown that it is possible to reproduce a robotic experiment at different sites, reproducing the same effect. The chapter has presented a workflow that provides end-to-end support for researchers wanting to reproduce a certain experiment. In this particular case, the workflow was based on the CITK toolkit developed at CITEC. In the particular case, the experiment involved a reproduction of the well-known Joint Simon effect known from psychology research. Using the end-to-end experimental workflow described in this chapter it was possible for a psychologist from Indiana University not expert in robotics to reproduce an experiment originally carried out at Bielefeld University. We regard this as a clear success story of experimental reproducibility and see this as a best practice of reproducibility.

## References

[1] Florian Lier, Marc Hanheide, Lorenzo Natale, Simon Schulz, Jonathan Weisz, Sven Wachsmuth, and Sebastian Wrede. Towards Automated System and Experiment Reproduction in Robotics. In Wolfram Burgard, editor, *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016.

[2] Florian Lier, Johannes Wienke, Arne Nordmann, Sven Wachsmuth, and Sebastian Wrede. The cognitive interaction toolkit–improving reproducibility of robotic systems experiments. In *International Conference on Simulation, Modeling, and Programming for Autonomous Robots*, pages 400–411. Springer, 2014.

[3] Florian Lier, Ingo Lütkebohle, and Sven Wachsmuth. Towards automated execution and evaluation of simulated prototype HRI experiments. Proc. 2014 ACM/IEEE Int. Conf. on Human-robot interaction, pages 230–231. ACM, 2014.

[4] Severin Lemaignan, Marc Hanheide, Michael Karg, Harmish Khambhaita, Lars Kunze, Florian Lier, Ingo Luetkebohle, and Gregoire Milliez. Simulation and hri recent perspectives with the morse simulator. LNAI Lecture Notes in Artificial Intelligence. Springer, 2014.

[5] S. Meyer zu Borgsen, P. Renner, F. Lier, T. Pfeiffer, and S. Wachsmuth. Improving human-robot handover research by mixed reality techniques. In *Proceedings of VAM-HRI 2018. The Inaugural International Workshop on Virtual, Augmented and Mixed Reality for Human-Robot Interaction*, 2018.

[6] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5, 2009.

[7] Herman Bruyninckx. Open robot control software: the orocos project. In *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, volume 3, pages 2523–2528. IEEE, 2001.

[8] Giorgio Metta, Paul Fitzpatrick, and Lorenzo Natale. Yarp: yet another robot platform. *International Journal on Advanced Robotics Systems*, 3(1):43–48, 2006.

[9] Emmanuel Pot, Jérôme Monceaux, Rodolphe Gelin, and Bruno Maisonnier. Choregraphe: a graphical tool for humanoid robot programming. In *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*, pages 46–51. IEEE, 2009.

[10] Sven Wachsmuth, Dirk Holz, Maja Rudinac, and Javier Ruiz del Solar. Robocup@home - benchmarking domestic service robots. In *AAAI*, 2015.

[11] Pedro U. Lima, Daniele Nardi, Gerhard K. Kraetzschmar, Rainer Bischoff, and Matteo Matteucci. Rockin and the european robotics league: Building on robocup best practices to promote robot competitions in europe. In Sven Behnke, Raymond Sheh, Sanem Sarıel, and Daniel D. Lee, editors, *RoboCup 2016: Robot World Cup XX*, pages 181–192, Cham, 2017. Springer International Publishing.

[12] F Amigoni, A Bonarini, G Fontana, M Matteucci, and V Schiaffonati. Benchmarking through competitions. In *European Robotics Forum–Workshop on Robot Competitions: Benchmarking, Technology Transfer, and Education*, 2013.

[13] Ana Huaman Quispe, Heni Ben Amor, and Henrik Christensen. A taxonomy of benchmark tasks for robot manipulation. In *Robotics Research*, pages 405–421. 01 2018.

[14] Fabio Bonsignorio. Reproducible research in robotics: Current status and road ahead. `http://www.heronrobots.com/EuronGEMSig/gem-sig-events/icra2017workshoprrr`, May 2017. (Accessed on 06/03/2018).

[15] Anna Stenzel, Eris Chinellato, Maria A Tirado Bou, Ángel P del Pobil, Markus Lappe, and Roman Liepelt. When humanoid robots become human-like interaction partners: corepresentation of robotic actions. *Journal of Experimental Psychology: Human Perception and Performance*, 38(5):1073, 2012.

[16] Florian Lier, Phillip Lücking, Joshua R. de Leeuw, Sven Wachsmuth, Selma Sabanovic, and Robert Goldstone. Can we reproduce it ? toward the implementation of good experimental methodology in interdisciplinary robotics research. In *ICRA 2017 Workshop on Reproducible Research in Robotics: Current Status and Road Ahead*, 2017.

[17] P. Lücking, F. Lier, J. Bernotat, S. Wachsmuth, S. Sabanovic, and F. A. Eyssel. Geographically distributed deployment of reproducible hri experiments in an interdisciplinary research context. pages 181–182, Chicago, IL,USA, 2018. ACM.

[18] Thomas Dolk, Bernhard Hommel, Lorenza S Colzato, Simone Schütz-Bosbach, Wolfgang Prinz, and Roman Liepelt. The joint simon effect: a review and theoretical integration. *Frontiers in Psychology*, 5, 2014.

[19] Joshua R De Leeuw. jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, 47(1):1–12, 2015.