Towards Reproducible Language-Based HRI Experiments: Open-Sourcing a Generalized Experiment Project

Uthman Tijani
University of South Florida
Tampa, Florida, USA
uthmantijani@usf.edu

Hong Wang
University of South Florida
Tampa, Florida, USA
hongw@usf.edu

Zhao Han
University of South Florida
Tampa, Florida, USA
zhaohan@usf.edu

ABSTRACT

We are witnessing increasing calls for reproducibility and replicability in HRI studies to improve reliability and confidence in empirical findings. One solution to facilitate this is using a robot platform that researchers frequently use, making it easier to replicate studies to verify results. In this work, we focus on a popular, affordable, and rich-in-functionality robot platform, NAO/Pepper, and contribute a generalized experiment project specifically for conducting language-based HRI experiments where a robot instructs a human for a task, including objective data collection.

Specifically, we first describe a concrete workflow from an existing experiment and how it is generalized. We then evaluate the generalized project with a case study to show how adopters can quickly adapt to their specific experiment needs. This work provides inspiration for HRI researchers to not only provide their experiment code as supplementary material but also generalize them to benefit other researchers to advance empirical research in HRI. The generalized Choregraphe project with documentation, demo, and usage notes is available under MIT license on GitHub at https://github.com/TheRARELab/langex. We welcome questions by posting GitHub issues and pull requests to share adapted packages.

CCS CONCEPTS

• **Human-centered computing** → **Systems and tools for interaction design**; • **Computer systems organization** → **Robotics**.

KEYWORDS

Open science; reproducibility; replicability; language-capable robots; Choregraphe; user studies; experiment design

ACM Reference Format:
Uthman Tijani, Hong Wang, and Zhao Han. 2024. Towards Reproducible Language-Based HRI Experiments: Open-Sourcing a Generalized Experiment Project. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3610977.3637483

HRI '24, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0322-5/24/03
https://doi.org/10.1145/3610977.3637483

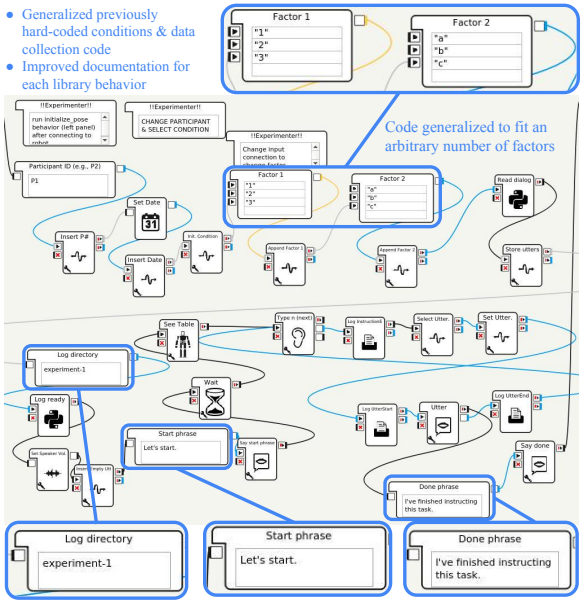

Figure 1: The generalized Choregraphe experiment project, with major non-code modifications circled in blue. It is common for experimenters to implement their own software for experiments. This work focuses on the most popular robot platforms, NAO/Pepper, contributing a generalized project for experimenters to reuse to improve reproducibility.

1 INTRODUCTION

To advance empirical research in fields like human-robot interaction (HRI), tightly controlled [18] or in-the-wild experiments [19] have been the cornerstone for contributing to human's knowledge of the perception and interaction of robots. To evaluate robot design, interactions, or algorithms, experimenters must develop not only them, the base condition, and a state-of-the-art method to compare, but also data collection for data analysis. To carry out evaluations, researchers often need to develop their own data collection methods. While there are well-validated subjective metrics [48] with reusable survey questions, oftentimes objective data, such as task completion time, is collected in their own code that roboticists must implement on their own. This distracts researchers from their focus on answering the research questions at hand and increases time spent on the procedure rather than the research itself.

Moreover, similar to the Psychology field that is currently experiencing a replication crisis [43, 47], the HRI field starts seeing increasing calls for reproducible and replicable studies [3, 7, 12, 15, 25, 45]

to increase the confidence and trust in the findings of the user studies as well as allow researchers to easily build upon each other's work to facilitate faster advancement of the field.

One of the barriers to replicability is caused by the use of different robot platforms [3] that have different appearances and physical limitations, leading to the generalizability issue where researchers are not able to replicate on their own robot [12]. While this issue can be solved by conducting experiments on different robot platforms, the cost factor of current robot platforms severely limits the number of robot platforms researchers could afford to evaluate their work.

One solution to the robot platform difference issue is to use one particular robot, e.g., the widely-used small NAO robot [39], manufactured by Aldebaran [38]. In a review article, Amirova et al. [1] found over 300 works in the literature from 2010 to 2020 that used NAO as their research platform of choice in a wide variety of domains like education (e.g., [9, 23]), healthcare (e.g., [24]), and industry [1, 31]. Indeed, thanks to NAO's affordability and broad functionality, it has become the de facto standard platform for conducting social robotics research [1], in addition to being the standard robot for the Standard Platform League of RoboCup [21].

Yet, despite NAO's popularity, there have been no software packages readily available for experimenters to modify and reuse to conduct experiments, including collecting common objective metrics [48] such as task completion time. Providing such an experiment package will not only help individual experimenters but also facilitate reproducibility and replicability through a standardized effort on a *de facto* standard robot platform, benefiting the whole HRI community at large.

In this work, we contribute a Choregraphe[1] package (Figure 1) generalized from two language-based HRI experiments where a robot instructed a human to finish a multi-step task. The scope and applicability lie in robot instruction tasks, which are a common type of task in HRI within the domains where robots use language [44]. Tellex et al. [44] have identified several important domains. In educational settings [4], a robot can be a storyteller to children [26], assist with language development [5, 46], or be an intelligent tutor, e.g., solving puzzles for problem-solving skills [29]. In socially assistive robotics, robots can also instruct older adults to engage in physical exercises [10, 13]. In collaborative robotics, robots need to generate ordered instructions for correct and efficient assembly [22, 41]. Robots in those domains work with people, and human-subjects evaluations are needed. For a full review of robots that use language, we refer readers to Tellex et al. [44].

With this work, in addition to facilitating conducting experiments and reproducibility, we hope to promote the reuse of one-off experiment code by inspiring other HRI researchers on how to generalize their experiment code for reuse, in addition to providing them as is in supplementary experiment material.

2 CONCRETE LANGUAGE-BASED HRI STUDY

We first discuss a specific workflow from our published language-based study [16] that retains the core elements necessary for conducting language-based HRI experiments. In the user study, the authors evaluated a computational referring form selection model in a live HRI scenario by replacing existing referring forms in a dialog with predicted references. For the task, a Pepper robot instructed a human to construct buildings using repeated wood blocks in nine shapes (cf. [27]). The model was compared with a random baseline and a full description condition. As part of their supplementary material [17] to the paper itself, a Choregraphe project was open sourced and its Choregraphe flow diagram is shown in Figure 2.

To help understand how the experiment works, we extracted a flow chart (Figure 3) from the Choregraphe project on how the experiment was prepared (top row) and how instructions were confirmed and delivered (middle row and bottom left group), including data collection nodes with a blue background. This flow chart also

[1]Choregraphe is a software to design robot behavior for NAO/Pepper [14], with modular boxes for pre-defined and custom behaviors programmable by Python.

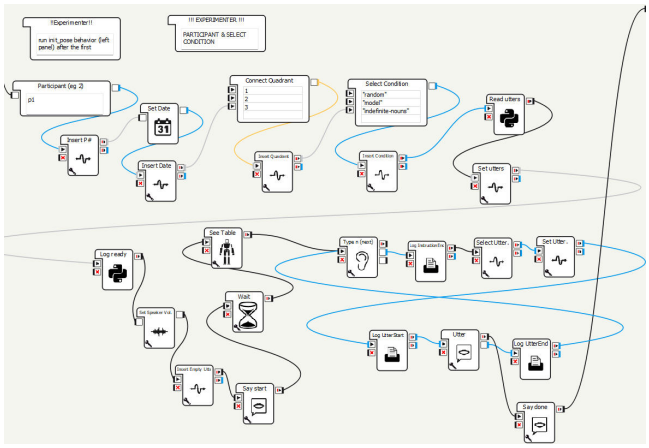

Figure 2: The as-is Choregraphe flow diagram [34] for the referring form experiment project [17], simplified in Figure 3. The enlarged, original version is available at https://osf.io/7akvz. As readers go over Section 2, it would be beneficial and convenient to reference the enlarged version.

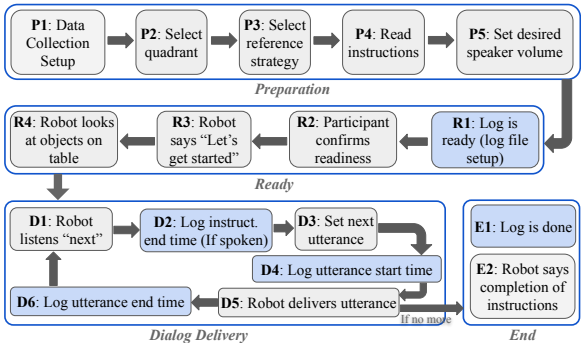

Figure 3: The specific workflow of a user study by Han and Williams [16]. Nodes in blue are for data collection (See Section 3). The first row with nodes' names prefixed P (for Prepare) contains preparation nodes. The second row with nodes' names prefixed R (for Ready) contains getting-ready nodes. The third circular row contains nodes where an interactive dialog was given utterance by utterance, with D for Dialog. The last E2 node indicates the completion of a dialog.

serves for comparison purposes to compare with the generalized workflow that will be discussed in Section 4. Below we examine the workflow in detail. To keep us focused on the general workflow, the data collection nodes will be discussed in Section 3.

(1) Data collection setup (P1): First (P1 and the top left nodes in Figure 3), a few metadata needs to be set for the current participant so relevant data like instruction completion time can be associated with the participant. The metadata includes anonymized participant ID and the date/time when the instruction ends and the utterance starts to calculate the utterance completion time.

(2) Select Condition & Read Instructions (P2-P4): After the participant meta information is stored, two factors in the experiment conditions are set, i.e., quadrant and referring form selection strategy. As seen from Figure 2, each factor has three levels, making a total of nine combinations, although in the experiment, Latin square was used to use fewer combinations to counterbalance ordering and learning effects. Under the hood, a different dialog was chosen for the robot to utter based on the resulting condition.

(3) Set Desired Speaker Volume (P4): Effective communication in an HRI scenario partially relies on audio clarity, and setting the desired speaker volume makes sure that bad clarity does not confound the findings. In the experiment, the authors adjusted the robot's volume level based on the distance between the robot and the participant sitting in a particular quadrant, ensuring that the spoken instructions were audible while not being too loud. This setting was used throughout the experiment.

(4) Participant and Robot Get Ready (R2–R4): Before the robot delivers the first utterance, these steps ensure the participant confirms their readiness by speaking "Ok" (R2), the robot responds with a start phase (i.e., "Let's get started"), and finally the robot moves to an initial pose like looking at the objects on the table (R4). R4 accepts a pose relative to, e.g., the robot's torso.

(5) Robot Listens for "next" (D1): Then the robot transitions into a listening mode, attentively awaiting a predefined "wake-up word" or trigger from the participant, such as "next". However, it is important to note that, rather than relying on the robot's autonomous speech recognition, a Wizard of Oz technique [30] (experimenters manually control but with participants believing the agent's autonomy) was used for the next instruction. This choice was made due to the complexities and challenges associated with implementing reliable speech recognition, which may lead to confounding results caused by imperfect speech recognition, i.e., adding extra time for instruction completion. In Han and Williams [16]'s work, experimenters used a text input to control the robot to move to the next utterance after hearing "ok".

(6) Robot Instructs (D3 and D5): This step involves the iterative process of the robot's instruction delivery and the participant's confirmation before the robot speaks the next utterance. The robot initiates the interactive dialog by providing instructions to the participant. These instructions are generated based on the condition selected at P3 and P4 in Figure 3. The iterative nature of this process allows for the sequential progression of the dialog, with each instruction followed by participant confirmation to the next one.

(7) Completion of Dialog (E2): The experiment ends with the last instruction for the task. Once all instructions have been completed, the robot announces, "Done." In Han and Williams [16]'s experiment, the robot says, "I have finished all instructions for this building." This clear indication signals the conclusion of the interaction, providing experimenters the opportunity to ask participants to conduct post-condition surveys.

Table 1: Sample Raw CSV Data With Headers

PID	Condition	Utterance	Event	Timestamp
P1	Random	"Bring in a red arch."	UtterStart	*(omitted)*
P1	Random	"Bring in a red arch."	UtterEnd	*(omitted)*
P1	Random	"Bring in a red arch."	InstructEnd	*(omitted)*
...	...	...	...	...

3 BUILT-IN OBJECTIVE DATA COLLECTION

The data collection approach in our generalized project mainly focuses on objective metrics such as task performance, including task completion time. That is, how long the participants were able to complete the task, implicitly answering the extent of the instructions given by the robot was understandable for them to carry out. The time taken by each participant for each utterance completion is recorded and stored in CSV format on the robot.

As seen in Figure 4, the "stopwatch" starts right before the instruction has been delivered (D4 in Figure 4) and ends after the utterance was delivered (D6 in Figure 4). There is another node (D2) that logs instruction end time. Each time log line (See Table 1) is appended to a separate file, right after a node was executed, for each participant with their corresponding utterances. The log is then used to analyze the time differences (i.e., $t^{InstructEnd} - t^{UtterEnd}$) in instruction completion for each condition, excluding the time spent on speaking the utterance (i.e., $t^{UtterEnd} - t^{UtterStart}$). An R script that combines all log files and calculates the instruction completion time is in the "proprocess-log" folder within the repository.

Subjective Data Collection: While the data collection effort was mostly focused on collecting objective data in the software implementation, subjective metrics like understandability can also be collected through questionnaires as usual after participants experience a condition, i.e., summative assessments. For formative assessments seen in mixed-method design, e.g., participants' emotions, video and audio data collection is recommended for qualitative coding [2] by setting up an external camcorder because NAO [37] and Pepper [32] have a limited horizontal field of view ($< 70°$).

4 GENERALIZATION

The workflow in Figure 3 shows the robot's behavior and interaction with participants during Han and Williams [16]'s experiment. In one of our current works under the experimentation phase, we were able to generalize the project to fit our language-based HRI experiment, ending up with another specific Choregraphe project. This inspired us to contribute a generalized Choregraphe package that is not tied to specific experiments like those two.

In this section, we describe how we generalized from the original Choregraphe project and visit a case study embedded in this general project on how experimenters can adapt the generalized project to another language-based experiment.

Figure 4 shows the generalized version of the workflow in Figure 3, and Figure 1 shows the corresponding Choregraphe flow diagram. The difference at the surface level is the replacement of

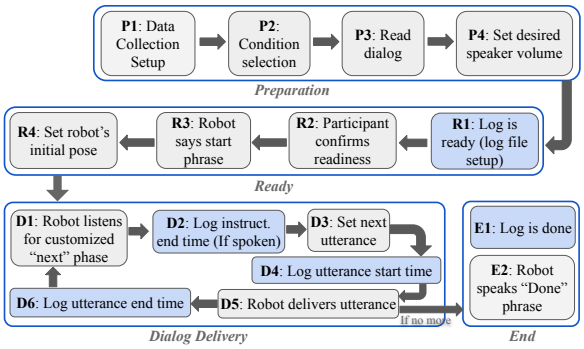

Figure 4: Generalized workflow for conducting language-based HRI user studies. At a high level, the condition selection is generalized to add an arbitrary number of factors. Other hard-coded values can be customized (See Section 4).

the quadrant and referring form strategy selection nodes (P2 and P3 in Figure 3), as well as removals of hard-coded values.

The quadrant selection node is particular to the experimental setup used in [16]. In that user study, participants sat in three of four quadrants and were instructed by the robot to complete three building construction tasks in each quadrant. The constituting blocks for each building were distributed to all three quadrants to test references to blocks not present in the scene.

In the generalized package, we now have two factor nodes, as seen in Figure 1 top right. When the "Factor 2" node is duplicated, the package allows for an arbitrary number of factors. Under the hood, we removed the hard-coded quadrant and referring form selection variable throughout the Python code in each workflow node and added code to concatenate and memorize different factors for use in later nodes in the workflow. For example, The "Log is ready" node (R1 in Figure 4) retrieves the condition string to create individual log files.

Additionally, we have added the start phrase (R3), end phrase (E2), and "log directory" nodes for easy customization, as seen in the bottom row of the Choregraphe flow diagram in Figure 1.

Along the generalization process, we also added documentation for each node, whereas the previous documentation only had placeholder text. For example, the "Insert P#" node has this description: "Stores participant ID to Exp/Participant in NAOqi's shared memory." On a related note, all data was prefixed with "Exp" and can be monitored in the memory watcher panel [36].

4.1 Case Study: How to Use and Adapt

To use this general package, one first follows the system requirement and installation guide in the GitHub repository. Then the project can be opened in Choregraphe to start the customization process. The current project is named "general-experiment" and it should be changed in the property of the project and by editing the "manifest.xml" file. Then, experimenters should have a list of factors in hand and the levels of each factor. With them, the "Factor 1" and "Factor 2" nodes should be changed both in their name and the list content. If there are more than three factors, the "Factor 2" *and* "Append Factor 2" should be duplicated to ensure the extra factors appear in the log filename, log header, and log rows.

Once these factors are specified in the Choregraphe flow diagram, the dialog content needs to be placed in the "utterances" directory at the project root. Each condition should have a dialog file named "factor1-factor2-factor3.txt". Then, to specify where the log directory is, simply change the content of the "Log directory" node. For start and end phrases, they can also be changed in the two text box nodes (See the bottom row in Figure 1).

One last change is the robot's initial behavior, which can be found in the "initialize_pose" folder. To change the robot's pose, e.g., to look at the table in front of it, after it speaks the start phase, the node "See Table" (See mid-left of Figure 1) can be used for this purpose by changing its parameters, the offset values (x, y, z) and the reference frame, e.g., torso.

To test the program, it follows the same procedure as running a hello-world Choregraphe application [35]. As the program starts running, log files will be created in "/home/nao/[log directory]/" on the robot and the R script can be used to analyze them. A video[2] is available to show the expected output.

5 LIMITATIONS

Human evaluation takes a variety of forms, and this work has focused on the specific experimental research [18] with an objective data collection component and a brief discussion on mixed-method design in Section 3. While experiments are good at isolating factors of interest and controlling confounding factors, we advocate, as Fischer [11], methodological pluralism and acknowledge other approaches that provide better ecological validity (representing the wider world or highly dynamic, contextualized real-world human interactions), e.g., participatory design [42] and contextual analysis [8]. In these non-experimental approaches, our evaluation-related code and data collection may need to be further customized. Yet, it is hard to find one solution that fits all, or fully confirm the superiority of one particular method, and the experimental method is part of methodological pluralism. Future work should investigate how to provide a generalized package for these kinds of evaluations. Additionally, we hope to see more generalized experiment work for other robot platforms, using the popular ROS framework [28], or in non-language and multimodal domains, including emotion [20], non-verbal cues [6], and non-dyadic group environments [40].

6 CONCLUSION

We contribute a generalized experiment package along with the generalization process of a language-based experiment. It can be used on the most popular NAO/Pepper platform (cf. [1]) to facilitate replication and comparison of findings, which is important before robots' cost factor decreases. A case study showed the adaption of experiment conditions, utterances, and initial behavior. Additional behaviors can be supported by Choregraphe through its behavior design functionalities [33]. With this work, we provide inspiration for and raise awareness of the practice of generalizing HRI researchers' one-off experiment code. So researchers, especially those in need of evaluating robots that use language, can easily build upon each other's work to better advance the HRI field.

[2] https://www.youtube.com/shorts/5plLXZN3D0E

REFERENCES

[1] Aida Amirova, Nazerke Rakhymbayeva, Elmira Yadollahi, Anara Sandygulova, and Wafa Johal. 2021. 10 years of human-nao interaction research: A scoping review. *Frontiers in Robotics and AI* 8 (2021), 744526.

[2] Carl Auerbach and Louise B Silverstein. 2003. *Qualitative data: An introduction to coding and analysis*. Vol. 21. NYU press.

[3] Shelly Bagchi, Patrick Holthaus, Gloria Beraldo, Emmanuel Senft, Daniel Hernandez Garcia, Zhao Han, Suresh Kumaar Jayaraman, Alessandra Rossi, Connor Esterwood, Antonio Andriella, et al. 2023. Towards Improved Replicability of Human Studies in Human-Robot Interaction: Recommendations for Formalized Reporting. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 629–633.

[4] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social robots for education: A review. *Science robotics* 3, 21 (2018), eaat5954.

[5] Cynthia Breazeal, Paul L Harris, David DeSteno, Jacqueline M Kory Westlund, Leah Dickens, and Sooyeon Jeong. 2016. Young children treat robots as informants. *Topics in cognitive science* 8, 2 (2016), 481–491.

[6] Elizabeth Cha, Yunkyung Kim, Terrence Fong, Maja J Mataric, et al. 2018. A survey of nonverbal signaling methods for non-humanoid robots. *Foundations and Trends® in Robotics* 6, 4 (2018), 211–323.

[7] Julia R Cordero, Thomas R Groechel, and Maja J Matarić. 2022. A review and recommendations on reporting recruitment and compensation information in HRI research papers. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1627–1633.

[8] 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*. 224–233.

[9] Jan de Wit, Arold Brandse, Emiel Krahmer, and Paul Vogt. 2020. Varied human-like gestures for social robots: Investigating the effects on children's engagement and language learning. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*. 359–367.

[10] Juan Fasola and Maja J Matarić. 2015. Evaluation of a spatial language interpretation framework for natural human-robot interaction with older adults. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 301–308.

[11] Kerstin Fischer. 2021. Effect confirmed, patient dead: A Commentary on Hoffman & Zhao's Primer for Conducting Experiments in HRI. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 1 (2021), 1–4.

[12] Marlena R Fraune, Iolanda Leite, Nihan Karatas, Aida Amirova, Amélie Legeleux, Anara Sandygulova, Anouk Neerincx, Gaurav Dilip Tikas, Hatice Gunes, Mayumi Mohan, et al. 2022. Lessons learned about designing and conducting studies from hri experts. *Frontiers in Robotics and AI* 8 (2022), 772141.

[13] Binnur Görer, Albert Ali Salah, and H Levent Akın. 2017. An autonomous robotic exercise tutor for elderly people. *Autonomous Robots* 41 (2017), 657–678.

[14] Erico Guizzo. 2014. How Aldebaran robotics built its friendly humanoid robot, Pepper. *IEEE Spectrum* (2014).

[15] Hatice Gunes, Frank Broz, Chris S Crawford, Astrid Rosenthal-von der Pütten, Megan Strait, and Laurel Riek. 2022. Reproducibility in human-robot interaction: furthering the science of HRI. *Current Robotics Reports* 3, 4 (2022), 281–292.

[16] Zhao Han and Tom Williams. 2023. Evaluating Cognitive Status-Informed Referring Form Selection for Human-Robot Interactions. In *2023 Annual Meeting of the Cognitive Science Society (CogSci)*.

[17] Zhao Han and Thomas Williams. 2023. Supplementary Materials for "Evaluating Cognitive Status-Informed Referring Form Selection for Human-Robot Interactions"'. https://doi.org/10.17605/OSF.IO/NZWVF

[18] Guy Hoffman and Xuan Zhao. 2020. A primer for conducting experiments in human–robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 1 (2020), 1–31.

[19] Malte Jung and Pamela Hinds. 2018. Robots in the wild: A time for more robust theories of human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)* 7, 1 (2018), 1–5.

[20] Rachel Kirby, Jodi Forlizzi, and Reid Simmons. 2010. Affective social robots. *Robotics and Autonomous Systems* 58, 3 (2010), 322–332.

[21] Hiroaki Kitano, Minoru Asada, Yasuo Kuniyoshi, Itsuki Noda, and Eiichi Osawa. 1997. Robocup: The robot world cup initiative. In *Proceedings of the first international conference on Autonomous agents*. 340–347.

[22] Ross A Knepper, Todd Layton, John Romanishin, and Daniela Rus. 2013. Ikeabot: An autonomous multi-robot coordinated furniture assembly system. In *2013 IEEE International conference on robotics and automation*. IEEE, 855–862.

[23] Elly A Konijn, Brechtje Jansen, Victoria Mondaca Bustos, Veerle LNF Hobbelink, and Daniel Preciado Vanegas. 2022. Social robots for (second) language learning in (migrant) primary school children. *International Journal of Social Robotics* (2022), 1–17.

[24] Namyeon Lee, Jeonghun Kim, Eunji Kim, and Ohbyung Kwon. 2017. The influence of politeness behavior on user compliance with social robots in a healthcare service setting. *International Journal of Social Robotics* 9 (2017), 727–743.

[25] Benedikt Leichtmann, Verena Nitsch, and Martina Mara. 2022. Crisis ahead? Why human-robot interaction user studies may have replicability problems and directions for improvement. *Frontiers in Robotics and AI* 9 (2022), 838116.

[26] Iolanda Leite, Marissa McCoy, Monika Lohani, Daniel Ullman, Nicole Salomons, Charlene Stokes, Susan Rivers, and Brian Scassellati. 2015. Emotional storytelling in the classroom: Individual versus group interaction between children and robots. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*. 75–82.

[27] Melissa & Doug. 2019. 100 Piece Wood Blocks Set. https://www.melissaanddoug.com/products/100-piece-wood-blocks-set. Accessed: 2023-11-17.

[28] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, Andrew Y Ng, et al. 2009. ROS: an open-source Robot Operating System. In *ICRA workshop on open source software*, Vol. 3. Kobe, Japan, 5.

[29] Aditi Ramachandran, Chien-Ming Huang, Edward Gartland, and Brian Scassellati. 2018. Thinking aloud with a tutoring robot to enhance learning. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*. 59–68.

[30] 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.

[31] Adam Robaczewski, Julie Bouchard, Kevin Bouchard, and Sébastien Gaboury. 2021. Socially assistive robots: The specific case of the NAO. *International Journal of Social Robotics* 13 (2021), 795–831.

[32] Aldebaran Robotics. 2018. 2D Cameras — Aldebaran 2.5.11.14a documentation. http://doc.aldebaran.com/2-5/family/pepper_technical/video_2D_pep.html.

[33] Aldebaran Robotics. 2018. Choregraphe - Tutorials — Aldebaran 2.5.11.14a documentation. http://doc.aldebaran.com/2-5/software/choregraphe/tutos/index.html.

[34] Aldebaran Robotics. 2018. Flow diagram panel — Aldebaran 2.5.11.14a documentation. http://doc.aldebaran.com/2-5/software/choregraphe/panels/flow_diagram_panel.html.

[35] Aldebaran Robotics. 2018. Hello World 1 - using Choregraphe — Aldebaran 2.5.11.14a documentation. http://doc.aldebaran.com/2-5/getting_started/helloworld_choregraphe.html.

[36] Aldebaran Robotics. 2018. Memory watcher panel — Aldebaran 2.5.11.14a documentation. http://doc.aldebaran.com/2-5/software/choregraphe/panels/memory_watcher_panel.html.

[37] Aldebaran Robotics. 2018. Video camera — Aldebaran 2.5.11.14a documentation. http://doc.aldebaran.com/2-5/family/robots/video_robot.html.

[38] Aldebaran Robotics. 2023. Aldebaran | Humanoid and programmable robots. https://www.aldebaran.com/en/.

[39] Aldebaran Robotics. 2023. NAO the humanoid and programmable robot | Aldebaran. https://www.aldebaran.com/en/nao.

[40] Sarah Sebo, Brett Stoll, Brian Scassellati, and Malte F Jung. 2020. Robots in groups and teams: a literature review. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–36.

[41] Kevin Spevak, Zhao Han, Tom Williams, and Neil T Dantam. 2022. Givenness hierarchy informed optimal document planning for situated human-robot interaction. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 6109–6115.

[42] Laura Stegner, Emmanuel Senft, and Bilge Mutlu. 2023. Situated participatory design: A method for in situ design of robotic interaction with older adults. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[43] Wojciech Świątkowski and Benoît Dompnier. 2017. Replicability crisis in social psychology: Looking at the past to find new pathways for the future. *International Review of Social Psychology* 30, 1 (2017), 111–124.

[44] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems* 3 (2020), 25–55.

[45] Daniel Ullman, Salomi Aladia, and Bertram F Malle. 2021. Challenges and opportunities for replication science in hri: A case study in human-robot trust. In *Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction*. 110–118.

[46] Rianne Van den Berghe, Josje Verhagen, Ora Oudgenoeg-Paz, Sanne Van der Ven, and Paul Leseman. 2019. Social robots for language learning: A review. *Review of Educational Research* 89, 2 (2019), 259–295.

[47] Bradford J Wiggins and Cody D Christopherson. 2019. The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology* 39, 4 (2019), 202.

[48] Megan Zimmerman, Shelly Bagchi, Jeremy Marvel, and Vinh Nguyen. 2022. An analysis of metrics and methods in research from human-robot interaction conferences, 2015–2021. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 644–648.