Towards Improved Replicability of Human Studies in Human-Robot Interaction: Recommendations for Formalized Reporting

Shelly Bagchi*
U.S. National Institute of Standards and Technology
Gaithersburg, MD, USA

Patrick Holthaus
University of Hertfordshire
Hatfield, United Kingdom

Gloria Beraldo
National Research Council of Italy
Rome, Italy

Emmanuel Senft
Idiap Research Institute
Martigny, Switzerland

Daniel Hernández García
Heriot-Watt University
Edinburgh, United Kingdom

Zhao Han
Colorado School of Mines
Golden, CO, USA

Suresh Kumaar Jayaraman
Cornell University
Ithaca, NY, USA

Alessandra Rossi
University of Naples Federico II
Naples, Italy

Connor Esterwood
University of Michigan
Ann Arbor, MI, USA

Antonio Andriella
Pal Robotics
Barcelona, Spain

Paul Pridham
University of Michigan
Ann Arbor, MI, USA

ABSTRACT

In this paper, we present a proposed format for reporting human studies in Human-Robot Interaction (HRI). We call for details which are often overlooked or left out of research papers due to space constraints, and propose a standardized format to contain those details in paper appendices. Providing a formalized study reporting method will promote an increase in replicability and reproducibility of HRI studies and encourage meta-analysis and review, ultimately increasing the generalizability and validity of HRI research. Our draft is the first step towards these goals, and we welcome feedback from the HRI community on the included topics.

CCS CONCEPTS

• **Computer systems organization** → **Robotics**; • **Human-centered computing** → **User studies**; • **Information systems** → Data replication tools.

KEYWORDS

human studies, reporting, replicability, reproducibility, guidelines

ACM Reference Format:
Shelly Bagchi, Patrick Holthaus, Gloria Beraldo, Emmanuel Senft, Daniel Hernández García, Zhao Han, Suresh Kumaar Jayaraman, Alessandra Rossi, Connor Esterwood, Antonio Andriella, and Paul Pridham. 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 (HRI '23 Companion), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3568294.3580162

*Corresponding author: shelly.bagchi@nist.gov. All authors contributed equally.

HRI '23 Companion, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9970-8/23/03…$15.00
https://doi.org/10.1145/3568294.3580162

1 INTRODUCTION

Standardization of methods is an ongoing effort in the Human-Robot Interaction (HRI) community, demonstrated by the ACM/IEEE HRI Conference's annual Workshop on Test Methods & Metrics[1], currently in its 5th iteration. Consistent interest in this workshop led its participants to form an IEEE Study Group on Standards for HRI in 2020 to explore the readiness of the state of research.

The IEEE Standards Association (SA) publishes standards documents and enables working groups to meet and collaboratively construct draft standards. In late 2021, after results from the study group were promising, the IEEE SA and Robotics & Automation Society (RAS) approved the creation of their first two standards groups in HRI: IEEE P3107, "Standard Terminology for Human-Robot Interaction"[2], and IEEE P3108, "Recommended Practice for Human-Robot Interaction Design of Human Subject Studies"[3].

IEEE P3108 aims to lay out a set of suggested guidelines for HRI researchers conducting human studies, as well as best practices for the design of HRI studies. In particular, some of the major goals of IEEE P3108 are to enable "human-subject studies to maintain standards for reporting, enable reproducibility and verification/validation studies, and to maximize the likelihood of results and methodologies being leveraged by other studies" [1]. In order to accomplish this goal, a subgroup of P3108 drafted a study reporting form to record important details that may not be included in

[1] https://hri-methods-metrics.github.io/
[2] https://standards.ieee.org/ieee/3107/10709/
[3] https://standards.ieee.org/ieee/3108/10710/

paper texts due to space considerations. This could enable better replicability and reproducibility of HRI studies, while also allowing both readers and reviewers to quickly assess the various significant elements of a study's design and results. The resulting form would be highly encouraged to be submitted in parallel to every study-centric submission in HRI-related venues.

In this paper, we present a preliminary draft of the reporting form and explain the included fields. By publishing the draft version, we hope to obtain some feedback from the HRI community regarding additional items or changes that would be of use for study reporting.

2 MOTIVATION

As the field of HRI is maturing, significant effort has been dedicated in the last decade to formalize what is meant by HRI research and how it should be conducted [2]. The need to create guidelines for designing and reporting HRI studies has been discussed in the literature over the last decade [3, 5, 17, 20] and is still a source of debate inside the community [8, 14, 15, 18]. Previous attempts have tried to identify the common aspects among diverse studies, especially in the field of social robotics.

An example of one such attempt is seen in Fraune et al., [10]. Within this paper, the authors summarize the central insights gained from the recent *Workshop Your Study Design* (WYSD) workshop held at the 2021 International Conference on Human-Robot Interaction. In particular, they contribute greatly to the question of how one can conduct experiments in HRI, but fall short in one important area. Namely, the authors offer no guidance or standards for the reporting of study designs, analyses, metadata, and results. While this is not entirely unexpected given the scope of the WYSD workshops, it is nonetheless important as without such standards, any cumulative research dependent on replication, reproducibility, or meta-analysis cannot reasonably be conducted.

Indeed, reporting the expected information in published papers is also connected to the important issue of replicability and reproducibility in HRI studies [11]. In addition to validating existing studies, replication efforts can provide novel insights on the generalizability of existing findings [19]. However, as mentioned by Cordero et al. [5], current HRI works lack essential information to reproduce and generalize their findings. This problem could depend on four reasons: a) the lack of a standard to address a multidisciplinary community composed of computer scientists, psychologists, engineers, philosophers, educators, and researchers belonging to other disciplines; b) the lack of meta-data that describes the demographics of participants, the study design, various variables like independent, dependent and control variables, and more, characterizing new research; c) the unsustainable practice of linking code and other support materials in the published papers, e.g., personal websites likely inaccessible in future; d) the absence of shared metrics to be applied in various scenarios. Moreover, other factors such as specific hardware, demographics, cultural identity, and mutable testing conditions might represent a barrier toward the standardization, replication, and reproducibility of consistent results across studies because of their influence on the interaction. It is worth clarifying that this work aims to improve reporting of human studies in order to facilitate the replication of the study's methods and research questions, rather than achieving the exact results.

In addition to replicability and reproducibility, standards also hold great promise for meta-analysis. Meta-analyses offer the field of HRI a way to overcome the limitations of any single study by leveraging the results of multiple studies in order to estimate a broad relationship between variables (i.e., effect size) [4, 16]. This method has gained popularity in the field of HRI in recent years (see: [6, 7, 12, 13]) but one key limitation of this methodological approach is that studies must report sufficient data to be included in such analyses. For example, [6] found 121 studies relevant to their meta-analysis but could only use 26 of these papers - less than 25%. Similarly, [20] reviewed metrics and scales across six years of HRI conferences. Although the vast majority used custom scales (which is itself a concern for the comparability of methods), the authors could not find the text of most surveys, which further damages the replicability of a study. By developing these standards and presenting them to the HRI community, this paper can further empower meta-analysis and evidence synthesis, and allows for more frequent and wider ranging meta-analyses in the field of HRI.

3 RECOMMENDED REPORTING FORMAT

The most recent draft of the full reporting form can be found at https://tinyurl.com/hri-study-reporting. In the following subsections, we will describe each section of the form and explain the included fields. Due to space constraints, we list the majority of items in text, but the actual document is intended to contain tables of questions and entry fields that can be filled in by a researcher and adjusted to meet their study parameters.

The specific fields were selected over multiple meetings of the IEEE P3108 subgroup on study reporting. Many of the items were motivated based on key study details that are lacking in current conference papers. The group reviewed a sample of papers and determined which additional details might assist with replicability. In addition, some fields were motivated by a desire for stronger, more consistent methodology in HRI studies. We observed recurring issues, such as small sample sizes without relevant power analysis or other justification. Including such relevant fields in our form may encourage researchers to think critically about elements that are overlooked in their reporting.

3.1 General Information

The reporting form begins with some basic information to ground the topic and context of the study. We understand that some of these fields may initially need to be redacted for anonymized submissions, but the goal version is presented here.

Study Title The title of the paper in which the human study is described.

Author Names Authors in the same order as on the paper.

Author Institutions Authors' place of work while conducting the study, such as universities, research institutes, or companies.

Study Setting Type of setting(s) that the research was conducted in: lab, field, online/crowd-sourced, etc.

City, Country of Research Geographical location(s) where the study was conducted.

Dates Conducted Period of time over which the study was conducted (possibly multiple date ranges).

Table 1: Suggested reporting format for detailed demographics

Category	*Justification* (why was this needed?)	*Breakdown/Statistics of Participants* (e.g. range, mean, max/min)	*Statistics of Target Population* (if known)
e.g., Age	To obtain a variety of viewpoints	18-45; mean 26 years	global mean 31 years
Gender	To obtain a representative sample; observe potential gender-specific effects	20 female (50%)	49.58% female
...	...	...	...

Institutional/Ethical Review Board Name of the board which reviewed the study and the approval number, if required by the researcher's institution, or reasons why it is not required.

Approval or Study Registration Link If available, link to the approved IRB/Ethical Review application or other study registration venue.

3.2 Participants/Recruitment

This section of the reporting form reviews how participants were handled during the study. The motivation is ethical handling of subjects, as well as revealing potential biases in the results due to the sample selection criteria or demographics.

Number of Participants Total number of participants across groups. If multiple iterations of the study were conducted, list the numbers of participants separately. If any participants voluntarily dropped out of the study, indicate this number/rate as well.

Recruitment Pool Characteristics of the people who were contacted to participate in the study. For example, university students or children, or your crowd-sourced participant selection criteria.

Recruitment Method Briefly explain how participants were contacted. For example, mailing list, bulletin board posting, class announcements, and personal solicitation.

Justification for the Number of Participants Briefly explain your power analysis or other justification for the number of participants in your study.

Compensation / Rewards If the participants received any compensation for their participation, indicate the amount and in what form.

Consent Type Briefly explain how participants consented to participate in the study. For example, informed consent form, consent by guardian or proxy, etc., or information sheet for survey/other exempt study.

Link to consent form or information sheet If applicable, include the link to the full consent form or information sheet. Although requirements may vary by institution, this provides a baseline for what the participants were informed of.

Demographics and Rationale behind the Collection If demographics were collected, list which categories were asked. For each category, provide justification for why they were needed in this study. Include a link to the demographics questionnaire, if possible, for wording replication. **Table 1** shows an example of how we recommend demographics reporting should be done.

Descriptive Statistics List the corresponding descriptive statistics of participants (e.g., mean and standard deviation of age; number of male and female participants).

Target Population of the Work The population(s) and/or research communities that the authors intend their research to be applied to or used by. For example, seniors, children with autism, manufacturing employees, etc.

Demographics of the Target Population For the criteria above, list the statistics of the target population, e.g., applications for the elderly can use the Organisation for Economic Co-operation and Development (OECD)'s elderly population data [9]. This can be easily compared to the participant pool demographics.

Pre-study Inclusion/Exclusion Criteria If participants were included/excluded based on a pre-survey, authors should indicate the criteria and how many were selected/eliminated per criterion.

3.3 Study Design

The Study Design section should contain the main technical details of the study that was conducted. This should be sufficiently detailed for an independent researcher to recreate the experimental structure.

Apparatus The platform(s) used in the study (i.e., what robot(s), device(s), and/or survey platform(s) were used).

Control Method (if applicable) Indicates how the system has been controlled (e.g., autonomous, Wizard of Oz, hybrid) and the control variables in the environment.

Interaction Method Details how participants were able to interact with the system (e.g., live interaction, simulation, remote interaction, video, audio, images).

Structure Report the structure of the experiment, for example whether the experiment was Within Subjects (participants performed more than one condition) or Between Subjects (participants performed only one condition). Additionally, how many participants per group or condition, and how many conditions. For example: $N \times M$ Between Subjects, where N is the number of participants and M is the number of groups (conditions). The conditions will be expanded upon in the following subsection.

Duration / Repeats / Intervals The duration of each individual trial, the number of repeats for each participant, and the time between trials.

Post-Study Exclusion Criteria Any criteria for exclusion, such as erroneous trials (e.g., system failure, unexpected participant behaviours, external reasons). Also include the corresponding number of excluded participants and/or trials per criterion.

Subjective and Objective Measures List all subjective measures (e.g., user preference, trust) and objective measures (e.g., timings, response rates, task performance). For subjective measures, list the questions and/or scales used for each measure. We recommend listing the measures along with the relevant hypotheses, as mentioned in the following subsection.

Description of Modifications to Existing Scales (if applicable) List how existing scales have been adapted to the study and whether they have been validated.

3.3.1 Design Elements. We imagine several elements of Study Design will require separate reporting tables, specific to the experimental structure, outlined below.

Table 2: Suggested reporting formats for hypotheses

Hypotheses	*Was it supported?*	*Statistical tests used*
e.g., 1. A slower robot is perceived as safer than a faster robot. 2. A slower robot is perceived as less performant than a faster robot.	Yes / No / Partially (explain or reference paper)	e.g. ANOVA
Predictions (i.e., measurable differences between conditions)	*Was it supported?*	
e.g., 1. Safety$_{CondA}$ > Safety$_{CondB}$	Yes / No / Partially (explain or reference paper)	p-value
2. Performance$_{CondA}$ > Performance$_{CondB}$	Yes / No / Partially (explain or reference paper)	p-value

Table 3: Suggested reporting format for design variables

Independent Variables (conditions)	*Parameters*
e.g., type of control	joystick, tablet, verbal
...	...
Dependent Variables (measures)	*Metrics Used*
e.g., task performance	% correct
mental workload	NASA TLX
...	...
Control Variables	*Metrics Used*
e.g., environmental noise	sound meter (dB)
time of day	clock
...	...

Hypotheses: List the hypotheses your experiment attempted to verify, your predicted results, and whether they were supported. Also list which statistical tests were used to determine significant results. **Table 2** shows an example of hypotheses reporting.
We additionally ask authors to answer:
Were your hypotheses pre-registered? Yes (provide link) / No
Were your predictions pre-registered? Yes (provide link) / No

Variables: List the Independent Variables (conditions), Dependent Variables (measures), and any other Control Variables relevant to the experiment. **Table 3** shows an example of how we visualize the reporting of design variables should be done.

Group Layout: Report the way in which your participants were grouped, within to the experimental structure reported previously. For Within Subjects studies, the order of conditions within each group should be listed as well. If possible, additionally report the demographic statistics of each individual group. An example of a structured table for reporting this can be found in the full reporting form: https://tinyurl.com/hri-study-reporting.

Statistical Tests Summary. In this free-form section, the authors should include information to help other researchers reproduce their statistical tests. For example, ANOVA (Analysis of Variance) tables.

3.4 Code & Data Release

This section collects information about where the study's code and/or collected datasets can be found. This can include algorithms and scripts for data collection during the study, scripts for post-study data cleaning and analyses, or any other code that would be valuable for replication. This enables researchers to more quickly build upon others' work. Although it is understandable that in some cases proprietary code cannot be shared, pre-processed and anonymized data as well as code/software used to generate statistics are highly encouraged to allow reproduction of results.

Code Availability Information on the code availability (e.g., open source, available by request, proprietary/not publicly available). The authors should indicate how code can be obtained by other researchers (if available). They may also include licensing information for code reuse. If authors were unable to include online access to these items, they may instead point to a file path within a supplemental archive.

Repositories or notebooks Link to collections of code files, such as GitHub repositories, Google Colab, Jupyter Notebook, or R Markdown files.

Data collection scripts These can include motion capture code and audio/video collection scripts, in addition to *verbal/text* scripts for study instructions given by the administrators.

Data cleaning and analyses scripts If the study's data cleaning and analyses scripts are available, they can be linked here for use by other researchers to ensure that post-processing and results are found in similar ways.

Online datasets If available, link to the dataset(s) used. For example, a Zenodo or other Digital Object Identifier (DOI) link that should be persistent.

4 CONCLUSION & FUTURE WORK

In this paper, we introduced the first draft of the IEEE P3108 standards committee's recommendations for study reporting. Better reporting is needed to increase the transparency of HRI research. A standardized set of items and format for reporting will facilitate study comparisons and enable reproduction of human studies in HRI. Furthermore, its use as a paper appendix or supplemental material can free up space in the main paper for more detailed technical methods, results, and discussions. We imagine that, in time, as the form becomes more comprehensive, researchers may also consult this as a guide when planning a study to remind them of factors to consider, increasing the strength of their initial results.

Our community currently lacks replication studies due to the difficulty of obtaining information about past studies; documentation is inconsistent, and often knowledge is lost as researchers transition into other positions. Facilitating replication will additionally lead to better validity of HRI research, strengthening methods and results.

In the future, the group aims to create an online database where HRI researchers can register their study details by filling out a standard form. This will allow the community to browse studies, quickly find relevant work, and increase replicability. Creating a database of studies would also enable large-scale meta-analyses to be done automatically. We believe there is a significant advantage in moving towards more standardized methods in HRI, and our proposed study reporting form is a first step in that direction.

REFERENCES

[1] IEEE Standards Association. 2021. *IEEE myProject P3108 PAR Details.* https://development.standards.ieee.org/myproject-web/public/view.html#pardetail/9314

[2] Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Šabanović. 2020. *Human-robot interaction: An introduction.* Cambridge University Press.

[3] Paul Baxter, James Kennedy, Emmanuel Senft, Severin Lemaignan, and Tony Belpaeme. 2016. From characterising three years of HRI to methodology and reporting recommendations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 391–398.

[4] Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. 2011. *Introduction to meta-analysis.* John Wiley & Sons.

[5] 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.

[6] Connor Esterwood, Kyle Essenmacher, Han Yang, Fanpan Zeng, and Lionel Peter Robert. 2021. A meta-analysis of human personality and robot acceptance in human-robot interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–18.

[7] Connor Esterwood, Kyle Essenmacher, Han Yang, Fanpan Zeng, and Lionel P Robert. 2022. A Personable Robot: Meta-Analysis of Robot Personality and Human Acceptance. *IEEE Robotics and Automation Letters* 7, 3 (2022), 6918–6925.

[8] 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.

[9] Organisation for Economic Co-operation and Development (OECD). 2022. *Demography - Elderly population - OECD Data.* Retrieved 2022-12-21 from https://data.oecd.org/pop/elderly-population.htm

[10] 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* (2022), 401.

[11] 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 (Dec. 2022), 281–292. https://doi.org/10.1007/s43154-022-00094-5

[12] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[13] Peter A Hancock, Theresa T Kessler, Alexandra D Kaplan, John C Brill, and James L Szalma. 2021. Evolving trust in robots: specification through sequential and comparative meta-analyses. *Human factors* 63, 7 (2021), 1196–1229.

[14] 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.

[15] 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).

[16] Mark W Lipsey and David B Wilson. 2001. *Practical meta-analysis.* SAGE publications, Inc.

[17] 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.

[18] Johanna Seibt, Christina Vestergaard, and Malene F Damholdt. 2021. The Complexity of Human Social Interactions Calls for Mixed Methods in HRI: Comment on "A Primer for Conducting Experiments in Human-robot Interaction," by G. Hoffman and X. Zhao. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 1 (2021), 1–4.

[19] 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.

[20] 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 *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (Sapporo, Hokkaido, Japan) *(HRI '22).* IEEE Press, 644–648.