An Analysis of Metrics and Methods in Research from Human-Robot Interaction Conferences, 2015-2021

Megan Zimmerman, Shelly Bagchi, Jeremy Marvel, Vinh Nguyen
U.S. National Institute of Standards and Technology
Gaithersburg, MD, USA
{megan.zimmerman,shelly.bagchi,jeremy.marvel,vinh.t.nguyen}@nist.gov

2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI) | 978-1-6654-0731-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/HRI53351.2022.9889595

***Abstract*—Standardized metrics and methods are critical towards wider adoption of HRI technologies in real-world applications. However, the interdisciplinary nature of HRI creates an inherently decentralized research paradigm that limits the use of standardized metrics for baseline comparisons among studies. This limitation restricts both the real-world adoption and academic replicability of HRI solutions developed by the research community. To identify specific opportunities for reuse of metrics and methods in HRI, this paper presents a comprehensive survey of 1464 papers from the ACM/IEEE International Conference on Human-Robot Interaction (HRI) and the IEEE International Conference on Robot and Human Interactive Communication (Ro-Man) over seven years. By providing a holistic perspective of the metrological tools leveraged in the current state-of-practice of HRI research, we find that a significant portion of HRI studies use custom surveys, thus limiting baseline comparison. Hence, the analysis in this work aims to advance the field of HRI by identifying specific barriers to adoption of HRI technologies in addition to proposing solutions to overcome existing limitations in the context of metrics and methodologies.**

***Index Terms*—performance metrics, test methods, HRI metrology, research trends, reproducibility, replication, robotics**

I. Introduction

Research in human-robot interaction (HRI) relies on interdisciplinary collaboration from a host of fields including Computer Science, Psychology, Engineering, Ethics, Biology, Economics, and others. While HRI research benefits from such interdisciplinary teamwork through the development of innovative ideas and solutions, the field also experiences a lack of centralization in test methodologies and verification metrics. This limitation restricts end-users from successful adoption of technologies developed in HRI research due to an inability to compare individual HRI solutions among each other. This paper intends to overcome this limitation by conducting a survey of the current state-of-practice in HRI research to identify specific opportunities to create novel solutions in standardizing HRI test methods and metrics.

Specifically, this paper provides a survey of the proceedings from ACM/IEEE International Conference on Human-Robot Interaction (HRI) and the IEEE International Conference on Robot and Human Interactive Communication (Ro-Man). We specifically call to attention the metrics, test methodologies, and technology trends associated with the state-of-practice in HRI research. Through this survey, we highlight the current state of metrology within the field of HRI.

Initially, the overall structure of the proceedings, including basic statistics of the studies are presented. Then, the use of metrics, tools, and test mechanisms for assessing HRI performance are discussed. Additional observations are also discussed followed by proposed solutions and conclusions.

A. The Conferences

The HRI and Ro-Man conferences are preeminent venues for emerging and ongoing research in the field of HRI. Providing a collegial platform to introduce and discuss studies focusing on human-robot, human-machine, and human-computer interaction, these conferences present a unique glimpse of the state-of-practice in HRI research. Both the HRI Conference and Ro-Man Conference are highly selective of the papers accepted for presentation, with typical acceptance rates of 15-25% and 35-50%, respectively, for main track proceedings. Therefore, the conference proceedings are reflective of the community's values and present a useful snapshot with regards to the test methods, metrics, and technologies being leveraged for the current state of HRI research.

This work analyzes a total of 1464 papers through examination of seven years of HRI Conference proceedings (2015 through 2021) and six years of Ro-Man Conference proceedings (2015 through 2020) to present a holistic snapshot of the state of methods and metrics in HRI research. Workshop proposals, late-breaking results, keynote talk abstracts, and demo presentations were omitted from this study due to varying methodologies and shorter study timeframes. The list of proceedings analyzed can be found in the supplemental materials, which can be accessed via the following link: **https://zenodo.org/record/5789410#.Ydc98WjMKUk** [1].

The subsequent work conducts aggregate analysis of all the conferences, as opposed to analyzing time series trends among the individual conferences, due to limited space. Future work will examine the data with more granularity.

II. General Statistics

A. Researcher Demographics

Of the 1464 papers, 74.80% of authors were from academia, 6.97% were from research institutions, and 2.25% were from

industry. The remaining 15.98% were collaborations between academia and either research institutions or industry. These statistics show that the state of HRI publications is still dominated by fundamental academic research. In addition, the research described in the proceedings were done in 51 countries, with 27.56% of the papers' first authors from the United States, 17.10% from Japan, 7.11% from Germany, and 5.88% from Italy. The total percentage breakdown is shown in Fig. 1. We observe that the majority of studies are concentrated in few countries, and thus results cannot easily be applied to other populations. Implicit cultural differences are difficult to pinpoint during development, so we believe that any social HRI study should be replicated across multiple countries for additional validity.

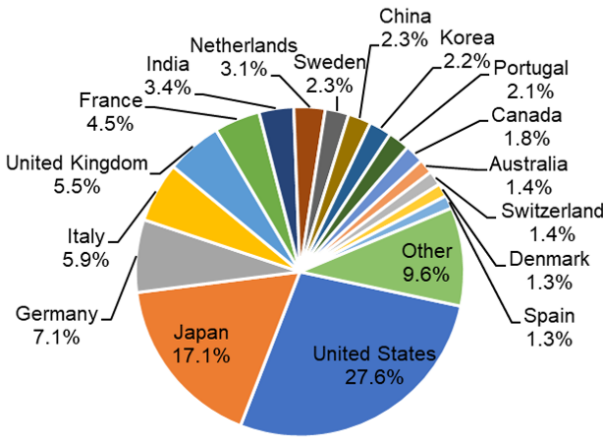

Fig. 1: Breakdown of the geographic location of first authors' institutions.

B. Human Subjects Demographics

We distinguish the study sample sizes between two subcategories, in-person sample size and online sample size. We categorize any data collected from physically-present humans as being "in-person," and data collected from people remotely as "online." There were a total of 1078 user studies with 936 in-person studies and 142 online studies, with some studies having both online and in-person sampling. In-person study sample sizes can be seen in Fig. 2. The spectrum of human participants in the online studies was generally higher than in-person by a factor of ten, with an average sample size of about 441.

Although we see some outliers, the majority of papers used under 50 participants. While a larger sample size would be ideal to verify claims in these papers, sometimes this approach is not feasible. More importantly, a significant majority of the proceedings do not justify the sample used in their study, whether they be too large or too small. Hence, standardization of sample sizes by requiring statistical power analysis [2] would significantly aid other researchers in comparative studies.

Demographics of the samples were not consistently reported, but a trend towards using populations of university students was seen. In addition, online studies using Mechanical Turk tend to not specify important demographics such as country of origin [3]. Although it is understandable that this would occur for pilot studies, making generalized conclusions on the bases of such restricted studies may not be valid.

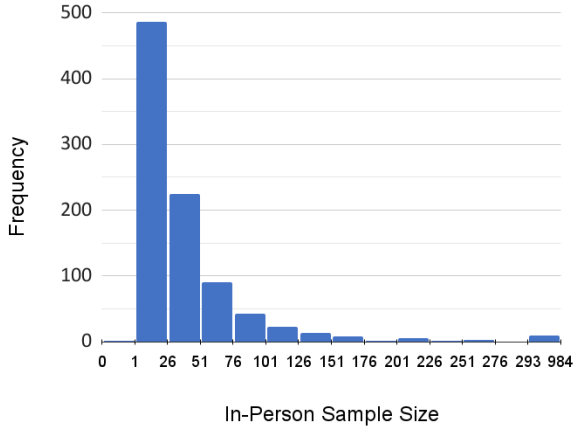

Fig. 2: Histogram of sample sizes for in-person user studies.

C. Robotic Systems

As part of our survey, we sought to determine the composition of the robots used in the various studies. Worth mentioning is that these robots, when present with human subjects, were operated along a broad spectrum of autonomy; some robots were allowed to operate fully autonomously, while others were controlled leveraging Wizard of Oz (WoZ, [4]). The types and frequencies of these robots are given in Fig. 3. Of the 1464 papers, 230 (15.7%) did not feature any representation of robots at all. Many of these papers focused on novel sensor systems, ontologies, and decision-making algorithms. Of the studies that did involve representative robots, humanoids represented the prominent robot classification with more than 388 (26.5%) instances. Considering avatars (platforms that emote responses based on simulated facial expressions, gaze, or other visual cues), non-human animals, and simulated robotic agents often embody anthropomorphic trains, the actual number of "human-like" robots is considerably higher. Mobile platforms constituted the second largest classification of robotic platforms (11.9%).

In total, 666 studies (45.5%) used some form of commercial robotic platform rather than a custom-built system. This is, generally speaking, a positive sign, as the use of robots that are (or were at one time) commercially available is conducive to supporting repeatable and reproducible HRI research. Specifically, the reliance on one-off robot platforms makes recreating the original testing conditions–and thus replicating and verifying the research results–virtually impossible. Although the prices and availability of these commercial robot platforms vary significantly, the use of widely-accessible and available robotic platforms are a benefit to the larger HRI community, as

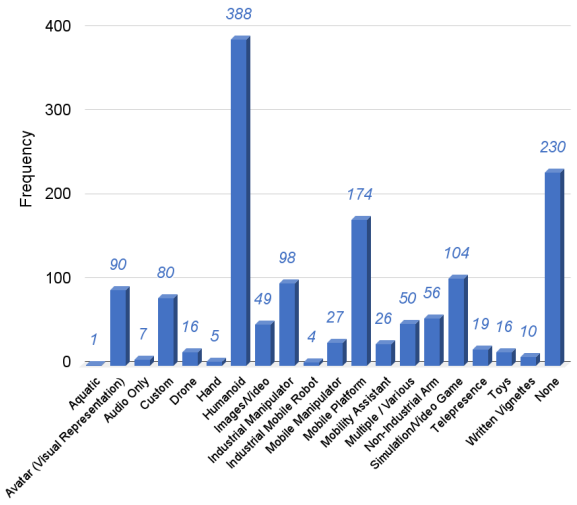

Fig. 3: Histogram of robot types used in studies.

they provide a solid basis for comparative studies in addition to being more representative of real-world environments.

III. Metrics and Test Methods

A major focus of this study is the identification and classification of the metrics and test methods used to assess system performance, extract user preferences, and verify models of behavior.

A. Objective and Subjective Measures

A point of interest of this study was with regards to the quantity and types of subjective and objective measures used in the various studies. Subjective measures are those that are traditionally self-reported by human participants, and center around the participants' opinions toward or about the robots they are evaluating. While these measures may be influenced by external factors, they most directly capture the individuals' thoughts and preferences. In contrast, objective measures tend to be independent of the user's feeling, and focus on quantifiable values such as interface utility, timing, and task performance. The top five most frequently used subjective measures (Table Ia) included evaluations capturing the human participants' directed emotion toward robot (13.30%), usability (12.25%), trust (12.13%), interface attribution (11.27%), and personal preference (10.65%). The objective measures (Table Ib) were principally focused on task performance (28.51%) and the utilization of interfaces.

Although subjective measures tend to capture the answers to many HRI research questions, the qualitative nature means that they are not always infallible. Subjective measures were generally captured through the use of customized pre- or post-task questionnaires. While it is necessary and important to capture feedback from the participants in a study, writing survey questions specifically for each study can result in leading questions if they are not subject to review. Use of objective, quantitative measures in addition can help ground the study and provide a basis for noticing outliers. Additionally, the full

TABLE I: Most frequently used Subjective and Objective Assessments

Subjective Measures	# Papers
Directed Emotion	216
Usability	199
Trust	197
Interface Attribution	183
Personal Preference	173

(a) Most common subjective measures

Objective Measures	# Papers
Task Performance	463
Task Timing	167
Raw Sensor Readings	145
Interface Performance	141
Accuracy	136

(b) Most common objective measures

text of study surveys was rarely reported, leading to difficulties with replication or re-use, let alone validation of the original study procedures. Several of the most-used validated surveys will be discussed in the next section.

B. Use of Surveys and Questionnaires

The majority of papers involving user studies included a subjective survey of the participants as a primary method for gathering metrics. There are a number of surveys commonly used in the community that we refer to as Named Surveys. These surveys are categorized based on whether or not they were referenced by a formal name (e.g., the System Usability Scale [5]) from a paper previously published on the topic. In all, there were 177 Named Surveys cited, with a total of 291 citations across 1078 papers with human studies. Of these Named Surveys, only 74 were cited more than once, meaning 66.9% of all cited surveys were used by only a single paper. This does not speak to the quality of these Named Surveys, but rather that there exists a myriad of such surveys to choose from. Indeed, some papers indicated that these surveys were selected because of specific properties or benefits over other surveys with similar content. However, the creation of additional surveys for a singular use results in an overwhelming number of options. Hence, there may be more benefit to refining or proposing additions to previously-created surveys.

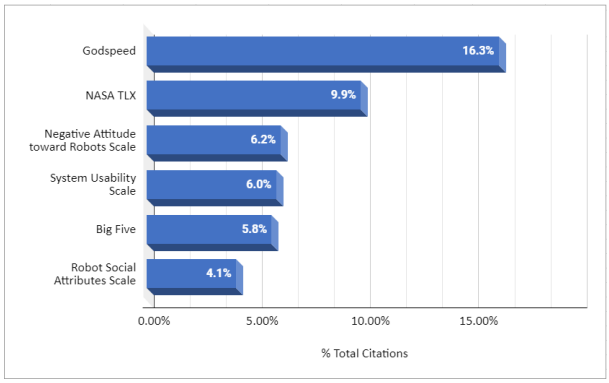

Fig. 4: Most frequently cited Named Surveys.

The most popular Named Surveys used were the Godspeed Questionnaire [6], NASA Task Load Index (NASA TLX) [7], the System Usability Scale [5], and Robot Social Attributes

Scale (RoSAS) [8]. While the NASA TLX was largely used unmodified, papers that used Godspeed and RoSAS tended to choose sub-scales that were most relevant for their needs, or mentioned modifying these scales without further details. Hence, these publications point to the benefit that some scales are flexible and can be adjusted to meet specific needs. In addition, we note that RoSAS was first published in the HRI 2017 proceedings, yet still made the list of most frequent citations with 20 papers over four years. This highlights the need for validated surveys in HRI research.

We define any other survey used in a study (i.e., surveys and questionnaires designed specifically for a given study and not previously validated) as a Custom Survey. Some studies use a combination of Named Surveys and a Custom Surveys, and may include Named Surveys that have been modified from their original versions specifically to fit a given study, technology, or demographic. The breakdown among studies that used surveys can be seen in Table II.

TABLE II: Usage of Custom vs. Named Surveys

Type of Survey	# Papers
Custom Questionnaire(s) Only	600
Named Questionnaire(s) Only	74
Combination of both	215

Many of the metrics in the Custom Surveys include concepts common across studies (e.g., trust and perceived capability), but are targeted toward their respective topics of research. This confounds reproducibility, and ensures the metrics are not broadly applicable. Moreover, because the criteria that these Custom Surveys use are tailored to their specific studies, research that use these surveys will be difficult–if not impossible–to objectively compare to other studies. Similarly, it should be noted that many of the subjective measures captured in the Custom Surveys are also present in the Named Surveys. Therefore, we recommend conducting Named Surveys as a primary option to facilitate standardization and comparison of results in HRI research. In addition, we found that there were situations where previously-published, validated surveys existed that encompassed the subjective topics used in other papers' custom questionnaires. For the purpose of encouraging replication, we recommend consulting the literature before creating an entirely new measure for a single study.

IV. Conclusions and Recommendations

In this paper, we presented a survey of the proceedings of seven years of the HRI Conference, and six years of the Ro-Man Conference. We provided an analysis of the test methods, metrics, and technologies used in the studies presented at the conference. A common theme throughout this paper is the concept of repeatable and reproducible research, specifically as it pertains to verifying and validating research results. We note a prevalent bias toward the use of non-commercial robotic platforms and surveys/questionnaires in the different studies. Similarly, the majority of studies leveraged subjective, qualitative measures, while less than half of the 1464 studies used some form of objective, quantitative metrics. The use of custom metrics and robot platforms with limited availability makes it difficult to directly compare different HRI studies, draw cross-cultural conclusions, and support repeatable and reproducible research within the HRI field.

For HRI technologies and fundamental concepts to mature and evolve,the experiments must be conducted in representative environments as opposed to strictly laboratory conditions. The use of standardized metrics can accommodate this transition of HRI research to these real-world environments. Hence, we propose the following solutions based on the specific knowledge uncovered in this work:

- **Create a centralized repository for study design and surveys.** Due to the interdisciplinary nature of HRI, it is understandable that published records of evaluation metrics and methods are decentralized and therefore it is difficult for researchers to understand existing methodologies for leveraging. Therefore, we propose creation of a centralized repository for researchers, and the survey study presented in this paper is an initial step in that direction.
- **Emphasize consistency and study design in HRI research curriculum.** As HRI courses are being increasingly included in academic curriculum, we recommend teaching human-subject study design with the introduction of methods specific to HRI. This research shows certain best-practices, including knowledge of available commercial robots and named surveys, which will aid early-career researchers with the fundamental knowledge to create consistent, replicable results.
- **Create and enforce official standards for implementation of HRI systems.** This paper shows that identification of the available test methods and metrics in HRI is possible. Hence, formal standards in the field are a possible solution towards enforcing replicability in HRI development. The initiation of new IEEE Standards Association Working Groups on HRI is a step in this direction, but will only be successful with significant participation from the research community.

The necessity for verifying study results and validating test methodologies cannot be understated. To justify the integration of developed technologies in the field, they must be evaluated with consistent and comparable metrics. Indeed, for any work to be truly impactful, its performance must be able to be validated lest it risk never seeing real-world implementation. While significant leaps are being made in terms of repeatability with regards to robot platforms and tools, there is much more progress to be made in regards to consistent metrics and test methods to inspire confidence for adoption in the field of human-robot interaction.

V. Acknowledgements

Special thanks to Xiang Li, Katherine Haas, and Victoria Koretsky for their assistance with data entry, without which this project would not have been possible.

REFERENCES

[1] M. Zimmerman, S. Bagchi, J. Marvel, and V. Nguyen, "Metrics and Methods used in IEEE/ACM HRI and IEEE RoMan Conference Proceedings from 2015 to 2021 Data," Jan. 2022, readMe provides usage information. [Online]. Available: https://doi.org/10.5281/zenodo.5789410

[2] J. Cohen, "Statistical power analysis," *Current Directions in Psychological Science*, vol. 1, no. 3, pp. 98–101, 1992.

[3] E. Minton, E. Gurel-Atay, L. Kahle, and K. Ring, "Comparing data collection alternatives: Amazon mturk, college students, and secondary data analysis," in *AMA Winter Educators' Conference Proceedings*, vol. 24, 2013, pp. 36–37.

[4] L. D. Riek, "Wizard of oz studies in hri: a systematic review and new reporting guidelines," *Journal of Human-Robot Interaction*, vol. 1, no. 1, pp. 119–136, 2012.

[5] J. Brooke *et al.*, "Sus - a "quick and dirty" usability scale," *Usability Evaluation in Industry*, vol. 189, no. 194, pp. 4–7, 1996.

[6] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009.

[7] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," *Adv. Psychology*, vol. 52, pp. 139–183, 1988.

[8] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The robotic social attributes scale (rosas): Development and validation," *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 254–262, 2017.