A Review and Recommendations on Reporting Recruitment and Compensation Information in HRI Research Papers

Julia R Cordero*, Thomas R Groechel*, and Maja J Matarić

***Abstract*— Study reproducibility and generalizability of results to broadly inclusive populations is crucial in any research. Previous meta-analyses in HRI have focused on the consistency of reported information from papers in various categories. However, members of the HRI community have noted that much of the information needed for reproducible and generalizable studies is not found in published papers. We address this issue by surveying the reported study metadata over the main proceedings of the 2021 IEEE International Conference on Robot & Human Interactive Communication (RO-MAN) and the past three years (2019 through 2021) of the main proceedings of the International Conference on Human-Robot Interaction (HRI) and alt.HRI. Based on the analysis results, we propose a set of recommendations for the HRI community that follow the longer-standing reporting guidelines from human-computer interaction (HCI), psychology, and other fields most related to HRI. Finally, we examine two key areas for user study reproducibility: recruitment details and participant compensation. We find a lack of reporting of both of these study metadata categories: of the 414 studies across both conferences and all years, 258 studies failed to report recruitment method and 255 studies failed to report compensation. This work provides guidance about specific types of needed reporting improvements for the field of HRI.**

I. Introduction

Reproducibility–the ability to duplicate a prior study's procedure and obtain the same results [1]– is critical for human-subject studies that aim to be *generalizable*, i.e., applicable to different populations and settings [2]. Human-robot interaction (HRI), similar to its related fields of human-computer interaction (HCI) and psychology, conducts studies in critical areas such as health and education, underscoring the importance of reproducibility and generalizability of its results. Reproducibility has been a prevalent topic for HRI; the International Conference on Human-Robot Interaction introduced the "Reproducibility in Human-Robot Interaction" track in 2020, focused on reproducing prior HRI work [3]. HRI researchers have conducted studies to replicate findings from previous work [4], [5]. Despite these efforts, a lack of reproducibility for studies has been called a "crisis" in psychology and more broadly [6], [7].

HRI falls under this call to action to bridge the gap in research reproducibility. However, in order for a study to be reproducible and generalizable, it must report metadata (e.g., participant demographics) comprehensively and in enough detail for other researchers to evaluate it in context and aim to reproduce the reported results. Hence, HRI must aim to meet a higher standard of rigor in study reporting.

This work is builds on our previous paper, "What and How Are We Reporting in HRI? A Review and Recommendations for Reporting Recruitment, Compensation, and Gender" [8], presented at the HRI 2022 workshop, "Fairness and Transparency in Human-Robot Interaction" [9]. As in that work, we took inspiration from the method of review presented in "Four Years in Review: Statistical Practices of Likert Scales in Human-Robot Interaction Studies" [10]. Accordingly, we first conducted a review of the reporting standards in related fields (HCI, psychology). Then, to evaluate the current level of study reporting in HRI, we collected data on and examined metadata being reported and not reported. We chose to focus on the metadata categories of recruitment and compensation because both describe the researchers' actions that determine the final studied population. Recruitment method, recruitment setting, participant inclusion criteria, data analysis exclusion criteria, and ethical considerations most directly shape this pool; compensation also influences the number and profile of participants who may otherwise decide for or against participation in a study [11].

We found that two metadata categories, participant recruitment and participant compensation, were reported infrequently or with an inadequate amount of contextual information. In HRI, having a contextualized understanding of what researchers did to shape the participant pool is critical for being able to reproduce a study and then generalize results to a broader, inclusive population.

This work contributes a review of the existing literature for reporting guidelines and standards in psychology, HCI, and general qualitative research. While some of the guidelines apply to specific types of studies (e.g., randomized control trials in medicine), all of the surveyed literature agreed on minimum details to be reported. We provide recommendations for user studies in the HRI community.

Second, we contribute a review of papers published in the main proceedings of the 2021 IEEE International Conference on Robot & Human Interactive Communication (RO-MAN), and those published in the International Conference on Human-Robot Interaction (HRI) from 2019 to 2021 (main proceedings and alt.HRI), both with regard to reported metadata. We reviewed 364 papers, 288 of which contained at least one study, totalling 416 studies. We found that 259 (≈62.3%) studies did not report details about participant recruitment and 257 (≈61.8%) studies did not report any de-

*Equal Contribution

This work was supported by the NSF NRI 2.0 grant for "Communicate, Share, Adapt: A Mixed Reality Framework for Facilitating Robot Integration and Customization", NSF IIS-1925083.

Julia Cordero, Thomas Groechel, and Maja J Matarić are all with the Interaction Lab, Department of Computer Science, University of Southern California, Los Angeles, CA 90089, USA
{jrcordero, groechel, mataric}@usc.edu

tails about participant compensation. Based on these findings, we argue that there is a need for reporting recommendations for the HRI community.

Nota Bene: Similar to [10], we have not employed best practices in our own prior work. Our goal is not to disparage the field, but to highlight the lack of reporting of study metadata. We hope to improve the rigor of HRI work so that we can all contribute more confidently to the reproducibility and generalizability of HRI studies.

II. Literature Review & Best Practices

HRI studies by definition involve human behavior. Each study participant's behavior is necessarily influenced by the participant's identity and experiences with the researcher and the experiment methodology and context. Hence, explicitly reporting what participants experience allows for a more complete interpretation of results and study reproducibility [12]. Reporting guidelines have been proposed and adopted in other fields that make use of qualitative user studies. When drawing inspiration from those fields, it is important to note that HRI is a distinct field from psychology and medicine because it requires interaction with technology, and it is distinct from HCI because of the embodied nature of robots. As a result, recommendations from other fields may be relevant but not directly applicable to HRI, and so may be modified to fit an individual researcher's specific study. Since HRI shares a dependency on human subjects, it can benefit from the examination of minimum information reporting already established in longer-standing related fields.

A. Recruitment

Recruitment Method Previous work has highlighted the importance of the way that researchers reach potential participants. The textbook *Qualitative HCI Research* states that the recruitment strategy may affect "quality, reliability, or generalizability" of a study and thus should be considered important to report. This acknowledges the role recruitment methods play in a study's methodology and, by extension, reproducibility [13].

Recruitment reporting is covered in various reporting guidelines; for example, Standards for Reporting Qualitative Research (SRQR) and Consolidated Criteria for Reporting Qualitative Research (COREQ), both written for qualitative health research [14], state that the manner of selecting and approaching participants needs to be explicitly reported [15], [16]. Describing the recruitment method is also part of the reporting checklist for psychology studies by the American Psychological Association (APA) [17] and Consolidated Standards of Reporting Trials (CONSORT) [14], [18].

Reporting the recruitment method has larger implications beyond simply detailing the flow of participants through the study; previous work suggests that the recruitment method should also include details about compensation [19].

Finally, Blandford et al. [13] suggest reporting whether compromises needed to be made during recruitment, and "the likely impact of this on the quality, reliability or generalisability" [13]. Two examples of papers that did so are "Multi-Modal Proactive Approaching of Humans for Human-Robot Cooperative Tasks" [20] and "Design of an Assistive Robot for Infant Mobility Interventions" [21]. For these works, reported compromises were due to the international SARS-CoV-2 pandemic and unavailability of participants with the targeted disability, respectively.

Recruitment Setting Not all guidelines provide recommendations about reporting recruitment settings (e.g., [15], [16]), CONSORT, the guidelines written for randomized control trials (RCTs) suggests a standard for reporting recruitment settings [18]. The American Psychological Association (APA) guidelines also include reporting the recruiting context (e.g., location, time period) in qualitative research [17]. We believe that reporting the setting in which participants are recruited provides important context about data sources for future researchers aiming to reproduce a study.

Participant Inclusion Criteria and Data Analysis Exclusion Criteria Inclusion criteria are defined as the broad definition of eligibility for a study, while exclusion criteria are defined as any criteria that remove any potentially eligible participants from the pool; both are intended to be applied before the study takes place [22]. Although this distinction is drawn clearly and recommended to be reported as two separate items by some works ([15], [16], [22], [23]), CONSORT states only that "eligibility criteria" should be reported [18]. When reporting decisions made for inclusion and exclusion criteria, SRQR recommends stating the justification for the decision in addition to stating the criteria themselves [15].

Ethical Approval SRQR recommends reporting ethical approval [15], while COREQ and CONSORT do not provide recommendations on that point [16], [18]. More broadly, Qualitative HCI recommends reporting ethical considerations and how they did or did not affect decisions in recruitment and study design [13]. This perspective also agrees with SRQR's recommendation to provide justification for study decisions, as well as to report the decision made [15].

Informed Consent In the reviewed literature, there is strong support for reporting the informed consent of participants. In "Reporting Qualitative Research in Psychology", reporting the recruitment process includes reporting "informed consent" [11]. Critical Appraisal Skills Program Qualitative Reporting Checklist also states that a research paper can be evaluated by the discussion of ethical issues, including participant consent [24]. Similarly, SRQR recommends reporting participant informed consent as part of the "ethical considerations relating to human subjects" [15]. However, neither COREQ nor Qualitative Research HCI state whether to report participant consent [16] [13].

Recommendation: All of the guidelines we reviewed stated which details should be included in a study report; SRQR made the additional recommendation that explanations accompany the reported study metadata so as to provide visibility into the researcher's role and the study's aims. While this was not a consensus among guidelines included in our literature review, we believe that the additional infor-

mation further enhances study reproducibility.

Based on the relevant literature review, we recommend:

- Reporting ethical approval from the relevant regional study approval board, and the corresponding consent process for participants;
- Explaining ethical considerations of the study design;
- Reporting the recruitment method and setting and explaining why the chosen recruitment strategy fits the aims of the study. (For an HRI conference paper example of reporting the strategy with an explanation, see "The Effects of a Robot's Performance on Human Teachers for Learning from Demonstration Tasks" [25].) If there is any reason for compromise during the recruitment phase, we further suggest reporting the reason and its possible impact on the study.
- Reporting both the inclusion criteria (study population) and exclusion criteria (causes for a member of the population to be disqualified from the study) and explaining the justification for choosing them for the study.

B. Compensation

While there is little specific guidance from various reporting guidelines on whether and how to report study participant compensation, there are broader recommendations to provide explanations of ethical considerations in study design [13]. In the healthcare research community, participant compensation is considered an ethical issue [26]–[28]. Additionally, previous work in HCI has called for more standardized reporting of participant compensation in order to increase study replicability [19].

Form and Amount of Compensation Some reporting standards do not provide recommendations on reporting participant payment (e.g., [15], [16], [18]). However, ethics boards (e.g., Institutional Review Boards) typically require that researchers report participant incentives, which casts compensation as an ethical factor to be reported [29]. The APA recommends that researchers "describe any incentives or compensation" in their study reports [17]; in HCI, a proposed standard is to report both the form and amount of participant compensation [15]. In the context of clinical trials for HIV research, there has been a call to record compensation for studies for the benefit of other researchers to maintain consistency when replicating the same study [26].

Location and Duration The proposed HCI standard [19] is to report location and duration along with compensation form and amount so as to contextualize the value of the incentives. The discussion of tracking participant compensation also references the role of location in contextualizing the amount of payment [26]. No explicit recommendations are provided in CONSORT on reporting study location or duration [18]; COREQ, however, suggests a standard of reporting the duration of time participants are studied [16]. SRQR guidelines recommend reporting context (defined as "setting/site and salient contextual factors") and the rationale for the context [15]; this recommendation agrees with APA's emphasis on the role of context in human subjects studies and the need to report the study context [17].

Indicators for Socioeconomic Status: Occupation and Education Although the reporting checklists in this literature review do not explicitly recommend reporting socioeconomic status (SES) in a participant pool [13] [15] [16], other studies in psychology and the health sciences have discussed SES as an influential characteristic in a sample. For example, Pater et al. [19] coded each manuscript for "economically disadvantaged" participants as part of the information that contextualizes the reporting of compensation schemes. Related is the recognition that a participant's economic situation can place them in a position vulnerable to undue influence [26] [30]. While SES can be difficult to distill, occupation has been used as a strong indicator of SES and education as helpful supplementary information [31] [32] [33].

Recommendation: Reporting compensation goes beyond reporting whether participants were paid; it is important to include amounts in order to understand the full impact of the incentive. Reporting the time that participants spent in the study and where they were located is just as important as part of the context. Although there is a lack of consensus of the importance of reporting participant compensation, we believe that the standards proposed in HCI can benefit HRI most due to the similarity in study structure. Based on the relevant literature review, we recommend:

- Reporting the form (e.g., voucher, course credit, gift card) and value of the compensation;
- Reporting the location of the study to further contextualize the value of compensation;
- Reporting the full study duration including data collection phases (e.g., "participants performed tasks with the robot for 20 minutes") and total study time (e.g., "the participant was at the research lab for 1.5 hours");
- Reporting participants' highest level(s) of education attained and the participants' occupations if applicable, in order to contextualize the form and value of compensation provided.

We acknowledge that following these recommendations can occupy a significant amount of space in space-limited papers. Researchers can opt to include a more thorough report in supplemental materials; for an HRI conference paper example see "Assessing and Addressing Ethical Risk from Anthropomorphism and Deception in Socially Assistive Robots" [34] for how to include such materials or see "What Should Robots Feel Like?" [35] for an example of how to report such information in a compact table.

Given the above recommendations, we analyzed papers from the International Conference on Human-Robot Interaction for the past three years (2019, 2020, 2021) and from the International Symposium on Robot and Human Interactive Communication in the last year (2021), coding for study metadata in an attempt to quantify what is, and what is not, being reported. This analysis aims to highlight the areas of specific improvements in HRI data reporting.

III. Review of RO-MAN and HRI Conferences

We reviewed and coded the papers from the 2021 IEEE International Conference on Robot & Human Interactive

Communication (RO-MAN). We also reviewed and coded the papers from the 2019, 2020, and 2021 International Conference on HRI main proceedings and alt.HRI (a paper track aimed at "pushing the boundaries of HRI", published in the companion proceedings for 2020 [36] and 2021 [37]). In total, we reviewed 364 papers, 288 of which contained at least one study, totalling 414 studies.

We examined one year of RO-MAN (2021) compared to three years of HRI (2019-2021) proceedings due to the larger volume of studies in RO-MAN. Oour previous work analyzed 174 papers with 236 studies across all three HRI conferences; here we add 190 papers with 178 studies from just RO-MAN'21, totalling 364 papers and 414 studies overall. (Note that a single paper may include 0, 1, or multiple studies.)

A. Coding Methodology and Criteria

We collected data by reading through each paper. For each paper, we performed a combined manual and automated key term search for different columns, as described below (e.g., "rewarded") and marked the columns based on the reported data found in the paper. Next, we read the paper fully to code any categories missed by the first search pass, and to check that all data recorded by the first search were correct.

We determined how many, if any, studies the paper contained. A study qualified if the paper reported participants, a study procedure was described, data were collected from those participants, and the data were used or reported. Pilot studies were included if they fit these criteria. We performed an analysis for three categories of study participant metadata: recruitment, compensation, and gender.

The recruitment method was coded and categorized if papers explicitly stated **how participants were recruited** (e.g., flyers, social media, an outside company). If a paper stated "we recruited university students" but omitted how they were recruited, it was coded as not having reported the recruitment method. For online studies conducted via an external online platform (e.g., Amazon Mechanical Turk, Prolific), the recruitment method was implied and coded as reported. Online studies run by a university had to state how their participants were recruited (e.g., emailed survey to engineering email list) to be coded as reported.

We also coded the population studied as convenience and non-convenience sampling. We define convenience sampling as any population strictly from a college or university, or recruited by an external online platform (e.g., Amazon Mechical Turk, Prolific). Convenience sampling was coded for anything explicitly stated as **"convenience sampling", college/university students (e.g., "university students"), or implied by the recruitment tool (e.g., "Amazon Mechanical Turk")**. Non-convenience sampling was coded when it was either explicitly stated or implied (e.g., "we worked with grade-school children, hired clinical experts"). Studies were coded as a "mixture" if they contained both university and non-university students (e.g., reported as recruiting from the university and surrounding area). Not reported was coded for studies that did not fall into any of the above categories and thus no supported assumptions could be made.

Acknowledgement of ethics board approval was also coded for when a paper explicitly stated **any acknowledgement of ethics board approval**. While encouraged, specific ethics board approval identifiers (e.g., "all study materials were reviewed and approved by a University ethics board under application UP-123456") were not required; a statement regarding the approval of the study was considered sufficient. Informed consent was coded as reported when authors included any statement regarding participants giving "consent" or "permission".

Compensation was coded and categorized when papers explicitly stated **any direct benefit to the participant or a statement of no direct benefit**. Automated search terms included "reward", "award", "receive", "given", "compensated", and "paid". The study procedure sections were also carefully read to check for other wordings of compensation or statements about the participants not being paid. If none of those were found, compensation was coded as not reported.

Participant inclusion criteria were coded and categorized when there was any statement about what characteristics researchers looked to have in the participant pool. In some papers, this was reported as a list of requirements (such as "18 years of age and fluent in English"), and in others this was reported as a general goal (such as "We aimed to recruit non-experts in programming"). Similarly, data analysis exclusion criteria were marked as reported when there was any statement present about why, if at all, researchers needed to discard a participant's data. In some cases, it was explicitly stated that no participant data were discarded, which we also coded as reported.

Online studies were also coded and categorized when papers explicitly stated they used an online platform (e.g., Amazon Mechanical Turk, "online study"). We analyzed two sets of data. The first set contained all studies. The second set examined the data when online studies were removed. There were a total of 416 studies of which 124 were online studies ($\approx$29.8%) and 294 were non-online ($\approx$70.7%). The four conferences had the following numbers of online and non-online studies: HRI'19 (11,50); HRI'20 (23,80); and HRI'21 (39,34) while RO-MAN'21(51,127).

B. Inter-rater Reliability

One human coder annotated the HRI dataset. Two human coders annotated the RO-MAN dataset; their Cohen's kappa is shown in Table I. While all Cohen's kappa values reach the threshold of satisfaction (> 0.70), it is worth noting that it was occasionally difficult to determine whether metadata were reported or not. For example, a paper may report, "The participants were volunteers", which may imply they were not paid. This highlights the importance of explicit reporting in order to better support reproducibility efforts in the field.

C. Recruitment and Compensation Frequency

A total of 157 studies reported recruitment method. When online studies were removed, 62 studies reported recruitment method. For population studied, 217 were coded as convenience, 27 were coded as a mixture, 86 were coded

MetadataCategory	Cohen's kappa
RecruitmemtMethod	0.95
InformedConsent	1.0
InclusionCriteria	0.72
ExcludedData	0.75
EthicsBoard	1.0
SampleType	0.70
Compensation	0.86

TABLE I: Cohen's kappa values for categories where we coded for presence of reporting. This was calculated between two raters on ≈ 15.3% of the RO-MAN'21 study data.

as non-convenience, and 86 did not report their population. When online studies were removed, 113 were coded as convenience, 26 were coded as a mixture, 75 were coded as non-convenience, and 78 did not report their population. A total of 189 studies reported ethics board approval. When online studies were removed, 156 studies reported ethics board approval. Finally, 213 studies reported informed consent with 203 reported when online studies were removed. Recruitment reporting summaries can be found in Figs. 1, 2, 3, and 4.

A total of 313 studies reported participant inclusion criteria, of those 230 were not online studies. For data excluded from analysis, 284 studies reported. A total of 212 studies reported excluded data when online studies were removed. The results are found in Figs. 5 and 6.

A total of 118 studies reported compensation. When online studies were removed, 73 studies reported compensation. Compensation reporting summaries can be found in Fig. 7.

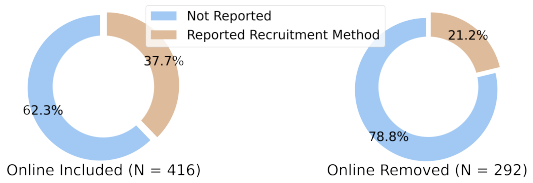

Fig. 1: Papers that explicitly stated how participants were recruited and papers that did not.

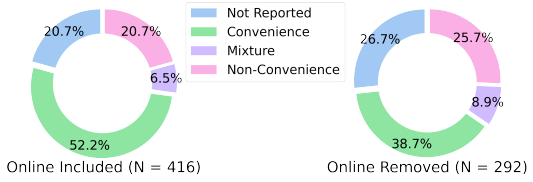

Fig. 2: Papers that reported the participant population (convenience, mixture, or non-convenience), either explicitly or implied via the recruitment method and those that did not.

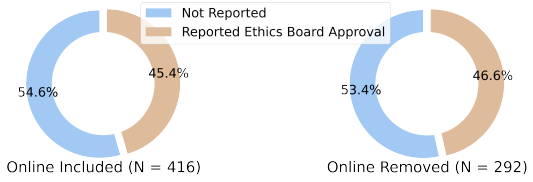

Fig. 3: Papers that explicitly stated an acknowledgement of ethics board approval and papers that did not.

D. Between Conference Analysis

A summary of Chi-square tests for independence [38] among the conferences (HRI'19 + HRI'20 + HRI'21 against RO-MAN'21) for each metadata category are in Table II. The p value measures the likelihood that the observed association between each independent variable (i.e., conference year) and the dependant variable (i.e., metadata category) is caused by chance. The HRI conferences were added together and compared to RO-MAN to find differences (RO-MAN vs. HRI). For within HRI results, please see [8]. The results are found in Table II.

Overall, HRI papers report informed consent, criteria for participant inclusion and exclusion, and compensation more than RO-MAN papers do. This can be seen in Table III. These results are consistent with and without online studies included. While HRI has a higher reporting ratio, both conferences fall short of reporting in all categories.

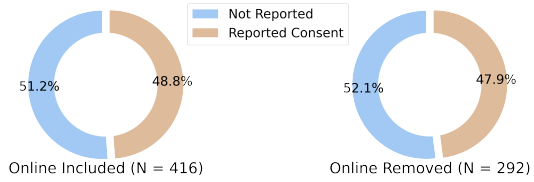

Fig. 4: Papers that stated that informed consent was given.

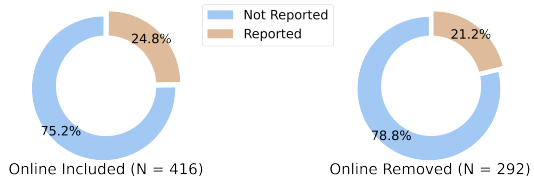

Fig. 5: Papers that stated participant inclusion criteria.

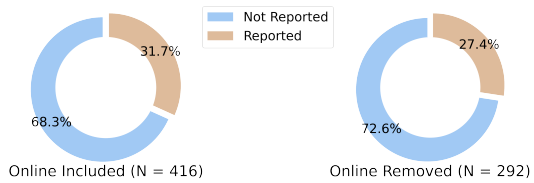

Fig. 6: Papers that explicitly stated if participant data were excluded from analysis.

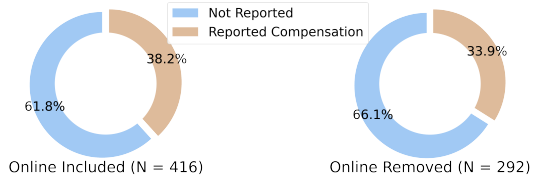

Fig. 7: Papers that stated direct benefit to the participants or no direct benefit and papers that did not.

MetadataCategory	N-OO	X^2	p	dof
RecruitmemtMethod		2.463581	.117	1
RecruitmemtMethod	Y	3.483210	.062	1
InformedConsent		24.416052	<.001***	1
InformedConsent	Y	16.885307	<.001***	1
InclusionCriteria		2.276732	.131	1
InclusionCriteria	Y	0.506571	.477	1
ExcludedData		16.331211	<.001***	1
ExcludedData	Y	16.387276	<.001***	1
EthicsBoard		1.607313	.205	1
EthicsBoard	Y	0.746356	.388	1
Compensation		29.278371	<.001***	1
Compensation	Y	17.048872	<.001***	1

TABLE II: Chi-square test for independence [38] results between each conference (HRI'19 + HRI'20 + HRI'21 against RO-MAN'21) for each category. N-OO indicates "non-online only" studies included within the test. Significance values used were $p < .05$ = *, $p < .01$ = **, $p < .001$ = ***.

MetadataCategory	Conference	N-OO	R	NR	Ratio
Informed Consent	HRI'(19,20,21)		62	116	1.87
Informed Consent	RO-MAN'21		142	96	0.68
Informed Consent	HRI'(19,20,21)	Y	43	84	1.95
Informed Consent	RO-MAN'21	Y	97	68	0.70
Excluded Data	HRI'(19,20,21)		37	141	3.81
Excluded Data	RO-MAN'21		95	143	1.51
Excluded Data	HRI'(19,20,21)	Y	19	108	5.68
Excluded Data	RO-MAN'21	Y	61	104	1.70
Compensation	HRI'(19,20,21)		41	137	3.34
Compensation	RO-MAN'21		118	120	1.02
Compensation	HRI'(19,20,21)	Y	26	101	3.88
Compensation	RO-MAN'21	Y	73	92	1.26

TABLE III: Differences in reporting (R) vs. not reporting (NR) ratios between HRI'19-'21 and RO-MAN'21. Only differences with significance for Chi-square test for independence [38] are reported. See Table II for details.

IV. Discussion and Limitations

We focus on recommendations for reporting HRI study metadata (Sec. II). The data are reported here to help support the idea that we, as HRI researchers, need to better report study metadata for more reproducible and generalizable studies. The goal is to encourage replicability studies such as "A Three-Site Reproduction of the Joint Simon Effect with the NAO Robot" [4] from HRI'20. Both conferences have recently taken steps towards supporting greater reproducibility and generalizability in HRI; RO-MAN 2021 held a special session called "Inclusive HRI", which encouraged submissions related to transparency, and HRI 2020 included a track for reproducibility. We hope to see similar focus in other conferences and journals.

Guideline for reporting metadata change. Reviewing studies from the past three years highlights the nuance needed for study categorizations and metadata. The changing landscape of user studies, such as the reproducibility crisis [7] or the international SARS-CoV-2 pandemic, indicate a need to reevaluate recommendations annually.

Finally, while we have attempted to make a case for reporting guidelines in HRI, it is also important to acknowledge the limitations of such a review. We only sourced papers from the 2021 International Conference on Robot & Human Interactive Communication (RO-MAN) and the International Conference on Human-Robot Interaction (HRI 2019 through 2021); differing trends from the ones we report on may be discovered in a review of different conferences or in journals. Additionally, as our review only covers the latest year of RO-MAN and the three most recent years of HRI, it is difficult to discern established trends in reporting HRI studies. Finally, we believe our coding is consistent and minimizes human error, but there is always possibility of human error in any manual search process, as shown in Table I.

V. Conclusion

Human participants are by definition a key component of human-robot interaction studies, and their behavior in studies will be at least in part impacted by the events they experience and their personal identities. Thus, to better support the reproducibility of studies and generalizability of a study's results, it is critical that who participants are and what they experience throughout the course of a study are reported. In this paper, we reviewed the reporting guidelines for participant recruitment and compensation in the related fields of psychology, medicine, and HCI. We then surveyed the papers published in RO-MAN 2021 and HRI 2019-2021 to examine reporting the same two categories and found that rates of reporting, both individually important and for understanding full study context, to be insufficient. We aim not to undermine the quality of any paper, but to suggest a higher rigor of reporting metadata in each study, including the authors' own, to allow for greater reproduction of results in HRI.

Acknowledgment

This work was supported by the National Science Foundation, IIS-1925083. The authors thank Karen Ly for help with annotation.

VI. Citation Diversity Statement

Recent work in several fields of science has identified a bias in citation practices such that papers from women and minority scholars are undercited [39]–[43]. We recognize this bias and have worked to reference appropriate papers with fair author inclusion. To raise awareness of this problem, we state the gender distribution of the references: 20% are published by a female/male team with a female lead, 10% by male/female with a male lead, 37.5% by a solely female team or writer, 30% by a solely male team or writer, and 2.5% by a male/non-binary person team with a male lead. While we used the available resources to determine author gender (e.g., personal websites and biographies), there is a possibility of misgendering, as we did not contact authors directly for confirmation. Additionally, due to lack of author-provided information regarding ethnicity, we have not included an ethnicity distribution for cited works.

References

[1] J. T. Cacioppo, R. M. Kaplan, J. A. Krosnick, J. L. Olds, and H. Dean, "Social, behavioral, and economic sciences perspectives on robust and reliable science," *Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences*, 2015.

[2] L. Carminati, "Generalizability in qualitative research: A tale of two traditions," *Qualitative health research*, vol. 28, no. 13, pp. 2094–2101, 2018.

[3] *HRI '20: Proc. of the 2020 ACM/IEEE Intl. Conf. on HRI*, (New York, NY, USA), Association for Computing Machinery, 2020.

[4] M. Strait, F. Lier, J. Bernotat, S. Wachsmuth, F. Eyssel, R. Goldstone, and S. Šabanović, "A three-site reproduction of the joint simon effect with the nao robot," in *Proc. of the 2020 ACM/IEEE Intl. Conf. on HRI*, pp. 103–111, 2020.

[5] D. Ullman, S. Aladia, and B. F. Malle, "Challenges and opportunities for replication science in hri: A case study in human-robot trust," in *Proc. of the 2021 ACM/IEEE Intl. Conf. on HRI*, pp. 110–118, 2021.

[6] M. Baker, "First results from psychology's largest reproducibility test," *Nature*, vol. 30, no. 10.1038, 2015.

[7] M. Baker, "Reproducibility crisis," *Nature*, vol. 533, no. 26, pp. 353–66, 2016.

[8] J. R. Cordero, T. R. Groechel, and M. J. Matarić, "What and how are we reporting in hri? a review and recommendations for reporting recruitment, compensation, and gender," *arXiv preprint arXiv:2201.09114*, 2022.

[9] H. Claure, M. L. Chang, S. Kim, D. Omeiza, M. Brandão, M. K. Lee, and M. Jung, "Fairness and transparency in hri," in *Proc. of the 2022 ACM/IEEE Intl. Conf. on HRI*, pp. 1244–1246, 2022.

[10] M. L. Schrum, M. Johnson, M. Ghuy, and M. C. Gombolay, "Four years in review: Statistical practices of likert scales in human-robot interaction studies," in *Companion of the 2020 ACM/IEEE Intl. Conf. on HRI*, pp. 43–52, 2020.

[11] H. M. Levitt, *Reporting qualitative research in psychology: How to meet APA style journal article reporting standards*. American Psychological Association, 2020.

[12] A. Birhane, "The impossibility of automating ambiguity," *Artificial Life*, vol. 27, no. 1, pp. 44–61, 2021.

[13] A. Blandford, D. Furniss, and S. Makri, "Qualitative hci research: Going behind the scenes," *Synthesis lectures on human-centered informatics*, vol. 9, no. 1, pp. 1–115, 2016.

[14] D. G. Altman, I. Simera, J. Hoey, D. Moher, and K. Schulz, "Equator: reporting guidelines for health research," *Open Medicine*, vol. 2, no. 2, p. e49, 2008.

[15] B. C. O'Brien, I. B. Harris, T. J. Beckman, D. A. Reed, and D. A. Cook, "Standards for reporting qualitative research: a synthesis of recommendations," *Academic Medicine*, vol. 89, no. 9, pp. 1245–1251, 2014.

[16] A. Tong, P. Sainsbury, and J. Craig, "Consolidated criteria for reporting qualitative research (coreq): a 32-item checklist for interviews and focus groups," *International journal for quality in health care*, vol. 19, no. 6, pp. 349–357, 2007.

[17] M. Appelbaum, H. Cooper, R. B. Kline, E. Mayo-Wilson, A. M. Nezu, and S. M. Rao, "Journal article reporting standards for quantitative research in psychology: The apa publications and communications board task force report.," *American Psychologist*, vol. 73, no. 1, p. 3, 2018.

[18] J. A. Bennett, "The consolidated standards of reporting trials (consort): Guidelines for reporting randomized trials," *Nursing Research*, vol. 54, no. 2, pp. 128–132, 2005.

[19] J. Pater, A. Coupe, R. Pfafman, C. Phelan, T. Toscos, and M. Jacobs, "Standardizing reporting of participant compensation in hci: A systematic literature review and recommendations for the field," in *Proc. of the 2021 CHI Conf. on Human Factors in Computing Systems*, pp. 1–16, 2021.

[20] L. Naik, O. Palinko, L. Bodenhagen, and N. Krüger, "Multi-modal proactive approaching of humans for human-robot cooperative tasks," in *2021 30th IEEE Intl. Conf. on Robot & Human Interactive Communication (RO-MAN)*, pp. 323–329, IEEE, 2021.

[21] A. Vinoo, L. Case, G. R. Zott, J. R. Vora, A. Helmi, S. W. Logan, and N. T. Fitter, "Design of an assistive robot for infant mobility interventions," in *2021 30th IEEE Intl. Conf. on Robot & Human Interactive Communication (RO-MAN)*, pp. 604–611, IEEE, 2021.

[22] J. R. Wright, "The importance of reporting patient recruitment details in phase iii trials," *J Clin Oncol.*, vol. 24, pp. 843–845, 2006.

[23] M. Toerien, S. T. Brookes, C. Metcalfe, I. De Salis, Z. Tomlin, T. J. Peters, J. Sterne, and J. L. Donovan, "A review of reporting of participant recruitment and retention in rcts in six major journals," *Trials*, vol. 10, no. 1, pp. 1–12, 2009.

[24] C. Treloar, S. Champness, P. L. Simpson, and N. Higginbotham, "Critical appraisal checklist for qualitative research studies," *The Indian Journal of Pediatrics*, vol. 67, no. 5, pp. 347–351, 2000.

[25] E. Hedlund, M. Johnson, and M. Gombolay, "The effects of a robot's performance on human teachers for learning from demonstration tasks," in *Proc. of the 2021 ACM/IEEE Intl. Conf. on HRI*, pp. 207–215, 2021.

[26] B. Brown, J. T. Galea, K. Dubé, P. Davidson, K. Khoshnood, L. Holtzman, L. Marg, and J. Taylor, "The need to track payment incentives to participate in hiv research," *IRB: Ethics & Human Research*, vol. 40, no. 4, pp. 8–12, 2018.

[27] H. F. Lynch, S. Joffe, H. Thirumurthy, D. Xie, and E. A. Largent, "Association between financial incentives and participant deception about study eligibility," *JAMA network open*, vol. 2, no. 1, pp. e187355–e187355, 2019.

[28] E. B. Ripley, "A review of paying research participants: It's time to move beyond the ethical debate," *Journal of Empirical Research on Human Research Ethics*, vol. 1, no. 4, pp. 9–19, 2006.

[29] "Ovpri."

[30] R. Klitzman, "How irbs view and make decisions about coercion and undue influence," *Journal of Medical Ethics*, vol. 39, no. 4, pp. 224–229, 2013.

[31] R. A. Miech and R. M. Hauser, "Socioeconomic status and health at midlife: a comparison of educational attainment with occupation-based indicators," *Annals of epidemiology*, vol. 11, no. 2, pp. 75–84, 2001.

[32] A. Karp, I. Kåreholt, C. Qiu, T. Bellander, B. Winblad, and L. Fratiglioni, "Relation of education and occupation-based socioeconomic status to incident alzheimer's disease," *American journal of epidemiology*, vol. 159, no. 2, pp. 175–183, 2004.

[33] C. W. Mueller and T. L. Parcel, "Measures of socioeconomic status: Alternatives and recommendations," *Child development*, pp. 13–30, 1981.

[34] K. Winkle, P. Caleb-Solly, U. Leonards, A. Turton, and P. Bremner, "Assessing and addressing ethical risk from anthropomorphism and deception in socially assistive robots," in *Proc. of the 2021 ACM/IEEE Intl. Conf. on HRI*, pp. 101–109, 2021.

[35] C. McGinn and D. Dooley, "What should robots feel like?," in *Proc. of the 2020 ACM/IEEE Intl. Conf. on HRI*, pp. 281–288, 2020.

[36] *HRI '20: Companion of the 2020 ACM/IEEE Intl. Conf. on HRI*, (New York, NY, USA), Association for Computing Machinery, 2020.

[37] *HRI '21 Companion: Companion of the 2021 ACM/IEEE Intl. Conf. on HRI*, (New York, NY, USA), ACM, 2021.

[38] M. L. McHugh, "The chi-square test of independence," *Biochemia medica*, vol. 23, no. 2, pp. 143–149, 2013.

[39] N. Caplar, S. Tacchella, and S. Birrer, "Quantitative evaluation of gender bias in astronomical publications from citation counts," *Nature Astronomy*, vol. 1, no. 6, pp. 1–5, 2017.

[40] M. L. Dion, J. L. Sumner, and S. M. Mitchell, "Gendered citation patterns across political science and social science methodology fields," *Political analysis*, vol. 26, no. 3, pp. 312–327, 2018.

[41] J. D. Dworkin, K. A. Linn, E. G. Teich, P. Zurn, R. T. Shinohara, and D. S. Bassett, "The extent and drivers of gender imbalance in neuroscience reference lists," *Nature neuroscience*, vol. 23, no. 8, pp. 918–926, 2020.

[42] D. Maliniak, R. Powers, and B. F. Walter, "The gender citation gap in international relations," *International Organization*, vol. 67, no. 4, pp. 889–922, 2013.

[43] S. M. Mitchell, S. Lange, and H. Brus, "Gendered citation patterns in international relations journals," *International Studies Perspectives*, vol. 14, no. 4, pp. 485–492, 2013.