Towards Formalizing HRI Data Collection Processes

Zhao Han and Tom Williams
MIRRORLab, Department of Computer Science, Colorado School of Mines, Golden, Colorado, USA 80401
zhaohan@mines.edu, twilliams@mines.edu

***Abstract*—Within the human-robot interaction (HRI) community, many researchers have focused on the careful design of human-subjects studies. However, other parts of the community, e.g., the technical advances community, also need to do human-subjects studies to *collect data to train their models*, in ways that require user studies but without a strict experimental design. The design of such data collection is an underexplored area worthy of more attention. In this work, we contribute a clearly defined process to collect data with three steps for machine learning modeling purposes, grounded in recent literature, and detail an use of this process to facilitate the collection of a corpus of referring expressions. Specifically, we discuss our data collection goal and how we worked to encourage well-covered and abundant participant responses, through our design of the task environment, the task itself, and the study procedure. We hope this work would lead to more data collection formalism efforts in the HRI community and a fruitful discussion during the workshop.**

I. Introduction

In a multidisciplinary field like HRI, it is important for researchers to leverage empirical research [1] to discover new knowledge from observations and experience. It is thus common to treat data collection solely within the lens of formal experimental design, to answer research questions by collecting, for example, qualitative data through interviews, surveys, or think-aloud protocols, and quantitative data from sensors or through coding qualitative data [2, 3, 4].

Moreover, while data collected through user studies is increasingly made publicly available, such data is rarely reused. Instead, researchers in HRI tend to build on past datasets through new experiments to replicate that past work either tightly or with carefully controlled deviations, e.g., with other robots ([5, 6]) or in different cultures ([7, 8, 9, 10]). This paradigm has led to substantial recent research seeking to formalize experimental design [4, 11] and analysis [12, 13] efforts within the unique contexts of HRI, with, unfortunately, data collection task design left behind.

Yet, other communities within HRI, such as the technical advances community, also collect human-subjects data, albeit for different purposes, such as collecting and modeling human data for more human-like and familiar interactions to improve robot experience [14, 15]. For example, to advance social navigation, researchers have collected human navigation data to predict human activity [16], human-motion trajectory data (Thör, [17]), and robot approaching behavior towards humans [18, 19]. Data in robots' view has also been collected to allow more practical robotics in unstructured environments [20, 21].

This work has been supported in part by the Office of Naval Research under N00014-21-1-2418.

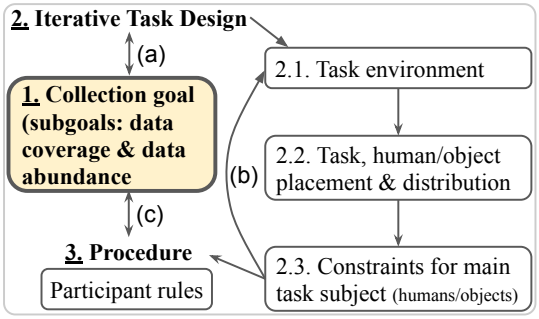

Fig. 1. A flowchart for the process of designing a data collection study, grounded in recent HRI literature (See Table I). Collection goal and data coverage/abundance subgoals are first identified, followed by an iterative task design and procedure for collecting more data. Task design includes three major elements: the environment, the task with human/object configurations, and the criteria for the main task subject identified to reach the collection goal. The two bi-directional arrows, (a) and (c), and an iteration arrow (b) underline the importance of the iterative process

In contrast to the experimental design and analysis works mentioned above, best practices or recommendations for collecting data for these types of purposes, i.e., machine learning modeling, have received less attention within HRI. In order to efficiently collect such data from human subjects, researchers must carefully design a task and a procedure to gather as many types of data as possible (i.e., data coverage) and solicit as much data as possible from participants (i.e., data abundance), as illustrated in Figure 1 left. While the latter is key for data-hungry machine learning techniques, the former is vital for robustness to cover real-world scenarios.

In this workshop paper, we thus make two contributions:

1) A process for designing a data collection study grounded in recent HRI literature
2) A concrete study design example applied the process for a data collection in the context of human-robot dialogue

This paper is organized as follows. We first give an overview of the process and its three steps in Section II with example works. We then detail the case study in Section III, whose task design is more specifically discussed in other recent work [22]. Following the workflow shown in Figure 1, Section III first describes the goal that guides us and defines data coverage in our domain. We then discuss how we designed the task to reach our goal, including task environment and task choice, and object placement and distribution. A brief discussion on the iterative process is also provided. Lastly, we conclude with insights as to how we used extra rules in our procedure to encourage more data to be collected from participants.

II. A Process for Data Collection Study Design

As shown in Figure 1, the process consists of three steps: goal, task design, and procedure. First, the **goal** of the data collection effort needs to be identified. Akin to hypotheses that guide the design of human-subject experiments, it is critical to articulate the precise goals that guide non-experimental data collection efforts. The goal is dependent on the domain or application of the proposed machine learning model. For example, in Taylor et al. [21]'s work, the goal was to collect egocentric color and depth (RGB-D) data of groups of people to predict social groups. In Yang et al. [18]'s work, the goal was to collect different reactions from humans when a robot approaches from different directions to join a conversation.

Concretely, the goal can be divided into two subgoals: data coverage and data abundance. **Data coverage** concerns different types of or forms of data that should be collected. For example, in Taylor et al. [21]'s work, the data was recorded in multiple crowded, sunny, outdoor environments, covering occlusion, shadow, lighting, and motion patterns to handle real-world challenges. In Yang et al. [18]'s work, the authors collected data from two group types, nine approaching directions, and three Wizard-of-Oz robot styles.

While data coverage addresses data quality, **data abundance** addresses data quantity. While Taylor et al. [21] do not explicitly discuss this, they collected 1.5 hours of 16,827 RGB-D frames. In Yang et al. [18]'s work on modeling conversational approaching behavior, the authors used 16 on-body cameras and a Motion Capture suit with 37 markers to gather more data from participants. As we show in our case study, data abundance can also be achieved by deliberately soliciting more data from participants.

Secondly, a **task design** specifically for data collection must be carefully constructed to reach the goal. The task design includes the environment, the task itself with human or object placement and distribution, and some criteria for the main task subject. Because Taylor et al. [21]'s work studies crowd behavior in a public environment, this step was skipped. In Yang et al. [18]'s work, the authors use a three-person "Who's the Spy" game with the robot being adjudicator to identify the spy. While without physical objects, the **task environment** consists of a marked circle that a triad of participants stands on; The robot stands at room corners outside of the circle. The **task** for each participant was to describe the material of the word on a card given to them. While objects are not the focus, standing participants face each other and were distributed in the center of a room. The **main task subject** is the robot that was constrained to only be teleoperated to approach in different directions to join the group when the spy is identified. In our case study, our main task subjects were buildings and we explicitly imposed additional constraints.

Lastly, a well-thought procedure needs to be in place to reach the collection goal. In Yang et al. [18]'s work, participants were asked to stand at fixed positions so they are in the field of view of the cameras. As Taylor et al. [21] studies public groups of people, no explicit procedure was given.

TABLE I
Sample works fitting into the proposed process (Task*: environment, task, and main subject constraints)

	Domain	Goal	Cover.	Abundance	Task*	Procedure
[21]	Group	✓	✓	✗	NA	NA
[18]	Navigation	✓	✓	✓	✓✓✓	✓
[19]	Service	✓	✓	✗	✓✓✓	✓
[17]	Navigation	✓	✓	✓	✓✓✓	✓
[23]	Trust	✓	✓	✗	✓✓✗	✓
[20]	Perception	✓	✓	✓	NA	NA
[24]	Tutoring	✓	✓	✗	✓✓✓	✓
[25]	Speech	✓	✓	✗	✓✓✓	✓

Figure 1's bi-directional arrows and cyclic nature emphasizes the iterative nature of the process centered around a primary collection goal. Yet this iterative process is not typically reported in the literature, similar to user studies where pilot studies may not be reported. In the following section, we will detail our case study and detail the iterative process we followed.

In addition to the two examples, Table I lists six more HRI conference, HRI workshop, and ACM THRI papers from 2018–2021 (filtered with the "data collection" keyword), and whether they fit the proposed process. It is worth noting that only eight papers were found, indicating the need for this work. Moreover, over half of them did not discuss data abundance.

III. A Case Study on Applying the Process

With the process defined with concise, grounded examples, we can now detail a case study from our recent work [22] that applied the described process in the human-robot dialog domain.

A. Collection Goal and Data Coverage

The goal in our example was to collect a corpus of verbal and nonverbal data that is rich and varied yet representative of typical human dialogue patterns. Second, we aimed to collect referring forms and gestures picking out both present, perceivable entities, as well as entities that are *not in the scene* or *were in the scene but are no longer visible* as humans are moved away from the scene. Third, we aimed to collect data that would cover a wide range of referring forms, including pronominal, deictic, and definite forms (e.g., it, this, that, this-N', that-N', and the-N'), as well as indefinite forms such as a-N'. Finally, we wished to collect data that was rich in nonverbal cues like gestures. The data abundance will be described in Section III-C of the procedure.

B. Iterative Task Design

Task Environment: Different from the single tabletop scenarios in previous robot dialogue research [26, 27, 28, 29] where all objects are present in a robot's operating environment, we used a four-quadrant tabletop scenario (Figure 2) by adjoining two tables [30] and separating them into four quadrants with two long foam boards [31], so objects can be hidden in

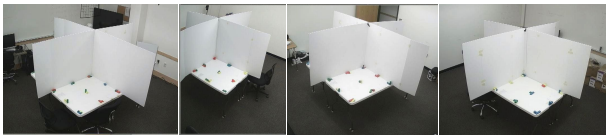

Fig. 2. The four-quadrant task environment, adapted from the 2×2 video multiplexer from four cameras installed at the corners of the room's ceiling. The environment is made by adjoining two tables and four foam boards, which are longer than the table to avoid participant looking into other quadrants.

Fig. 3. Four buildings to be constructed. To help participants identify individual blocks, two angles were provided. Each building has repeated blocks to reach the data collection goal, i.e., wider variety of referring forms.

different quadrants and referers can refer to both present and non-present objects.

Task: The task environment helps reach the goals of encouraging references to both visible and non-visible objects; similarly, the task, number of objects, and object distribution in quadrants should help us to collect more natural language references and gestures. To that end, we chose a series of collaborative *tower building* tasks [32] where instructor participants teach learner participants construct four buildings (Figure 3) from $18 \times 4 = 72$ blocks [33] in different quadrants. The repetitive elements during the building process increase use of reference forms, either in speech or with gestures.

Object Placement and Distribution: We used a number of block shapes, including triangles, cubes, cuboids, cylinders, arches, and half-circles, so they are not too complex to describe and participants can focus on referring to them in the same quadrant or previous quadrants. The blocks required to construct each building are randomly placed at the vertices of a 3×3 grid. This placement strategy leads to varying the physical distance between blocks and encourages referring to visible objects with "this" and "that" [34].

Criteria for Main Task Subject: To cover indefinite nouns (e.g., a N), we constrained the placement of the blocks used to construct buildings as follows: Half of the blocks needed for each building are distributed to the quadrant in which that building is to be constructed, and the other half of the blocks need to be evenly distributed in the other three quadrants. To meet this constraint, each building has an even number of 18 blocks. *Nine of them* are placed in the quadrant where the building is constructed, and each of the other three quadrants has 3 (i.e., $\frac{9}{3}$) blocks, depleting the remaining nine blocks.

Iterative Task Design Process: The task design is an iterative process, similar to interaction design [35]. Feedback and improvement should be incorporated before the design is

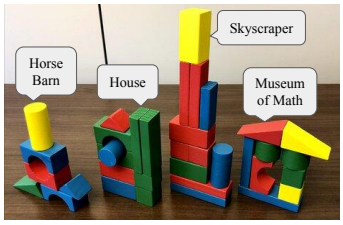

Fig. 4. The **first iteration** of simpler buildings. The task design is an iterative process by gathering feedback and incorporating improvements towards the goal of our collection effort.

finalized. Indeed, the buildings were previously much simpler and consisted of fewer blocks, as shown in Figure 4, but later made more complex to encourage the production of more references and gestures by participants in a single session. To gather feedback, we ran pilot studies and presented the task design in lab meetings.

C. Procedure Design

With the task design in place, we can describe how we designed our study procedure. Once seated, both participants were provided with rule cards as reminders. However, only instructors' cards included the building photos, not visible to the learner participants, encouraging more speech and gestures from instructor participants. Moreover, learners were asked not to speak unless absolutely necessary to proceed, limiting data provided on their part but significantly increasing the amount of language needed to be used by instructors. Similarly, instructors were asked not to touch any blocks, and to only ask learners to find blocks if those blocks were not found in the current quadrant. These tactics encouraged additional language and gestures by instructors, and encouraged instructors to visually search their quadrants before issuing instructions, so as to encourage a wider variety of referring forms.

IV. Discussion and Conclusions

As we have mentioned, our experimental design process was iterative, and was not perfect in the beginning. The flowchart we provide is not a precise recipe that must be followed exactly (as seen that some examples did not follow part of the process). But instead it is provided to make the logic of the design process more clear, serving as a clearer takeaway from this work. Indeed, the experiment design for data collection requires creativity, especially for the task. Hopefully, our work will inspire HRI researchers to step outside of the boundary of the well-established user studies and work more on data collection to develop human-like and familiar interactions.

In conclusion, we contribute a formalized process for a model data collection experiment, informed by recent HRI literature. Centered around reaching a high-level data collection goal, as well as sub-goals regarding data coverage, variety, and abundance, we followed the familiar task design and procedure elements used in traditional experimental designs. We provided a detailed account of the underlying design considerations, and a flow chart that visualizes the steps. In the future, we would

like to expand this workshop paper to a comprehensive meta-analysis of data collection work in HRI and a taxonomy for the the task and procedure design.

REFERENCES

[1] I. S. MacKenzie, *Human-computer interaction: An empirical research perspective*. Newnes, 2012.

[2] V. Eatough and J. A. Smith, "Interpretative phenomenological analysis," *The Sage handbook of qualitative research in psychology*, vol. 179, p. 194, 2008.

[3] C. Jost, B. Le Pévédic, T. Belpaeme, C. Bethel, D. Chrysostomou, N. Crook, M. Grandgeorge, and N. Mirnig, *Human-Robot Interaction: Evaluation Methods and Their Standardization*. Springer Nature, 2020, vol. 12.

[4] G. Hoffman and X. Zhao, "A primer for conducting experiments in human–robot interaction," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 1, pp. 1–31, 2020.

[5] G. Bejerano, P. Robinette, H. A. Yanco, and E. Phillips, "Back to the future: Opinions of autonomous cars over time," in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 157–161.

[6] D. Ullman, S. Aladia, and B. F. Malle, "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*, 2021, pp. 110–118.

[7] G. Trovato, M. Zecca, S. Sessa, L. Jamone, J. Ham, K. Hashimoto, and A. Takanishi, "Cross-cultural study on human-robot greeting interaction: acceptance and discomfort by egyptians and japanese," *Paladyn, Journal of Behavioral Robotics*, vol. 4, no. 2, pp. 83–93, 2013.

[8] K. S. Haring, D. Silvera-Tawil, Y. Matsumoto, M. Velonaki, and K. Watanabe, "Perception of an android robot in japan and australia: A cross-cultural comparison," in *International conference on social robotics*. Springer, 2014, pp. 166–175.

[9] S. Andrist, M. Ziadee, H. Boukaram, B. Mutlu, and M. Sakr, "Effects of culture on the credibility of robot speech: A comparison between english and arabic," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2015, pp. 157–164.

[10] 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 *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 103–111.

[11] C. L. Bethel, Z. Henkel, and K. Baugus, "Conducting studies in human-robot interaction," in *Human-Robot Interaction*. Springer, 2020, pp. 91–124.

[12] M. E. Bartlett, C. Edmunds, T. Belpaeme, and S. Thill, "Have i got the power? analysing and reporting statistical power in hri," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 11, no. 2, pp. 1–16, 2022.

[13] 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 International Conference on Human-Robot Interaction*, 2020, pp. 43–52.

[14] B. Siciliano and O. Khatib, "Humanoid robots: historical perspective, overview and scope," *Humanoid robotics: a reference*, pp. 1–6, 2018.

[15] M. Huggins, S. Alghowinem, S. Jeong, P. Colon-Hernandez, C. Breazeal, and H. W. Park, "Practical guidelines for intent recognition: Bert with minimal training data evaluated in real-world hri application," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 341–350.

[16] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, "Human motion trajectory prediction: A survey," *The International Journal of Robotics Research*, vol. 39, no. 8, pp. 895–935, 2020.

[17] A. Rudenko, T. P. Kucner, C. S. Swaminathan, R. T. Chadalavada, K. O. Arras, and A. J. Lilienthal, "Thör: human-robot navigation data collection and accurate motion trajectories dataset," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 676–682, 2020.

[18] F. Yang, W. Yin, M. Björkman, and C. Peters, "Impact of trajectory generation methods on viewer perception of robot approaching group behaviors," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 509–516.

[19] A. Nanavati, M. Doering, D. Brščić, and T. Kanda, "Autonomously learning one-to-many social interaction logic from human-human interaction data," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 419–427.

[20] R. Martín-Martín, M. Patel, H. Rezatofighi, A. Shenoi, J. Gwak, E. Frankel, A. Sadeghian, and S. Savarese, "JRDB: A dataset and benchmark of egocentric robot visual perception of humans in built environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[21] A. Taylor, D. M. Chan, and L. D. Riek, "Robot-centric perception of human groups," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 3, pp. 1–21, 2020.

[22] Z. Han and T. Williams, "A task design for studying referring behaviors for linguistic HRI," in *Companion of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 2022.

[23] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, "Trust-aware decision making for human-robot collaboration: Model learning and planning," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 2, pp. 1–23, 2020.

[24] O. Engwall, R. Cumbal, J. D. Águas Lopes, M. Ljung, and L. Månsson, "Identification of low-engaged learners in robot-led second language conversations with adults," *ACM Transactions on Human-Robot Interaction*, 2021.

[25] J. Novoa, R. Mahu, J. Wuth, J. P. Escudero, J. Fredes, and N. B. Yoma, "Automatic speech recognition for indoor hri scenarios," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 2, pp. 1–30, 2021.

[26] D. Roy, "Semiotic schemas: A framework for grounding language in action and perception," *Artificial Intelligence*, vol. 167, no. 1-2, pp. 170–205, 2005.

[27] K.-y. Hsiao, S. Vosoughi, S. Tellex, R. Kubat, and D. Roy, "Object schemas for responsive robotic language use," in *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, 2008, pp. 233–240.

[28] C. Matuszek, L. Bo, L. Zettlemoyer, and D. Fox, "Learning from unscripted deictic gesture and language for human-robot interactions," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[29] R. Scalise, S. Li, H. Admoni, S. Rosenthal, and S. S. Srinivasa, "Natural language instructions for human–robot collaborative manipulation," *The International Journal of Robotics Research*, vol. 37, no. 6, pp. 558–565, 2018.

[30] "Mainstays 6 foot fold-in-half table, white granite," https://www.walmart.com/ip/622822527, accessed: 2022-02-20.

[31] "Elmer's 36" x 48" tri-fold foam presentation board - white," https://www.target.com/p/A-13313406, accessed: 2022-02-20.

[32] M. F. Jung, D. DiFranzo, S. Shen, B. Stoll, H. Claure, and A. Lawrence, "Robot-assisted tower construction—a method to study the impact of a robot's allocation behavior on interpersonal dynamics and collaboration in groups," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 1, pp. 1–23, 2020.

[33] "100 piece wood blocks set," https://www.melissaanddoug.com/100-piece-wood-blocks-set/481.html, accessed: 2022-02-20.

[34] R. M. Dixon, "Demonstratives: A cross-linguistic typology," *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, vol. 27, no. 1, pp. 61–112, 2003.

[35] Y. Rodgers, H. Sharp, and J. Preece, "Interaction design: Beyond human-computer interaction," 2011.