Human Prediction of Robot's Intention in Reach Movements

by

Teerachart Soratana

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in the University of Michigan
2023

Doctoral Committee:
Professor Yili Liu, Co-Chair
Associate Professor Xi Jessie Yang, Co-Chair
Assistant Professor Maani Ghaffari Jadidi
Professor Nadine Sarter

Dedication

To my parents and friends

Acknowledgements

This Ph.D. journey was full of unexpected events; some smooth, some rocky. I would like to thank the people I have met and interacted with along this journey, who provide support, advice, and motivation. I acknowledge:

- my advisors, Dr. Yili Liu and Dr. Jessie Yang, for being a very supportive advisor. This work is made possible by their guidance. I am very grateful for their mentorship, knowledge, and patience.
- my dissertation committee members, Dr. Maani Ghaffari and Dr. Nadine Sarter, for their invaluable feedback and insights.
- members of the Interaction and Collaboration Research Lab (ICRL), both current and past members of ICRL. Special shoutout to Patrik, thank you for being an amazing friend for the whole PhD journey. Yaohui, Doowon, Jinyong, Shreyas, Hyesun, Tribhi, Na, Ruikun – thank you for making the lab a special place to be.
- my friends and cohort members in IOE and outside – Zhongzhu, Lauren, Arlen, Xinyu, Rohan, Joseph, Moyan, Yifan, Dom, Kam, Kati, Daniel, KP, Alexandra, Jingwen, Robert, Dory, Rosemary, and Xubo – thank you for many joyous and unforgettable memories. Special thanks to my best friends, Fah and Fon, for a countless numbers of work-together session, and Kay for proofreading some of my work.

- IOE staff - Rod Capp, Matt Irelan, Christopher Konrad, Valerie Martin, Olof Minto, Paul Pridham, Allyson Rodriguez, Krissy Stafiej, Tina Sroka, and Dr. Sheryl Ulin - for making the IOE a great place for students to develop professionally and personally.
- funding sources, including fellowships and Graduate Student Instructor positions (IOE 430) provided by the IOE department and Thai government scholarship.
- my parents, for their love and patience.

Table of Contents

List of Tables

List of Figures

Abstract

Humans can predict another person's intentions from observed movement patterns when they are working together. This intentions prediction allows team members to plan and perform actions in anticipation of other team members. However, the prediction of team members' intentions does not come naturally in human-robot teaming. The first objective of this dissertation research is to examine human prediction of robot's intentions by studying the effects of optimal feedback control laws in a robotic arm on its predictability and perceived human-likeness. Three in-laboratory studies were conducted to examine human prediction of robot's intentions by studying the effects of optimal feedback control laws in a robotic arm on its predictability and perceived human-likeness. These three studies differ in either the number of targets or path planning constraints. Participants observed a robotic arm as it moved toward an object on a shelf. The results showed that low energy expenditure may enable the participants to predict the robot's target quicker. The speed of predictions was significantly affected by the trajectory characteristics of the end-effector. To reach a more diverse group of participants, online studies with online crowdworkers were also proposed. This also brings up a question about the impact on experimental research results introduced by switching the format of interaction from observing a robot's physical movements to watching their videos. This question led to the second objective of the dissertation. An additional study was conducted and compared to one of the in-laboratory studies to investigate the similarities and differences in task quality, subjective experience, motivation, and perceived payment fairness between observing a robot's physical movements and watching their videos in the context of HRI judgment tasks. The findings showed that participants rated the robot's physical movement as less human-like and less life-like, but reported feeling safer, compared to watching the robot's movement through videos. Observing a robot's physical movements yielded better task quality, in terms of accuracy and the desired response time, than watching their video recordings. Additional studies were conducted to investigate similarities and differences between an in-laboratory study with college students as participants and online studies. The online studies utilized an online crowdsourcing platform, on which online crowdworkers were recruited to observe the movements of a robot. The findings revealed that students may produce higher-quality data than crowdworkers. Students were also motivated to complete a study out of interest while the crowdworkers were more focused on compensation. The use of an online crowdsourcing platform also supports the investigation of a research gap in macroergonomics related to the compensation issue on online crowdsourcing platforms, including compensation scheme and fairness and their effects on task performance and subjective experience. This investigation forms the third objective of the dissertation. Four studies were conducted on Amazon Mechanical Turk to explore the effects of participant location, payment method, and payment rate on online crowdwork task quality, subjective experience, motivation, and perceived payment fairness results showed that participant location affects the rate that participants followed written task descriptions. Task accuracy was comparable between the quota and piece-rate payment methods. High-paying tasks were perceived as more fair, however, it did not lead to a higher number of completed tasks in the piece-rate condition. A lower-paying task in the quota condition and a higher-paying task in the piece-rate condition attracted more fraudulent participants.

Chapter 1
Introduction

1.1 Motivation

Human Judgment of Robot's Movement Intention and Movement Quality

With the advances in machine learning and AI, robots working alongside human workers are becoming a reality. One key requirement for such human-robot teams to succeed is the ability of the human and robot team members to predict each other's target during collaboration. This ability would allow team members to plan and perform their actions in anticipation of other team members' actions (Becchio et al., 2012). In human-robot interaction (HRI), this applies to both human and robot team members. When the robot has more information about the task, either because it is programmed by a task expert or because it is designed to operate autonomously, the human observers should be able to predict the robot's target to improve the fluency of collaboration. This predictability is one of the important aspects of transparency in human-machine teaming (McDermott & Dominguez, 2019), which is the user's understanding of robot's future intentions, goals, and limitations.

In order for the human worker to accurately predict the robot's intention, the research community has been investigating methods to communicate the robot's targets using human kinematic models. These kinematic models describe movement behavior in mathematical form. Prior research indicates that the majority of studies addressing this problem considered biomimetics methods, i.e., by imitating movements of living things. These studies discussed human kinematic models in their proposed methods for designing movements for robots and evaluating the movement's human-likeness (Gulletta et al., 2020). Stulp et al. (2015) and Dragan et al. (2015) used models that optimize the prediction accuracy and speed. Their results suggest that in order to enable a human to predict a robot's targets quickly and accurately, instead of solely focusing on maximizing robot movement efficiency, other trajectory characteristics should also be considered. Cabrera (2018) generated human-like variability of a trajectory to be used in classifiers training for recognizing human gestures. The author used the jerk and energy expenditure minimization strategy, which is one of the common optimal feedback control laws. Koppenborg et al. (2017) studied the effects of movement speed and erratic movements on robot's predictability. The authors found that erratic movements led to reduced prediction performances, such as lower accuracy and prediction speed. Despite all the existing efforts, a research gap exists in identifying and understanding the trajectory characteristics to make a robot more predictable and in linking the findings back to the human kinematic models. Transparency of robotic systems may influence how the user places trust in the robot (Elprama et al., 2016).

The first research objective described in this dissertation work is to examine the effects of trajectory characteristics on the judgment of robot's intention, its perceived human-likeness, and its perceived safety. Chapter 3 discusses a series of studies investigated the first research question: how optimal feedback control laws affect human performance and attitude when humans working collaboratively with the robot in the context of object handling, which is a very common application of robotic arms. This work was not aimed to propose a new computational human behavior modeling. It only focuses on using the existing theoretical models in this area. This part aims to build a generalizable and human-interpretable solution to improve teamwork in HRI tasks and connect the findings to the existing knowledge on human voluntary arm movement models. The findings based on the human arm movement models are expected to be intuitive to human observers, thereby improving the fluency of collaboration between humans and collaborative robots.

The data collection in the laboratory environment described in Chapter 3 was originally proposed to be conducted in 2020 but was postponed due to the outbreak of the COVID-19 pandemic and subsequent university-wide temporary restrictions on human subject research. These circumstances set off a pressing need to adapt from a laboratory to an online study and a unique opportunity to investigate the similarities and differences between these two approaches to experimental studies in the context of HRI judgment task. Conducting online research is a common practice that has been utilized before the pandemic as related work addressing similar issues sometimes opts for video recordings of the movements to reach a wider range of participants outside of the laboratory setting (Bainbridge et al., 2011; Dragan et al., 2013; Kose-Bagci et al., 2009). The online experimental study allows researchers to collect data while the in-laboratory research restrictions were in effect while enriching the related literature on the comparability between in-laboratory and online experiments for academic research. However, designing an experimental testbed for an online study that is comparable to its corresponding laboratory study can be challenging due to the limitations of online platforms and differences in the participant pool.

In-laboratory and Online Research in HRI

Online experimental research has gained increasing attention and popularity among researchers as an alternative to in-laboratory experiments. It opens the door for researchers to reach out to a large number of participants with diverse demographics for their studies, usually faster and at a lower cost than laboratory experiments. For example, this alternative allows researchers to recruit non-college students, such as online survey takers or crowdworkers on a crowdwork platform, whose diversity is much higher than college students. Online crowdworkers can be recruited quickly, which reduces the time and financial resource needs in comparison with the traditional in-person laboratory environment (called "in-lab experiments" in this dissertation work). However, important issues must be addressed about how to compare online versus in-lab experiments with regard to their research design, implementation, and results. For example, online and in-lab studies may need to use different formats to deliver task information to the participants, online researchers may not have the same level of control over participants' activity during the experimental period, and the different experimental environments of online and in-lab studies may create confound effects on participants' behavior.

One approach to examining the issues above is by comparing findings from studies conducted online with the well-established findings conducted in-lab or in the field. This approach has been adopted by studies in the field of social psychology, cognitive psychology, and decision-making (Thomas & Clifford, 2017). Many have found that studies conducted online are as valid as their in-laboratory study counterpart (Buhrmester et al., 2011; Casler et al., 2013; Horton et al., 2011; Keith et al., 2022; Lewis et al., 2009; Paolacci et al., 2010), while having the benefit of recruiting diverse participants compared to traditional college student pool (Berinsky et al., 2012; Briones & Benham, 2017; Weigold & Weigold, 2022). However, these studies do not require the participants to interact face-to-face with a machine, a physical device, the researcher, or another participant (Casler et al., 2013). It is unclear whether the results of these studies can be generalized to human-robot interaction (HRI) tasks because HRI tasks require interactions between the participants and the robotic system being studied. See Figure 1-1 for the illustration of the gap between these two experimental designs. It is important to develop a better understanding of the factors that may affect the comparability of online versus in-lab HRI studies.

The second research objective described in this dissertation work is to examine the effects of (1) formats of interaction and (2) participant pools on the findings, task quality, subjective experience, and perceived payment fairness. Chapter 4 discusses an additional study investigating whether introducing changes to the format of interaction affects participants' perception of the robot's human-likeness, life-likeness, and its perceived safety. In Chapter 4, stimuli were presented in two different formats of interaction: a robot's physical movement or their videos. Participant pool and study location were held constant across both conditions. The videos condition is designed to be similar to the version for online crowdworkers. In the bigger scheme of experimental research, this issue may affect not only the outcome of the research but also other aspects such as task quality, subjective experience, and perceived payment fairness.

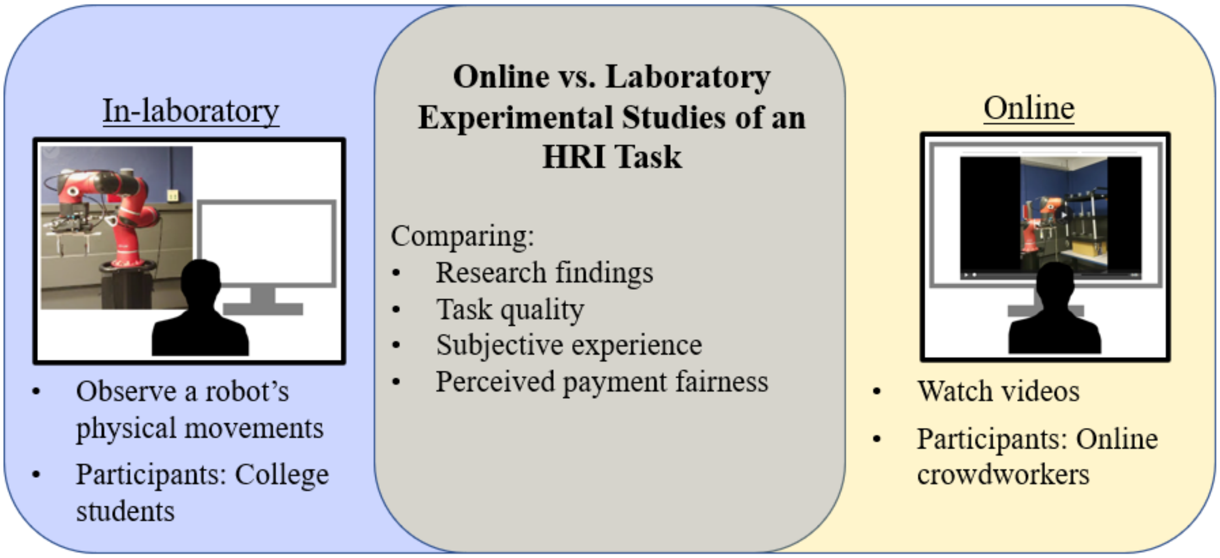

In-laboratory
Observe a robot's physical movements
Participants: College students
Online vs. Laboratory Experimental Studies of an HRI Task
Comparing:
Research findings
Task quality
Subjective experience
Perceived payment fairness
Online
Watch videos
Participants: Online crowdworkers

Figure 1-1: Bridging the gap between in-laboratory and online experimental studies.

Chapter 5 discusses studies conducted to compare the differences in task quality, subjective experience, and perceived payment fairness between two formats of interaction (a robot’s physical movements and their videos) and two participant pools (in-lab participants recruited from college student pool and online participants recruited from a crowdwork platform). The use of an online crowdsourcing platform also reveals a research gap in macroergonomics related to the task quality and subjective experiences on online crowdsourcing platforms, which was discussed in this dissertation as the third research objective.

Crowdworkers in Online Research

One of the emerging human-computer interaction (HCI) task environments is online crowdwork platforms, which provide online services that allow an organization, a company, or a person to outsource tasks that can be completed quickly by a large group of participants via the internet. Data requesters also use these online platforms to recruit participants for a survey or a controlled experiment because of the fast speed of recruiting participants, greater diversity of participants compared to the traditional in-lab participant pool, and the low time and labor investment needed for the requesters. These online platforms introduce a temporary contract between the requesters and participants in the form of gig work, in which participants expect compensation for the time spent working, while requesters expect valid data in return. However, online platforms are not without their problems. A small fraction of participants may not be fully truthful in completing the eligibility screening questions to enroll in a task (Kennedy et al., 2020) or are less attentive to the task requirements to maximize earnings by completing a task quickly (Hauser et al., 2019), which may put the validity of research results into question. The intended online transactions are sometimes left unfulfilled. Some requesters' payments were perceived as unfairly low because they were not required to pay minimum wage and they did not consider the time spent on task search or task rejection (Hara et al., 2018). These problems can also lead to a loss in the requester's effort, time, and money. Hence, there is a need for HCI research on the design of online crowdwork environments, including the crowdwork task payment methods, the participants it attracts, and tools that can support both participants and requesters.

The research community has started to investigate the effects of participant demographics and payment schemes (a combination of payment method and payment rate) on crowdwork work quality and perceived payment fairness to understand how different tasks posted on online crowdwork platforms attract participants with different motivations and crowdwork quality (Chandler & Kapelner, 2013; Irani & Silberman, 2013; Litman et al., 2015; Shaw et al., 2011). Some related studies have identified participants' location (Chandler & Kapelner, 2013; Irani & Silberman, 2013; Litman et al., 2015; Shaw et al., 2011), task payment rate (Yin et al., 2013), and payment method (Mason & Watts, 2009) as factors influencing participant's task performance and quality. To find a suitable payment policy that works for both participants and requesters, studies on the effects of payment rates and payment methods have been conducted (Callison-Burch, 2014; Ikeda & Bernstein, 2016; Litman et al., 2015; Marge et al., 2010). The quota payment method is a default payment method in most social studies and online work, where the participants must complete a certain number of tasks to receive compensation. An alternative method of payment is the piece-rate method, where the participants are paid based on the number of tasks completed. Although the piece-rate method has been proposed as one of the payment methods that can improve task quality and flexibility for the online participants, studies applying it to an online crowdwork platform are limited (Ikeda & Bernstein, 2016; Mason &

Watts, 2009). A research gap exists around understanding the effects of participant location, payment method, and payment rate on task quality, subjective experience, motivation, and perceived payment fairness.

This issue leads to the third research objective discussed in this dissertation. Chapter 6 discusses a series of four studies that were conducted on Amazon Mechanical Turk (AMT) platform to investigate the effects of participant location, payment method, and payment rate on task quality, subjective experience, motivation, and perceived payment fairness.

1.2 Research Questions

The research questions being investigated are as follows:

1.2.1 Human Judgment of Robot's Movement Intention and Movement Quality

RQ1: How do human-like optimal feedback control laws in robotic arms affect (1) human performance in predicting robot's movement intention and (2) perceived human-likeness of the robot?

Hypotheses:

- **H1a** Humans can predict trajectories that minimize energy expenditure, torque, and jerk quicker and more accurately than those relatively deviated from these goals.
- **H1b** Trajectories that minimize energy expenditure, torque, and jerk quicker would be perceived as more human-like.

1.2.2 In-laboratory and Online Research in Human-Robot Interaction (HRI)

RQ2: Do in-lab and online participants provide data of the same quality?

Hypothesis:

- H2: In-lab participants would differ from online participants in terms of the data quality they provide.

1.2.3 Crowdworkers in Online Research

RQ3: How does the payment rate and payment method affect the online workers' task quality, subjective experience, motivation, and perceived payment fairness?

Hypotheses:

- **H3a**: Piece-rate payment method may attract workers with different motivations from those using the quota payment method.
- **H3b**: Piece-rate payment method may negatively affect how the workers perceived the payment fairness of the task.

1.3 Scientific Merits and Practical Impacts

Scientific Merits

- This work will enhance readers' understanding of the interaction between humans and collaborative robots. The research results can inform future researchers on how to design a robot's movements to make them more predictable and more human-like as perceived by their human coworkers.

- This work enriches the related literature and helps develop a deeper understanding of the factors that may affect the comparability of online and in-lab studies, particularly in the domain of human-machine interaction.
- This work will enhance readers' understanding of how online payment methods and the attributes of online crowdsourcing workers affect their task quality, subjective experience, motivation, and perceived payment fairness. The findings can provide information for online crowdwork platform owners to improve their policies on payment distribution and for requesters to improve the design of their tasks based on the workers' attributes.

Practical Impacts

- This work would help achieve a higher level of acceptance of technology, especially for people working closely with robots in the manufacturing sector.
- This work benefits the online crowdwork platform owners, requesters, and workers. It can provide useful information for online crowdwork platform owners to improve their policies on payment distribution and for requesters to improve the design of their tasks.
- This work can provide researchers and requesters on online crowdwork platforms with a more suitable payment scheme than the status quo. A proper payment scheme for crowdwork workers may improve their overall satisfaction and performance.
- This work can raise awareness of payment fairness issues among researchers and online crowdwork platform users (both requesters and workers) with the hope that further discussions of this topic by the stakeholders will help lead to a fairer workplace.

1.4 Overview of the Chapters

This dissertation is organized as follows: After this introduction chapter, Chapter 2 describes a literature review of the background and related work for the series of studies that are reported in this dissertation work. Chapter 3 describes a series of studies that investigate the effects of optimal feedback control laws on robot’s movement predictability and human-likeness. This chapter discusses the experimental design, procedure, trajectory generation, method of analysis, and the results of the studies. Chapter 4 describes a series of studies that investigate the effects of interaction format on robot’s movement predictability and human-likeness, which compareds participants’ judgment of a robot’s movement by either observing the physical robot movement or watching videos of the same robotic movements. Chapter 5 discusses a series of studies comparing an online experiment with its traditional in-lab counterpart by examining the similarities and differences in task quality, subjective experience, motivation, and perceived payment fairness between online crowdworkers and an in-lab experiment. This chapter also discusses a comparison between two formats of interaction: observing a robot’s physical movements vs. their videos. Chapter 6 describes a series of studies that investigate the effects of participant location, payment method, and payment rate on online crowdwork task quality, subjective experience, motivation, and perceived payment fairness. Lastly, Chapter 7 provides the conclusions of this dissertation, limitations, and future works.

1.5 References

Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, *3*, 41–52. https://doi.org/10.1007/s12369-010-0082-7

Becchio, C., Manera, V., Sartori, L., Cavallo, A., & Castiello, U. (2012). Grasping intentions: From thought experiments to empirical evidence. *Frontiers in Human Neuroscience*, *6*, 117. https://doi.org/10.3389/fnhum.2012.00117

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. Com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368. https://doi.org/10.1093/pan/mpr057

Briones, E. M., & Benham, G. (2017). An examination of the equivalency of self-report measures obtained from crowdsourced versus undergraduate student samples. *Behavior Research Methods*, *49*(1), 320–334. https://doi.org/10.3758/s13428-016-0710-8

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *In A. E. Kazdin (Ed.), Methodological Issues and Strategies in Clinical Research (p. 133–139). American Psychological Association*. https://doi.org/10.1037/14805-009

Cabrera Ubaldi, M. E. (2018). *Gist of a Gest: Learning Gestures for the First Time* [PhD Thesis]. Purdue University. Open Access Dissertations. 1911. https://docs.lib.purdue.edu/open_access_dissertations/1911

Callison-Burch, C. (2014). Crowd-workers: Aggregating information across turkers to help them find higher paying work. *Second AAAI Conference on Human Computation and Crowdsourcing*.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156–2160. https://doi.org/10.1016/j.chb.2013.05.009

Chandler, D., & Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, *90*, 123–133. https://doi.org/10.1016/j.jebo.2013.03.003

Dragan, A. D., Bauman, S., Forlizzi, J., & Srinivasa, S. S. (2015). Effects of robot motion on human-robot collaboration. *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 51–58. http://doi.org/10.1145/2696454.2696473.

Dragan, A. D., Lee, K. C. T., & Srinivasa, S. S. (2013). Legibility and predictability of robot motion. *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 301–308. https://doi.org/10.1109/HRI.2013.6483603

Elprama, S. A., Makrini, I. E., Vanderborght, B., & Jacobs, A. (2016). Acceptance of collaborative robots by factory workers: A pilot study on the importance of social cues of anthropomorphic robots. *International Symposium on Robot and Human Interactive Communication*, 7.

Gulletta, G., Erlhagen, W., & Bicho, E. (2020). Human-like arm motion generation: A Review. *Robotics*, *9*(4), 102. https://doi.org/10.3390/robotics9040102

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., & Bigham, J. P. (2018). A data-driven analysis of workers' earnings on Amazon Mechanical Turk. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3174023

Hauser, D., Paolacci, G., & Chandler, J. (2019). Common concerns with MTurk as a participant pool: Evidence and solutions. In *F. R. Kardes, P. M. Herr, & N. Schwarz (Eds.), Handbook of research methods in consumer psychology* (pp. 319–337). Routledge/Taylor & Francis Group. https://doi.org/10.4324/9781351137713-17

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, *14*(3), 399–425. https://doi.org/10.1007/s10683-011-9273-9

Ikeda, K., & Bernstein, M. S. (2016). Pay it backward: Per-task payments on crowdsourcing platforms reduce productivity. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4111–4121. https://doi.org/10.1145/2858036.2858327

Irani, L. C., & Silberman, M. S. (2013). Turkopticon: Interrupting worker invisibility in amazon mechanical turk. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 611–620. https://doi.org/10.1145/2470654.2470742

Keith, M. G., Stevenor, B. A., & McAbee, S. T. (2022). Scale mean and variance differences in MTurk and non-MTurk samples: A meta-analysis. *Journal of Personnel Psychology*. https://doi.org/10.1027/1866-5888/a000309

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. (2020). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods*, *8*(4), 614–629. https://doi.org/10.1017/psrm.2020.6

Koppenborg, M., Nickel, P., Naber, B., Lungfiel, A., & Huelke, M. (2017). Effects of movement speed and predictability in human–robot collaboration. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *27*(4), 197–209. https://doi.org/10.1002/hfm.20703

Kose-Bagci, H., Ferrari, E., Dautenhahn, K., Syrdal, D. S., & Nehaniv, C. L. (2009). Effects of embodiment and gestures on social interaction in drumming games with a humanoid robot. *Advanced Robotics*, *23*(14), 1951–1996. https://doi.org/10.1163/016918609X12518783330360

Lewis, I., Watson, B., & White, K. M. (2009). Internet versus paper-and-pencil survey methods in psychological experiments: Equivalence testing of participant responses to health-related messages. *Australian Journal of Psychology*, *61*(2), 107–116. https://doi.org/10.1080/00049530802105865

Litman, L., Robinson, J., & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US-and India-based workers on Mechanical Turk. *Behavior Research Methods*, *47*(2), 519–528. https://doi.org/10.3758/s13428-014-0483-x

Marge, M., Banerjee, S., & Rudnicky, A. I. (2010). Using the Amazon Mechanical Turk for transcription of spoken language. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 5270–5273. https://doi.org/10.1109/ICASSP.2010.5494979

Mason, W., & Watts, D. J. (2009). Financial incentives and the" performance of crowds". *Proceedings of the ACM SIGKDD Workshop on Human Computation*, 77–85. https://doi.org/10.1145/1600150.1600175

McDermott, P., & Dominguez, C. (2019). Human Machine Teaming (HMT): Design and Evaluation Methods. *Workshop at the 63rd HFES Annual Meeting, Seattle, WA*.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*(5), 411–419. https://doi.org/10.1017/S1930297500002205

Shaw, A. D., Horton, J. J., & Chen, D. L. (2011). Designing incentives for inexpert human raters. *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, 275–284. https://doi.org/10.1145/1958824.1958865

Stulp, F., Grizou, J., Busch, B., & Lopes, M. (2015). Facilitating intention prediction for humans by optimizing robot motions. *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1249–1255. https://doi.org/10.1109/IROS.2015.7353529

Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, *77*, 184–197. https://doi.org/10.1016/j.chb.2017.08.038

Weigold, A., & Weigold, I. K. (2022). Traditional and modern convenience samples: An investigation of college student, Mechanical Turk, and Mechanical Turk college student samples. *Social Science Computer Review*, *40*(5), 1302–1322. https://doi.org/10.1177/08944393211006847

Yin, M., Chen, Y., & Sun, Y.-A. (2013). The effects of performance-contingent financial incentives in online labor markets. *Twenty-Seventh AAAI Conference on Artificial Intelligence*. https://doi.org/10.5555/2891460.2891626

Chapter 2
Literature Review

2.1 Implicit Communication

2.1.1 Implicit Communication in Human-Human Interaction

Humans can communicate their intention while they are performing a task together without words to achieve a common goal (Becchio et al., 2012). The form of communication embedded in the standard action with pragmatic purpose is called “sensorimotor communication” (Pezzulo et al., 2019) e.g., pushing an object; this action has a pragmatic purpose of moving the object, but also communicates the pusher’s intention to move it in a certain direction. Pezzulo et al. (2019) provided two main assumptions for sensorimotor communication: (1) the actions performed by the communicator can be sensed by the observer, and (2) the communicator has a communicative intention when performing the task. Vesper et al. (2011) found that when a pair of participants were instructed to perform a task as synchronously as possible and only one of the participants knew the task objective, the participant reduced the variability in their response time and perform the task faster. Sacheli et al. (2013) found that participants adjusted the hand trajectory to better communicate their intentions to the observers. Arm and hand were generally mentioned as a medium for sensorimotor communication, especially for object-handling tasks (Pezzulo et al., 2019). Zhou and Wachs (2018) investigated in scrub nurse scenario where human observers predict the surgeon’s next action to pick and hand over surgical equipment for the surgeon’s next action. They found that human observers could predict the surgeon’s next action after a few iterations of training with 90% accuracy after watching half of the surgeon’s actions to request the tool. They found that the top motion-based features to predict the intention are head orientations, arm orientations, and arm acceleration. In another word, the surgeon’s sensorimotor communication is reflected within their head/neck posture, eye gaze, and arm positions/orientations.

2.1.2 Implicit Communication in HRI

Sensorimotor communication embedded itself within arm positions/orientations or neck postures can implicitly communicate the robot’s intention to the observers, and thus has been proposed as one of the means to communicate the robot’s targets. Takayama et al. (2011) incorporated head/neck posture to communicate the humanoid robot’s intent to perform a task to observers. They used animation of engagement (moving a slight distance toward or away from the area of interest), confidence, and timing on the motion of a robot to show forethought before the robot performs a task. They reported that these movements increase the anticipation of the task that is about to be performed.

Besides the research on humanoid robots, researchers have also investigated how humans could predict the movement of robotic arms, either an industrial robot or a collaborative robot. The sensorimotor communication was embedded in the kinematic models of the arm to generate trajectories that can be predicted quicker and more accurately. Stulp et al. (2015) proposed a model to reduce prediction error and amount of time observers took to predict a robot’s intention. Through trial-and-error interactions between human and robot, Stulp et al. (2015) modeled a movement that minimizes (1) trajectory duration, (2) time between the start of robot trajectory and human prediction (3) incorrect prediction, and (4) sum of jerk in joint space (Stulp et al., 2015). Prediction accuracy was given the highest weight. They found that the model generated a trajectory with higher jerkiness over time while reducing the time an observer took to predict a robot's intention. Dragan et al. (2015) compared trajectories generated in three different ways: functional, predictable, and legible. A functional trajectory generated with the Rapidly-exploring Random Trees (RRT) planner (Kuffner & LaValle, 2000) was erratic, inefficient, and produced an unnatural pose for the robotic arm. A predictable trajectory, generated by minimizing integral over squared velocities. This cost function produced a smooth and short path toward the target. A legible trajectory maximized the probability that an observer would make a correct prediction quicker. Both Stulp et al. (2015) and Dragan et al. (2015) provided similar insights: there are other pronounced features to enable the observer to make a correct prediction quicker that deviate from minimizing the trajectory's jerk or maximizing efficiency.

However, these methods described above require multiple trainings with the observer(s) to get prediction accuracy and prediction time. Hence, it may not be able to produce a generalizable model (Stulp et al., 2015). Dragan et al. (2015) also reported that a functional trajectory can come across as tricky to predict because of unexpected changes of direction. Similarly, Koppenborg et al. (2017) defined their movements with low predictability condition as "erratic movements with unexpected breaks and changes of direction coupled." The previously mentioned work brought up descriptions of trajectory characteristics as part of their approaches. Stulp et al. (2015) defined their "energy" term as the sum of jerk in joint space. Dragan et al. (2015) defined their "efficiency" as minimizing squared velocities. Koppenborg et al. (2017) characterized their "predictable" movements as ones without "erratic movements." The usage of these trajectory characteristics as part of the investigation for a kinematic model can be linked to a human-centered technique for trajectory planning (Gulletta et al., 2020), which is supported by theories of human optimal feedback control law.

2.2 Human Optimal Feedback Control Laws

2.2.1 Trajectory Characteristics

Human optimal feedback control law describes a low-level motor control that takes in the target's location and produces motor commands to the muscle while taking in sensory feedback from the muscle to adjust its output signal (Scott, 2004). Generally, when someone wants to reach for an object, they are conscious of their intent to reach for it, but the exact detail of this execution would be controlled by this low-level motor control. This function would produce "features" of human movement. For example, the reaching movements are relatively straight and have bell-shaped velocity profiles (Scott, 2004). Scott (2004) argued that this feedback control law can be defined with a mathematical expression. However, there are many theories on how to define optimal feedback control law, both abstractly and mathematically. Note that these optimal feedback control laws are context and task-dependent (Schwartz, 2016). An optimal feedback control law that works on reaching movements may not work for other contexts, such as drawing.

The candidates for the cost function of human optimal feedback control laws are two-thirds power law (Gulletta et al., 2020; Lacquaniti et al., 1983; Todorov & Jordan, 1998; Viviani & Flash, 1995), jerk minimization (Flash & Hogan, 1985; Glasauer et al., 2010; Todorov & Jordan, 1998), energy minimization (Uno et al., 1989), and torque minimization (Schwartz, 2016; Uno et al., 1989). Two-thirds power law theorized that there is a non-linear relationship between end-effector curvature and speed when humans perform reaching tasks (Gulletta et al., 2020). Lacquaniti et al. (1983) applied two-thirds power law and supported this theory on a tracking task on a 2D space. Similarly, Viviani and Flash (1995) argued that the two-thirds power law can be derived from a minimum jerk model. Their work was validated using a drawing task, which is a 2D tracking task. The minimum jerk model is one of the common theories for human arm path planning. Flash and Hogan (1985) validated this model by comparing the human arm trajectories for a planar movement without external forces to a mathematical model that minimizes the jerk of the hand trajectory. The authors argued that the minimum jerk model is preferred because it minimized abrupt changes in force (Flash & Hogan, 1985). Todorov and Jordan (1998) also proposed a minimum jerk model and tested with scenarios such as 2D without viapoint, 2D with viapoint, and 3D task. They found that their models, which focus on the smoothness constraint, produce fewer errors in the prediction compared to the two-thirds power law. Glasauer et al. (2010) validated the minimum jerk model in the handover task. They compared human-human handover with human-robot handover tasks. They found that the minimum jerk model produced a trajectory that required less time to predict compared to trapezoidal profiles. Uno et al. (1989) proposed minimizing torque model as the feedback loop. This torque minimization model was described as producing a smooth torque profile and trajectory. The rationale provided for this model is it reduces tear and wear on the musculoskeletal system (Uno et al., 1989). They compared human arm trajectories in planar movement under external forces with the trajectory generated under minimizing torque change model (Uno et al., 1989). The trajectory generated with minimizing torque model was reported to have relatively low energy expenditure and hence avoid unnecessary force.

2.2.2 Psychological Aspect of Optimal Feedback Control Laws

These theories on human optimal feedback control laws were used not only in research on the prediction of robot's intention but also in the work related to psychological aspect of HRI. Arai et al. (2010) pointed out that the human observers' state of mind, such as fear and surprise, can be affected by the robot's trajectory characteristics when the robots are working hand-in-hand with humans. They reported that when the robot was moving at a high speed (1 m/s), the observers reported higher fear and surprise than at lower speeds (0.25 m/s and 0.5 m/s). Koppenborg et al. (2017) found that movements with low predictability (erratic movements with sudden changes in directions) coupled with high movement speed led to higher workload and lower task performance and perceived safety. Kulic and Croft (2007) found that there is a strong correlation between speed and anxiety, agitation, and surprise. They found that a trajectory with a high speed may negatively affect perceived safety. This perceived safety score was described using combinations of terms in affective states, such as comfort, fear, anxiety, and surprise (Rubagotti et al., 2022). To measure subjective safety and other subjective metrics of observer's attitude toward a robot, Bartneck et al. (2009) produced a questionnaire to measure robot's perceived anthropomorphism, perceived animacy, perceived intelligence, perceived likability, and perceived safety. Anthropomorphism was defined as having the attributes of a human's form, behavior, or characteristics. Animacy was defined as the quality of being life-like. Epley et al. (2007) argued that the degree of anthropomorphism can be improved by an increase in incentive to understand and predict the other. In the context of collaborative tasks with a robot, this assumption on incentive is applicable when humans are working with a robot that they did not personally program, and hence benefit from understanding and predicting the robot's target. Castro-González et al. (2016) found that movements with sudden changes in direction lead to lower perceived animacy compared to smooth movements in a bimanual robot. However, when a robot was in one-arm configuration, there is no difference between perceived animacy between smooth and sudden changes movements (Castro-González et al., 2016). Tremoulet and Feldman (2000) found that a point robot scored higher in perceived animacy when the robot change its direction with larger angular magnitude (closer to perpendicular) than smaller angular magnitude. They also found that the change in speed also affected perceived animacy score; a point robot scored higher in animacy when it gained speed than when it stayed at a constant speed or slowed down (Tremoulet & Feldman, 2000).

Based on this literature review, two assumptions have been made: (1) optimal feedback control laws are associated with human arm motion planning and execution, and (2) humans can predict the intention of other humans by observing body movement (Becchio et al., 2012). This literature review section points to a gap in research on the effects of trajectory characteristics related to optimal feedback control laws on the human prediction of robot's intention in reach movements, perceived human-likeness, and perceived safety of the robot.

2.3 Robot's Physical Movements vs. Their Videos for HRI Research

2.3.1 Online vs. In-laboratory Research

Online studies provide some advantages not offered in the in-lab environment, such as requiring less labor, time, and money (Buhrmester et al., 2011; Kraut et al., 2004). Many have found that online crowdworkers and college students behave similarly (Casler et al., 2013; Horton et al., 2011; Keith et al., 2022; Lewis et al., 2009; Paolacci et al., 2010). However, some important issues remain to be addressed. For example, recruiting convenient samples, such as students or AMT participants, may affect the validity of the research outcome because the sample pools and environments may not represent the targeted population and task setting. Landers and Behrend (2015) argued that both samples have their own positive and negative implications for the external and internal validities of research, which concern whether the findings are generalizable to the real-life scenarios and whether the design of the study answers the intended research question.

External validity is ensured when the study's findings can be generalized to the settings and the participants in the targeted scenario (Horton et al., 2011). AMT participants are more demographically diverse than college students (Berinsky et al., 2012; Briones & Benham, 2017; Weigold & Weigold, 2022) while college students are dubbed as WEIRD (Western, Educated, Industrialized, Rich, and Democratic) (Henrich et al., 2010). The WEIRD sample is not an ideal group to draw generalizable findings from because they are different from the rest of the population.

Internal validity is ensured when the design and procedure of the experiment truly examine the research question being investigated (Horton et al., 2011). The main concerns on internal validity for online study include the possibility that data are collected from inattentive participants and under other confounded influences of environmental factors outside the researcher's control. Another concern for internal validity is the format of information presentation employed in the experiment. AMT participants are restricted to an online environment with relevant media for information presentation (such as online questionnaires, still images, or videos), which might not simulate an actual real-world task scenario. In a lab environment, the choice of media is less limited than in online studies and extraneous variables such as distraction or inattention are more easily observed and controlled.

2.3.2 Online Crowdworkers vs. College Students

Some online studies have been compared with the traditional in-lab studies to identify the similarities and differences between these two participant pools in terms of demographic diversity, motivation, task performance, and behavior. In terms of demographic diversity, AMT participants generally have higher average age with a larger age range (Berinsky et al., 2012; Buchheit et al., 2018; Kees et al., 2017; Lewis et al., 2009; Weigold & Weigold, 2022), they also have more work experience, but fewer years of education compared to college students (Buchheit et al., 2018). AMT participants outside the US reported to have more years of education compared to the US-based AMT participants and US-based general household samples (Smith et al., 2016). In terms of task performance, college students were found to perform better on a general reasoning task than AMT participants. Buchheit et al. (2018) found that graduate students scored higher in the fluid intelligence measures [Cognitive reflection test (Frederick, 2005) and Raven Progressive Matrices (RPM) (Raven, 2003)] than both crowdworkers on AMT and undergraduate students. AMT participants also reported to expend more effort on the same task compared to college students (Buchheit et al., 2018).

The motivation and incentive to complete tasks also differ between the two sample pools. AMT participants enrolled in a research study online for money (Mason & Suri, 2012) while traditional college students may enroll in a research study for course credits, as seen in many related works in psychology (Dandurand et al., 2008; Hamby & Taylor, 2016; Hauser & Schwarz, 2016; Lewis et al., 2009; Weigold & Weigold, 2022), accounting (Buchheit et al., 2018), and marketing research (Kees et al., 2017). The percentage of AMT participants who reported that they would not participate without compensation was higher than college students completing a task via an online survey link, which can be completed outside research laboratory environment (Kees et al., 2017). In cases where monetary incentives were provided for both sample pools, college students generally received a better payment rate than AMT participants (Casler et al., 2013; Horton et al., 2011).

The percentage of participants who completed a research study once they have started it was also different. Paolacci et al. (2010) found that college students in-lab were more likely to complete the task (98.6%) than AMT participants (91.6%). This might be because AMT participants do not receive a penalty to their work reputation for not completing a task. AMT participants have the option to withdraw their participation without penalty but they would not receive compensation (Amazon Mechanical Turk Inc., n.d.a) and need to spend time looking for another task, which is itself an unpaid task.

In terms of behavior, AMT participants behave similarly to college students (Casler et al., 2013; Horton et al., 2011; Keith et al., 2022; Lewis et al., 2009; Paolacci et al., 2010). Horton et al. (2011) found that AMT participants behave similarly to college students and well-known psychology experiments on priming and framing effects. AMT participants demonstrated similar pro-social behavior to that of college students. Paolacci et al. (2010) found that US-based AMT participants and college students have displayed similar behavior of framing effect, and outcome bias, but not for the conjunction fallacy. The authors justify this finding as conjunction fallacy has high expected variability of results, consistent with the literature on conjunction fallacy (Charness et al., 2010). Casler et al. (2013) found no significant differences between AMT and college students in terms of behavior on an object selection and categorization task. Buchheit et al. (2018) found that the task quality on reasoning tasks was comparable between undergraduate and AMT participants because AMT participants are more familiar with the questionnaire than undergraduate students. However, they also found that graduate students performed better on the reasoning task than AMT participants and undergraduate students. Keith et al. (2022) performed a meta-analysis on publications in a psychology journal and found that AMT participants have similar mean compared to non-AMT participants in many measures of interest in the psychology field, albeit tended to have higher variance compared to non-AMT participants. Keith et al. (2022) asserted that approval ratings, location qualification, and English proficiency in AMT participants can display different behavior than that of non-AMT participants. Location of AMT participants was also found to affect the number of participants who untruthfully report a demographic (known as location spoofing) (Keith et al., 2022; Soratana et al., 2022) or use automated software to complete survey (Moss & Litman, 2018). This brings up another concern about the effects of participants' focused attention on the internal validity and how to screen out participants who paid insufficient attention to the study's instruction.

2.3.3 Focused Attention on Experimental Task

Participants who do not understand the study's instructions may damage the internal validity as the observations would consist of arbitrary answers, hence methods to screen out participants who paid too little focused attention on the study's instructions are required. A common method to quantitatively measure participants' engagement with experimental task is through "attention check questions," which are generally in one of the four forms described as follows (Thomas & Clifford, 2017). First, the participants are asked a question to check the participants' comprehension of study-related material (Berinsky et al., 2012). Second, they are asked a question with a known answer unrelated to the experimental context, known as bogus item (Maniaci & Rogge, 2014). Third, a textual instruction embedded within a large passage of text asks participants to perform an unintuitive task, known as Instructional Manipulation Check or IMC (Oppenheimer et al., 2009). For an example of an IMC, a large block of text is provided and occasionally a short text passage would appear to instruct the participants to perform a task outside of the posted question, e.g., writing a line “I read the instruction” as a response instead of answering the posted question. Fourth, consistency is calculated between the same questions asked at a different time or session, known as test-retest reliability (Buhrmester et al., 2011).

An investigation conducted by Paolacci et al. (2010) utilized attention check questions with known answers to compare the rate that participants failed to notice these questions between AMT participants, college students, and online discussion board participants. The authors did not find significant differences in the number of failed attention check questions across the three subject pools. Kees et al. (2017) also used the same approach to attention check questions, but they found that AMT participants were less likely to fail attention check questions, followed by students who participated in-lab, and followed by students who completed the survey online via a link and other online panels. This finding pointed out that survey-taking experiences may affect the rate at which participants detect attention check questions.

Hauser and Schwarz (2016) utilized the original IMC questionnaire from Oppenheimer et al. (2009) and variants of IMC to investigate the effects of presentation format on rate of detecting attention check questions between AMT participants and college students. The other two formats included (1) a slightly reworded version of the original IMC questionnaire and (2) the IMC questionnaire on the introduction page instead of having it on the same page as the survey. They found that AMT participants were more attentive to the instructions than college students in all three different formats of attention checks.

Smith et al. (2016) utilized test-retest reliability as a metric to measure whether participants pay attention to the questionnaire or not. They found that AMT participants from outside of the US performed scored lower in the test-retest reliability than US-based AMT participants and US-based household participants. The authors asserted that AMT participants from outside of the US were less attentive to the questionnaire.

From the literature, there seem to be at least three factors that affect the rate of detecting attention-check questions: experience, language proficiency, and the environment of the experiment. First, experiences in survey taking may allow the participants to detect attention-check questions more often. AMT participants have more experience as expert survey takers and were found to detect attention check questions more often than college students (Hauser & Schwarz, 2016; Kees et al., 2017). Note that experience in survey taking alone does not necessarily mean the experienced participants would detect attention check questions more often. Weigold and Weigold (2022) found that more than 80% of college students and AMT participants who are not enrolled in a university did not miss any attention check, but only about 53% of AMT participants who are enrolled in a university did not miss any attention check. Second, non-native English speakers are less likely to detect attention check questions, which require language comprehension skills. Smith et al. (2016) have found that non-native English speakers on AMT missed attention check questions more than US-based AMT participants and US-based household participants. Hauser and Schwarz (2016) also found that AMT participants who are non-native English speakers were less likely to detect attention check questions than college students who are native English speakers. Third, a laboratory environment with researchers' presence may induce higher attention to tasks from participants (Kees et al., 2017; Oppenheimer et al., 2009). Kees et al. (2017) found that students in the lab were more likely to detect attention check questions than students who completed the survey online via a link and online panels. Oppenheimer et al. (2009) also expected that the presence of in-lab researchers leads college students to take the experiment session more seriously.

2.4 Online Crowdworkers

2.4.1 Research Using Online Platforms and Online Participants

Online crowdwork platforms allow a person or organization to post their work online and outsource the work to participants who sign up to work. This type of work is generally called gig work, where the workers are paid to complete a small task rather than traditionally employed by the person or organization. This activity is usually conducted via an online crowdwork platform that acts as the middleman; these online platforms provide an interface where requesters can outsource tasks and workers can look up the list of tasks. Examples of these online crowdwork platforms are AMT, CrowdFlower, Qualtrics, etc. Online crowdwork platforms have become popular in the research community due to the fast speed of participant recruitment, greater diversity of participants, lower investment in money and time from the researcher, and providing acceptable data (Buhrmester et al., 2011). However, online workers are diverse in many aspects such as motivation, skills, and cultures. Requesters also have diverse needs for the task to be outsourced, such as quality requirements and samples that represent the targeted population pool. Finding a payment scheme that works best for both sides is still a work in progress. This literature review discusses the aspects mentioned above, as well as the challenges that come with designing a payment scheme that supports both workers and requesters.

Many research findings pointed out that AMT participants complete tasks for money (Hara et al., 2018; Martin et al., 2014), which is how the platform is promoted (Antin & Shaw, 2012). However, money might not be the only reason that participants stay and work on these online crowdwork platforms (Mason & Watts, 2009). Brewer et al. (2016) conducted studies with participants aged 60 - 80 years old. Most of the participants answered that the main motivation to complete a task online was either for money or for fun. This points out that AMT participants can be categorized into different types based on their motivation (Oppenlaender et al., 2020): some workers are motivated to perform the tasks for money as their main objective, some for fun, and some for learning. Contribution to society is also part of the motivation (Chandler & Kapelner, 2013; Kaufmann et al., 2011) and can affect the quality of work. Chandler and Kapelner (2013) found that meaningful treatment of the collected data increases the quality of the work output. The work presented with good contexts (e.g., research in healthcare) would produce higher quality work compared to the same task but without context or the same task that was told the data were set to be discarded.

Buhrmester et al. (2011) found that increasing the payment does not affect data quality, but increases the recruitment rate. AMT participants also have comparable data quality to that of standard internet across different levels of payment. In this group of AMT participants, they reported that their main motivations were to enjoy doing interesting tasks, followed by to kill time, to have fun, to make money, and lastly to learn. The amount of payment reported in the study of Buhrmester et al. (2011) ranges from $0.04 - $6.00/hour. Participants who seek tasks that pay much lower than the “minimum payment,” according to Guidelines for Academic Requesters (Dynamo, 2014) compiled by Dynamo users (Salehi et al., 2015), may have a different motivation than to complete a task on AMT. In another study conducted with low payment (approximately $1.20 - $1.80/hour), Kaufmann et al. (2011) found that participants were motivated to complete tasks on AMT mainly for receiving payment, which is followed by autonomy to choose a task and to use skill set, then by improving a skill and completing a task as a pastime. Litman et al. (2015) used the same motivation questionnaires as Buhrmester et al. (2011); however, their result showed that the US-based and India-based participants rated their motivation to make money the highest, followed by enjoying interesting tasks, learning, to have fun, and to kill time. There is clearly a need for further research on how to design a payment scheme that aligns with participants' motivation while paying a suitable amount to the participants for their work.

2.4.2 Payment Methods

Pay-per-time is common for hourly employment for in-person tasks and is similarly adopted by the AMT platform with payment at the end using the estimated time taken to complete a task (also called Human Intelligence Task or HIT on AMT). Due to the limitation of the AMT site, the requesters set the price per task and the workers only see the price per task when they look up the list without an estimate of the hourly payment rate (Callison-Burch, 2014). This payment per task can be based on the estimated time taken to complete the task, which is based on the requester's estimation. This estimate may bias in the favor of the requester, leading to low payment (Hara et al., 2018). To make the payment more fair, the requester must have an accurate estimation of the amount of time the task needs to take, including the time to search for the task. For a computer task, estimating the time taken to complete an online task is a challenge. The estimation of task time was usually set by the requesters. Alternative suggestions include using the time recorded by the browser that the workers take to complete the task (Callison-Burch, 2014), or the time reported by the participants (Whiting et al., 2019). However, these methods can be biased because the participant can overestimate their time spent working.

The quota payment method is the default payment method in most social studies and AMT, where the participants must complete a certain number of tasks to receive compensation. This payment method is also called pay-in-bulk. This method fits the task in which a certain number of trials are required for each participant, and the participants need to spend the amount of time proportional to the number of trials. For online crowdwork, this method provides less flexibility for the participants as they must allocate more time to complete the whole task. Withdrawal before completing all the tasks means losing the compensation for the time they have already spent on the task. The Belmont Report (National Commission for the Protection of Human Subjects of Biomedical - Behavioral Research, United States, 1978) suggested that research studies should inform the participants that they may withdraw from the research study at any point. Although this report did not make any suggestion on the amount of compensation entitled to the participants when they withdraw, some research studies indeed compensate participants who withdraw from the study early, depending on the progress of the completed study session. Partial payment for early withdrawal is somewhat similar to the piece-rate method described below, where the participants are paid based on the number of tasks completed. However, the quota method requires the participants to complete all the tasks, even though some studies may offer partial payment for early withdrawal.

The piece-rate payment method is a system where the participants are told explicitly that they are paid based on the number of tasks completed. This system is also called “pay-per-unit.” The piece-rate payment method was implemented in the manufacturing line as described in the report by Lazear (2000). They found that the piece-rate payment method increased the productivity of an auto glass installer task. In the case of an in-person task, manufacturing procedures and quality control processes have clear guidelines on quality control and the workers are also trained. For the online crowdwork platform, Ikeda and Bernstein (2016) studied the quota payment method compared to the piece-rate payment method. They found that participants preferred the piece-rate method over the quota method, but there were no significant differences in task quality. However, the authors also found that the quota method has a higher completion rate than the piece-rate method. A limitation of the study is that Ikeda and Bernstein (2016) provided payment in the form of Amazon credits and coupons for telephone bills, which may attract a different group of participants than the one on the AMT platform. In a similar work, Mason and Watts (2009) suggested that the piece-rate payment method may produce a lower quality of work. They found that the pay-per-puzzle (similar to the quota payment method) produces higher work quality than the pay-per-word (similar to the piece-rate payment method) in a simulated task of a word search puzzle.

For both the quota and the piece-rate payment methods, the challenge is how to estimate the amount of time it takes to complete a task. The tasks described in Lazear (2000) are physical tasks conducted in person. The length of a physical task can be estimated using MODular Arrangement of Predetermined Time Standards (MODAPTS). However, estimating the time to complete a computer task is difficult (Bedny et al., 2012; Oppenlaender et al., 2020; Sengupta et al., 2011). It can vary depending on the level of concentration required, decision-making involved in the tasks (Bedny et al., 2012), the complexity and learnability of the tasks (Sengupta et al., 2010), and workers' expertise. This last point is relevant to studies on AMT as participants from different backgrounds and skills gather on the online crowdwork platform. This aspect has been mentioned in research work related to creative tasks, where the amount of time to complete a task can vary between participants (Oppenlaender et al., 2020; Wu et al., 2014). When the time estimation is off in the favor of the requester, the payment that participants received would be unfair, as discussed in the following section.

2.4.3 Payment Fairness

One of the notable issues on AMT is that the median hourly payment is low compared to the minimum hourly wage in the US (Callison-Burch, 2014; d'Eon et al., 2019; Hara et al., 2018; Irani & Silberman, 2013; Martin et al., 2014; Salehi et al., 2015; Whiting et al., 2019) even though payment is one of the main motivations for participants to complete tasks online. Hara et al. (2018) found in their study that the median wage on AMT was around $2/hour. This is due to several reasons: (1) requesters post a low-pay online task as it is the status quo on the AMT, (2) it takes time for the participants to search for a task – which is an unpaid part of working on AMT, and (3) rejection of the tasks by requesters. Requestors may feel inclined to pay less because posting online tasks on AMT is costly. AMT charges at least 20% fee for a requester to post an online task, with an additional 20% if you recruit 10 people or more (Amazon Mechanical Turk Inc., n.d.b). If a requester pays 10 participants for 1 hour for $10 per hour per participant, only $7.14 would be paid to each participant with AMT's additional fees. This also does not account for the time spent searching for the tasks or when a participant's submitted tasks are rejected by the requester (Hara et al., 2018). Thus, the participant would receive less than $7.14/hour even though the requestor pays $10/hour per participant. For a fair transaction, two main aspects must be designed appropriately: autonomy and reciprocity (Sandel, 2009). In the context of an online crowdwork platform, the autonomy aspect of the design focuses on power difference between participants and requesters, i.e., both sides should have equal bargaining power. The reciprocity aspect focuses on paying the appropriate amount of payment for the output by the participants. This also translates into the two main approaches taken to address this fairness issue: empower the participants or adjust the payment scheme of the requesters.

Information-sharing sites and tools are the main approaches to empowering participants. MTurknation, MTurkforum, and Turkopticon (Irani & Silberman, 2013) were websites where participants can share information and rate requesters. An argument against information sharing is that it may help cheaters and speeders (participants who produced unacceptable data) to find the right requirement to enroll in a task for which they are not qualified. Another issue is that information-sharing websites can direct participants toward higher pay jobs, but the issue of unfair payment persists due to a large percentage of tasks providing low payment. Dynamo (Salehi et al., 2015) was a website for collective action that created guidelines for requesters (Dynamo, 2014) and drafted an open letter to the CEO of Amazon.com, which was a step toward empowering the participants. Crowd-Workers (Callison-Burch, 2014) was a web tool that helped participants find higher-paying jobs. This tool tracked hourly wages by recording the task duration and reward. Turkbench (Hanrahan et al., 2019) was a web tool that crawled and sorted the AMT site to find better-paying jobs and reduce unpaid tasks (such as manually crawling the site). However, it was not able to keep up with the pace of the site because of the large volume of tasks being posted.

Another standpoint is that requesters should make changes to their payment scheme to properly reciprocate the worker’s effort and output. Guidelines for Academic Requesters (Dynamo, 2014) compiled by Dynamo users (Salehi et al., 2015) suggest that the recommended minimum payment is $0.10/minute or $6/hour. This recommended minimum payment also showed up in an ethnomethodological analysis on AMT participants’ forum by Martin et al.

(2014). The participants discussed that the current status quo payment rate on AMT is low, and $6/hour is an agreeable wage on AMT. However, estimating the time a computer task takes is a challenge. Whiting et al. (2019) built an online tool to survey the participants on how long the tasks take and then make a recommendation to the requester on how to set the payment rate per task. However, this system depends on the participants to report the time it takes to complete the task, which may vary between individual skills, hence requiring multiple observations to make an accurate estimation. Also, in this case, the participants have the incentive to overestimate the time taken to complete the task. Another approach is to make similar tasks pay the same amount of payment. Borromeo et al. (2017) supported a distributive fairness standpoint, such that the participants who contribute the same amount should receive the same payment.

Measuring payment fairness is also a challenge as the perceived value of a task is subjective, which depends on the initial task value shown at the beginning of the task (Mason & Watts, 2009; Yin et al., 2013) or on the cost of living (Litman et al., 2015). Mason and Watts (2009) found that there was an anchoring effect on the perceived task value. Participants would assign the perceived task value proportional to the initial task value shown to the participants when recruited even for the same task. Similarly, Yin et al. (2013) found that task quality increased when the work offered a subsequent task that paid more and decreased when the subsequent task paid less. Although increasing the payment rate may lead to higher perceived payment fairness, it may not affect the quality of the work. For example, in the case of the picture-ordering task and word-search puzzle, the increase in payment rate did not increase task quality (Mason & Watts, 2009). Participant demographics may also affect how each participant defines fair payment. The largest group of participants on AMT is from the US, followed by India (Difallah et al., 2018). These two countries are different in several aspects: minimum wage

(Litman et al., 2015), cost of living, and education level (Irani & Silberman, 2013). These differences would have strong impacts on what would consider a "fair payment" as the minimum wage in India is lower than in the US. However, in terms of performance, Shaw et al. (2011) found that US participants perform better in a content analysis task compared to India-based participants, which may also relate to the participant's English language proficiency. The payment rate of this study might be low, as they were paying $0.30/task, but the author did not specify the task duration. Chandler and Kapelner (2013) found that US-based participants perform with higher task quality than India-based participants in an image labeling task, but they scored similarly in attention check questions.

This literature review section presented above points to a gap in research on the effects of payment method, payment rate, and participant location on task quality, subjective experience, motivation, and perceived payment fairness. Hence, the goal of the four studies reported in Chapter 6 of this dissertation is to help address the issues surrounding the payment methods between the quota and piece-rate payment methods and participant location. The result of this work may lead to a better understanding of appropriate payment schemes suitable for the participants on the online crowdwork platforms.

2.5 Summary

This chapter presents a literature review on implicit communication between the human-human team and human-robot team, followed by human optimal feedback control laws. Then, it discusses related literature on a robot's physical movements vs. their videos as formats of interaction between humans and the robotic system being evaluated. Lastly, this chapter discusses a literature review of studies conducted on online crowdwork platforms.

2.6 References

Amazon Mechanical Turk Inc. (n.d.a). *Frequently Asked Questions*. https://www.mturk.com/worker/help

Amazon Mechanical Turk Inc. (n.d.b). *Pricing*. https://www.mturk.com/pricing

Antin, J., & Shaw, A. (2012). Social desirability bias and self-reports of motivation: A study of amazon mechanical turk in the US and India. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2925–2934. https://doi.org/10.1145/2207676.2208699

Arai, T., Kato, R., & Fujita, M. (2010). Assessment of operator stress induced by robot collaboration in assembly. *CIRP Annals*, *59*(1), 5–8. https://doi.org/10.1016/j.cirp.2010.03.043

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, *1*(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Becchio, C., Manera, V., Sartori, L., Cavallo, A., & Castiello, U. (2012). Grasping intentions: From thought experiments to empirical evidence. *Frontiers in Human Neuroscience*, *6*, 117. https://doi.org/10.3389/fnhum.2012.00117

Bedny, G. Z., Karwowski, W., & Bedny, I. S. (2012). Complexity evaluation of computer-based tasks. *International Journal of Human-Computer Interaction*, *28*(4), 236–257. https://doi.org/10.1080/10447318.2011.581895

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. Com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368. https://doi.org/10.1093/pan/mpr057

Borromeo, R. M., Laurent, T., Toyama, M., & Amer-Yahia, S. (2017). Fairness and transparency in crowdsourcing. *International Conference on Extending Database Technology (EDBT)*. https://doi.org/10.5441/002/edbt.2017.46

Brewer, R., Morris, M. R., & Piper, A. M. (2016). “Why Would Anybody Do This?”: Understanding Older Adults’ Motivations and Challenges in Crowd Work. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2246–2257. https://doi.org/10.1145/2858036.2858198

Briones, E. M., & Benham, G. (2017). An examination of the equivalency of self-report measures obtained from crowdsourced versus undergraduate student samples. *Behavior Research Methods*, *49*(1), 320–334. https://doi.org/10.3758/s13428-016-0710-8

Buchheit, S., Dalton, D. W., Pollard, T. J., & Stinson, S. R. (2018). Crowdsourcing Intelligent Research Participants: A Student versus MTurk Comparison. *Behavioral Research in Accounting*, *31*(2), 93–106. https://doi.org/10.2308/bria-52340

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon’s Mechanical Turk: A new source of inexpensive, yet high-quality data? *In A. E. Kazdin (Ed.), Methodological Issues and Strategies in Clinical Research (p. 133–139). American Psychological Association*. https://doi.org/10.1037/14805-009

Callison-Burch, C. (2014). Crowd-workers: Aggregating information across turkers to help them find higher paying work. *Second AAAI Conference on Human Computation and Crowdsourcing*.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon’s MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156–2160. https://doi.org/10.1016/j.chb.2013.05.009

Castro-González, Á., Admoni, H., & Scassellati, B. (2016). Effects of form and motion on judgments of social robots' animacy, likability, trustworthiness and unpleasantness. *International Journal of Human-Computer Studies*, *90*, 27–38. https://doi.org/10.1016/j.ijhcs.2016.02.004

Chandler, D., & Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, *90*, 123–133. https://doi.org/10.1016/j.jebo.2013.03.003

Charness, G., Karni, E., & Levin, D. (2010). On the conjunction fallacy in probability judgment: New experimental evidence regarding Linda. *Games and Economic Behavior*, *68*(2), 551–556. https://doi.org/10.1016/j.geb.2009.09.003

d'Eon, G., Goh, J., Larson, K., & Law, E. (2019). Paying Crowd Workers for Collaborative Work. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–24. https://doi.org/10.1145/3359227

Dandurand, F., Shultz, T. R., & Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods*, *40*(2), 428–434. https://doi.org/10.3758/BRM.40.2.428

Difallah, D., Filatova, E., & Ipeirotis, P. (2018). Demographics and dynamics of mechanical Turk workers. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 135–143. https://doi.org/10.1145/3159652.3159661

Dragan, A. D., Bauman, S., Forlizzi, J., & Srinivasa, S. S. (2015). Effects of robot motion on human-robot collaboration. *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 51–58. http://doi.org/10.1145/2696454.2696473.

Dynamo. (2014). *Guidelines for Academic Requesters (Version 1.1)*. https://irb.northwestern.edu/docs/guidelinesforacademicrequesters-1.pdf

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864. https://doi.org/10.1037/0033-295X.114.4.864

Flash, T., & Hogan, N. (1985). The coordination of arm movements: An experimentally confirmed mathematical model. *Journal of Neuroscience*, *5*(7), 1688–1703. https://doi.org/10.1523/JNEUROSCI.05-07-01688.1985

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Glasauer, S., Huber, M., Basili, P., Knoll, A., & Brandt, T. (2010). Interacting in time and space: Investigating human-human and human-robot joint action. *19th International Symposium in Robot and Human Interactive Communication*, 252–257. https://doi.org/10.1109/ROMAN.2010.5598638

Gulletta, G., Erlhagen, W., & Bicho, E. (2020). Human-like arm motion generation: A Review. *Robotics*, *9*(4), 102. https://doi.org/10.3390/robotics9040102

Hamby, T., & Taylor, W. (2016). Survey satisficing inflates reliability and validity measures: An experimental comparison of college and Amazon Mechanical Turk samples. *Educational and Psychological Measurement*, *76*(6), 912–932. https://doi.org/10.1177/0013164415627349

Hanrahan, B. V., Martin, D., Willamowski, J., & Carroll, J. M. (2019). Investigating the Amazon Mechanical Turk market through tool design. *Computer Supported Cooperative Work (CSCW)*, *28*(5), 795–814. https://doi.org/10.1007/s10606-018-9312-6

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., & Bigham, J. P. (2018). A data-driven analysis of workers' earnings on Amazon Mechanical Turk. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3174023

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29–29. https://doi.org/10.1038/466029a

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, *14*(3), 399–425. https://doi.org/10.1007/s10683-011-9273-9

Ikeda, K., & Bernstein, M. S. (2016). Pay it backward: Per-task payments on crowdsourcing platforms reduce productivity. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4111–4121. https://doi.org/10.1145/2858036.2858327

Irani, L. C., & Silberman, M. S. (2013). Turkopticon: Interrupting worker invisibility in amazon mechanical turk. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 611–620. https://doi.org/10.1145/2470654.2470742

Kaufmann, N., Schulze, T., & Veit, D. (2011). More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk. *Amcis*, *11*(2011), 1–11.

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of Advertising*, *46*(1), 141–155. https://doi.org/10.1080/00913367.2016.1269304

Keith, M. G., Stevenor, B. A., & McAbee, S. T. (2022). Scale mean and variance differences in MTurk and non-MTurk samples: A meta-analysis. *Journal of Personnel Psychology*. https://doi.org/10.1027/1866-5888/a000309

Koppenborg, M., Nickel, P., Naber, B., Lungfiel, A., & Huelke, M. (2017). Effects of movement speed and predictability in human–robot collaboration. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *27*(4), 197–209. https://doi.org/10.1002/hfm.20703

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological research online: Report of Board of Scientific Affairs' Advisory Group on the Conduct of Research on the Internet. *American Psychologist*, *59*(2), 105. https://doi.org/10.1037/0003-066X.59.2.105

Kuffner, J. J., & LaValle, S. M. (2000). RRT-connect: An efficient approach to single-query path planning. *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*, *2*, 995–1001. https://doi.org/10.1109/ROBOT.2000.844730

Kulić, D., & Croft, E. (2007). Physiological and subjective responses to articulated robot motion. *Robotica*, *25*(1), 13–27. https://doi.org/10.1017/S0263574706002955

Lacquaniti, F., Terzuolo, C., & Viviani, P. (1983). The law relating the kinematic and figural aspects of drawing movements. *Acta Psychologica*, *54*(1), 115–130. https://doi.org/10.1016/0001-6918(83)90027-6

Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology*, *8*(2), 142–164. https://doi.org/10.1017/iop.2015.13

Lazear, E. P. (2000). Performance pay and productivity. *American Economic Review*, *90*(5), 1346–1361. https://doi.org/10.1257/aer.90.5.1346

Lewis, I., Watson, B., & White, K. M. (2009). Internet versus paper-and-pencil survey methods in psychological experiments: Equivalence testing of participant responses to health-related messages. *Australian Journal of Psychology*, *61*(2), 107–116. https://doi.org/10.1080/00049530802105865

Litman, L., Robinson, J., & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US-and India-based workers on Mechanical Turk. *Behavior Research Methods*, *47*(2), 519–528. https://doi.org/10.3758/s13428-014-0483-x

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

Martin, D., Hanrahan, B. V., O'Neill, J., & Gupta, N. (2014). Being a turker. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 224–235. https://doi.org/10.1145/2531602.2531663

Mason, W., & Watts, D. J. (2009). Financial incentives and the "performance of crowds". *Proceedings of the ACM SIGKDD Workshop on Human Computation*, 77–85. https://doi.org/10.1145/1600150.1600175

Moss, A., & Litman, L. (2018). After the bot scare: Understanding what's been happening with data collection on MTurk and how to stop it. Retrieved February, 4, 2019. Retrieved from https://www.cloudresearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it/

National Commission for the Protection of Human Subjects of Biomedical - Behavioral Research, United States. (1978). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research* (Vol. 2). Department of Health, Education, and Welfare, National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. Retrieved from https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Oppenlaender, J., Milland, K., Visuri, A., Ipeirotis, P., & Hosio, S. (2020). Creativity on Paid Crowdsourcing Platforms. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376677

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*(5), 411–419. https://doi.org/10.1017/S1930297500002205

Pezzulo, G., Donnarumma, F., Dindo, H., D'Ausilio, A., Konvalinka, I., & Castelfranchi, C. (2019). The body talks: Sensorimotor communication and its brain and kinematic signatures. *Physics of Life Reviews*, *28*, 1–21. https://doi.org/10.1016/j.plrev.2018.06.014

Raven, J. (2003). Raven progressive matrices. In *Handbook of nonverbal assessment* (pp. 223–237). Springer, Boston, MA. https://doi.org/10.1007/978-1-4615-0153-4_11

Rubagotti, M., Tusseyeva, I., Baltabayeva, S., Summers, D., & Sandygulova, A. (2022). Perceived safety in physical human–robot interaction–A survey. *Robotics and Autonomous Systems*, *151*, 104047. https://doi.org/10.1016/j.robot.2022.104047

Sacheli, L. M., Tidoni, E., Pavone, E. F., Aglioti, S. M., & Candidi, M. (2013). Kinematics fingerprints of leader and follower role-taking during cooperative joint actions. *Experimental Brain Research*, *226*(4), 473–486. https://doi.org/10.1007/s00221-013-3459-7

Salehi, N., Irani, L. C., Bernstein, M. S., Alkhatib, A., Ogbe, E., & Milland, K. (2015). We are dynamo: Overcoming stalling and friction in collective action for crowd workers. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1621–1630. https://doi.org/10.1145/2702123.2702508

Sandel, M. J. (2009). *Justice: What's the Right Thing to Do*. Farrar, Straus and Giroux. (ISBN 13: 978-0374180652)

Schwartz, A. B. (2016). Movement: How the Brain Communicates with the World. *Cell*, *164*(6), 1122–1135. https://doi.org/10.1016/j.cell.2016.02.038

Scott, S. H. (2004). Optimal feedback control and the neural basis of volitional motor control. *Nature Reviews Neuroscience*, *5*(7), 532–545. https://doi.org/10.1038/nrn1427

Sengupta, T., Bedny, I., & Bedny, G. (2011). Microgenetic Principles in the Study of Computer-Based Tasks. *Human-Computer Interaction and Operators' Performance. Optimizing Work Design with Activity Theory*, 117–148. (ISBN 13: 978-0429105968)

Shaw, A. D., Horton, J. J., & Chen, D. L. (2011). Designing incentives for inexpert human raters. *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, 275–284. https://doi.org/10.1145/1958824.1958865

Smith, S. M., Roster, C. A., Golden, L. L., & Albaum, G. S. (2016). A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples. *Journal of Business Research*, *69*(8), 3139–3148. https://doi.org/10.1016/j.jbusres.2015.12.002

Soratana, T., Liu, Y., & Yang, X. J. (2022). Effect of Payment Methods in Crowdsourcing Platforms. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2022*. https://doi.org/10.1177/1071181322661135

Stulp, F., Grizou, J., Busch, B., & Lopes, M. (2015). Facilitating intention prediction for humans by optimizing robot motions. *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1249–1255. https://doi.org/10.1109/IROS.2015.7353529

Takayama, L., Dooley, D., & Ju, W. (2011). Expressing thought: Improving robot readability with animation principles. *Proceedings of the 6th International Conference on Human-Robot Interaction*, 69–76. https://doi.org/10.1145/1957656.1957674

Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, *77*, 184–197. https://doi.org/10.1016/j.chb.2017.08.038

Todorov, E., & Jordan, M. I. (1998). Smoothness maximization along a predefined path accurately predicts the speed profiles of complex arm movements. *Journal of Neurophysiology*, *80*(2), 696–714. https://doi.org/10.1152/jn.1998.80.2.696

Tremoulet, P. D., & Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception*, *29*(8), 943–951. https://doi.org/10.1068/p3101

Uno, Y., Kawato, M., & Suzuki, R. (1989). Formation and control of optimal trajectory in human multijoint arm movement. *Biological Cybernetics*, *61*(2), 89–101. https://doi.org/10.1007/BF00204593

Vesper, C., Van Der Wel, R. P., Knoblich, G., & Sebanz, N. (2011). Making oneself predictable: Reduced temporal variability facilitates joint action coordination. *Experimental Brain Research*, *211*(3), 517–530. https://doi.org/10.1007/s00221-011-2706-z

Viviani, P., & Flash, T. (1995). Minimum-jerk, two-thirds power law, and isochrony: Converging approaches to movement planning. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(1), 32. https://doi.org/10.1037/0096-1523.21.1.32

Weigold, A., & Weigold, I. K. (2022). Traditional and modern convenience samples: An investigation of college student, Mechanical Turk, and Mechanical Turk college student samples. *Social Science Computer Review*, *40*(5), 1302–1322. https://doi.org/10.1177/08944393211006847

Whiting, M. E., Hugh, G., & Bernstein, M. S. (2019). Fair Work: Crowd Work Minimum Wage with One Line of Code. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*(1), 197–206. https://doi.org/10.1609/hcomp.v7i1.5283

Wu, H., Corney, J., & Grant, M. (2014). Relationship between quality and payment in crowdsourced design. *Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 499–504. https://doi.org/10.1109/CSCWD.2014.6846895

Yin, M., Chen, Y., & Sun, Y.-A. (2013). The effects of performance-contingent financial incentives in online labor markets. *Twenty-Seventh AAAI Conference on Artificial Intelligence*. https://doi.org/10.5555/2891460.2891626

Zhou, T., & Wachs, J. P. (2018). Early prediction for physical human robot collaboration in the operating room. *Autonomous Robots*, *42*(5), 977–995. https://doi.org/10.1007/s10514-017-9670-9

Chapter 3
Effects of Optimal Feedback Control Laws on Robot's Movement Predictability and Human-Likeness

3.1 Introduction

This chapter discusses three studies conducted to investigate movement features in the robotic arm that enable human observers to predict the robot's intention quicker, more accurately, and perceive the robot as more human-like or safer. Three studies were conducted in which participants observed the robot's movement trajectories, predicted the robot's intention in reach movements and rated the robot's movement quality based on the subjective measures of human-likeness and safety. These three studies differ in either the number of targets or path planning constraints.

3.2 Study 1: Functional Trajectories with Rapidly-exploring Random Trees (RRT) Methods - Reaching to 3 Targets

The goal of Study 1 is to investigate the effects of trajectory characteristics on the prediction of robot's intention, its perceived human-likeness, and its perceived safety.

3.2.1 Methods

Participants

24 adults participated in the study (11 female, 12 male, and one prefer not to answer). Mean age = 24.14 years old, *SD* = 4.39. They are undergraduate and graduate students in Engineering at the University of Michigan. Participants were recruited via departmental mailing lists or bi-weekly college of engineering announcements.

Trajectory Generation

The trajectories was generated using RRT (Kuffner & LaValle, 2000) to avoid collision with the shelf's frame and the objects. All trajectories generated were given a fixed set of joint angles c_i for the initial pose, hence they all have the same initial position p_{init} and orientation q_{init}. Each target point m has its unique position p_m and orientation q_m. To generate different trajectories from the starting point to the target's location, RRT solver ($c_g = RRT\ (c_i, p_m, q_m, c_s)$ *Equation 3-1*) output a sequence of a set of joint angles c_g that could reach the specified position p_m and orientation q_m, given the joint position seed c_s. Joint position seed c_s was a set of joint angles given the RRT solver as the initial value to start the solver toward feasible solutions. By randomizing the joint position seed c_s, RRT solver can generate unique solutions, i.e., different c_g, to the same target position and orientation.

$$c_g = RRT\ (c_i, p_m, q_m, c_s) \qquad \textit{Equation 3-1}$$

A trajectory was generated by moving from the joint angle of the fixed initial pose c_i to the pose c_g generated by RRT solver without colliding with the obstacle in the scene. To compute the trajectory characteristics offline, robot's state was broadcasted into Robotic Operation System (ROS) topics during a trajectory execution. These ROS topics contained information about Sawyer's j^{th} joint position (denoted as θ_j), j^{th} joint angular velocity (denoted as ω_j), effort applied in the j^{th} joint (denoted as τ_j), end-effector coordinate in 3D Cartesian space (denoted as $\vec{p}$), and end-effector orientation (denoted as $\vec{q}$) along with their respective timestamp. These broadcasts were recorded at 100 Hz. The next subsection outlines the methods and equations that compute trajectory characteristics from the recorded robot's state. This method can put the robotic arm in unnatural poses and configurations, similar to a functional movement described by Dragan et al. (2015).

Equations for the Trajectory Characteristics

Savitzky-Golay smoothing filter (window length of 11 and 4th order smoothing polynomial) was applied to the robot's parameters collected to remove noises. Then, trajectory characteristics related to human movement were computed. Table 3-1 shows the equations of the computed trajectory characteristics.

Table 3-1: Equations of trajectory characteristics.

Trajectory Characteristics	Equations
Duration	T
Energy expenditure	$\int_0^T \sum_{j=1}^{n} \tau\omega \, \delta t$
Sum of torque-change (Uno et al., 1989)	$\frac{1}{2}\int_0^T \sum_{j=1}^{n} \left(\frac{d\tau_J}{\delta t}\right)^2 \delta t$
Joint-space sum of angular jerk-change	$\int_0^T \sum_{j=1}^{n} \left\lVert\frac{d^3\omega_j}{\delta t^3}\right\rVert^2 \delta t$
Joint-space sum of angular acceleration-change	$\int_0^T \sum_{j=1}^{n} \left\lVert\frac{d^2\omega_j}{\delta t^2}\right\rVert^2 \delta t$
Joint-space sum of angular velocity-change	$\int_0^T \sum_{j=1}^{n} \left\lVert\frac{d\omega_j}{\delta t}\right\rVert^2 \delta t$
Joint-space distance	$\int_0^T \sum_{j=1}^{n} \left\lVert\frac{d\theta_j}{\delta t}\right\rVert^2 \delta t$
End-effector space sum of jerk-change (Todorov & Jordan, 1998; Uno et al., 1989)	$\int_0^T \left\lVert\frac{d^4\vec{p}(t)}{\delta t^4}\right\rVert^2 \delta t$
End-effector space sum of acceleration-change (Broquere et al., 2008; Rozo et al., 2016)	$\int_0^T \left\lVert\frac{d^3\vec{p}(t)}{\delta t^3}\right\rVert^2 \delta t$
End-effector space sum of velocity-change (Broquere et al., 2008; Rozo et al., 2016)	$\int_0^T \left\lVert\frac{d^2\vec{p}(t)}{\delta t^2}\right\rVert^2 \delta t$
End-effector space distance	$\int_0^T \left\lVert\frac{d\vec{p}(t)}{\delta t}\right\rVert^2 \delta t$

Dependent Variables

1. Early prediction time (time between participant's final prediction input and the end of trajectory)
2. Subjective evaluation of human-likeness (perceived animacy, perceived anthropomorphism, and perceived safety) using a subset of Godspeed questionnaire (Bartneck et al., 2009) in 5-point scale.

The early prediction time is a measure for prediction efficiency. Longer early prediction time indicates that the participants have more time to plan their own task in anticipation of the robot's target. The prediction time here is defined differently from Busch et al. (2017) and Dragan et al. (2015). They defined the prediction time as the time from the start of the robotic arm's motion to the participant's response. In this chapter, early prediction time measures the amount of time a participant has before the robotic arm completed its reach motion.

Experimental Task

The robot used in this study was Rethink Robotics's Sawyer, which meets international safety requirements for industrial robots and is relatively safe to deploy in a human study. In this study, participants watched Sawyer reaching for an object on a shelf and used a computer-based interface to input their prediction about the robot's target. There were 9 possible object locations on the shelf arranged in three rows and three columns. The object locations were arranged such that they were coplanar.

Before the start of the session, the researcher informed the participant that the robot would reach for an object on the shelf and that their task is to predict the robot's target using the provided interface. The first training session consisted of nine trajectories, one for each object on the shelf. The training session aimed to familiarize the participants with the interface and the task. After the participant has completed the first training session and before the second training session, the participant was instructed as follows: (1) try to answer before the robot reaches the object, (2) answer as accurately and as fast as they can, (3) they can change their answer, and (4) their final input must be the correct label of the robot's end position. The second training session consisted of the same nine trajectories from the first training session, but the order was randomized. Figure 3-1 shows snippets of a training trajectory.

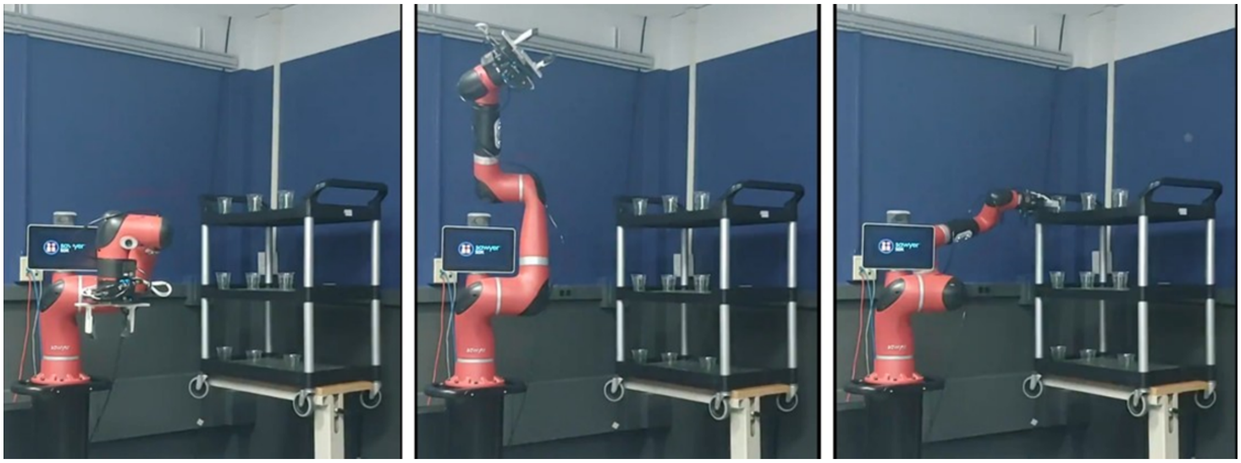


Figure 3-1: Snippets of a training trajectory.

(Left) Initial position, (Middle) during the trajectory, (Right) 'Top Left' position.

In the data collection session, the participant saw only three objects at a time. There were three conditions: Top row, middle row, and bottom row. Each condition has nine trajectories, exactly three trajectories for each object. The order of the conditions was randomized using Latin Square Design to counterbalance the order effect. The objects outside of the condition were removed from the scene (e.g., if the current condition was "top row," items on the middle and bottom rows were removed). For each trajectory observed, the participant performs a prediction task as trained during the training session. Then, they filled in the Godspeed questionnaire (Bartneck et al., 2009) on perceived animacy, anthropomorphism, and safety. In total, each participant observed 27 trajectories.

Statistical Analysis

Linear regression mixed-effect models were conducted with trajectory characteristics as predictors and each subject as the random grouping factor. By specifying the subject as the random grouping factor, this analysis method takes into account participants' individual differences in performance. This analysis used all the valid observations made by participants. For response time, only early predictions (input made before the robot reached the object) were included in the analysis. This is because once the robot stops at the target's position, predicting the robot's target is trivial. For subjective evaluation metrics, all the observations collected were included in the analysis regardless of whether the participant was able to answer before the robot reached the object. The trajectory characteristics were standardized before the analysis.

3.2.2 Results

Early Prediction Time

Table 3-2 shows the statistical analysis results. Participants made a prediction earlier when the trajectory was lower in end-effector distance ($p < 0.001$), lower in energy expenditure ($p < 0.001$), or lower in end-effector space sum of acceleration- change ($p < 0.001$). Participants made a prediction earlier when the trajectory was higher in the sum of torque-change ($p = 0.006$), higher in joint-space sum of jerk-change ($p = 0.001$), higher in end-effector space sum of speed-change ($p < 0.001$), or higher in end-effector space sum of jerk-change ($p < 0.001$).

Perceived Anthropomorphism

Table 3-2 shows the statistical analysis results. The robot scored higher perceived anthropomorphism when the trajectory was relatively lower in end-effector distance ($p = 0.010$), lower in energy expenditure ($p = 0.003$), lower in joint-space distance ($p = 0.023$), or lower in end-effector space sum of acceleration-change ($p = 0.014$). The robot also scored higher perceived anthropomorphism when the trajectory was higher in the joint-space sum of speed-change ($p = 0.033$) or higher in end-effector space sum of speed-change ($p < 0.001$).

Perceived Animacy

Table 3-3 shows the statistical analysis results. The robot scored higher perceived animacy when the trajectory was relatively lower in end-effector distance ($p = 0.036$) or higher in end-effector space sum of speed-change ($p = 0.003$).

Perceived Safety

Table 3-3 shows the statistical analysis results. The robot scored higher perceived safety when the trajectory was relatively lower in energy expenditure ($p = 0.001$), lower in joint-space distance ($p = 0.019$), lower in end-effector space sum of acceleration-change ($p < 0.001$), or higher in end-effector space sum of jerk-change ($p = 0.019$).

Table 3-2: (Study 1) Effects of trajectory characteristics on early prediction time and perceived anthropomorphism.

Predictors	**Early prediction time**			**Perceived anthropomorphism**		
	EC[1]	95% CI	*p*-value[2]	EC[1]	95% CI	*p*-value[2]
End-effector distance	-0.78	-1.15 – -0.42	<0.001 ***	-0.15	-0.26 – -0.04	0.010 *
Energy expenditure	-0.57	-0.78 – -0.36	<0.001 ***	-0.11	-0.18 – -0.04	0.003 **
Duration	-0.13	-1.37 – 1.12	0.844	-0.45	-0.84 – -0.06	0.023 *
Joint-space distance	0.31	0.09 – 0.53	0.006 **	0.02	-0.05 – 0.09	0.518
Sum of torque-change	-0.39	-2.01 – 1.23	0.635	0.57	0.05 – 1.09	0.033 *
Joint-space sum of speed-change	-0.32	-1.03 – 0.40	0.382	-0.06	-0.28 – 0.16	0.591
Joint-space sum of acceleration-change	0.93	0.37 – 1.50	0.001 **	0.06	-0.11 – 0.23	0.483
Joint-space sum of jerk-change	1.32	0.91 – 1.72	<0.001 ***	0.24	0.12 – 0.36	<0.001 ***
End-effector space sum of speed-change	-2.30	-2.98 – -1.62	<0.001 ***	-0.26	-0.47 – -0.05	0.014 *
End-effector space sum of acceleration-change	1.09	0.71 – 1.47	<0.001 ***	0.05	-0.07 – 0.17	0.420
End-effector space sum of jerk-change	-0.78	-1.15 – -0.42	<0.001 ***	-0.15	-0.26 – -0.04	0.010 *

[1] EC: estimate coefficient.

[2] *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, · $p < 0.1$)

Table 3-3: (Study 1) Effects of trajectory characteristics on perceived animacy and perceived safety.

Predictors	**Perceived animacy**			**Perceived safety**		
	EC[1]	95% CI	*p*-value[2]	EC[1]	95% CI	*p*-value[2]
End-effector distance	-0.10	-0.20 – -0.01	0.036 *	0.03	-0.08 – 0.15	0.555
Energy expenditure	-0.04	-0.10 – 0.02	0.152	-0.13	-0.20 – -0.05	0.001 **
Duration	-0.05	-0.37 – 0.27	0.765	-0.47	-0.86 – -0.08	0.019 *
Joint-space distance	0.01	-0.05 – 0.07	0.784	0.00	-0.07 – 0.08	0.923
Sum of torque-change	0.08	-0.35 – 0.51	0.714	0.41	-0.12 – 0.93	0.127
Joint-space sum of speed-change	-0.05	-0.23 – 0.13	0.590	0.07	-0.15 – 0.29	0.507
Joint-space sum of acceleration-change	0.07	-0.07 – 0.22	0.305	0.16	-0.02 – 0.33	0.076 .
Joint-space sum of jerk-change	0.16	0.05 – 0.26	0.003 **	0.02	-0.11 – 0.14	0.773
End-effector space sum of speed-change	-0.16	-0.33 – 0.01	0.065 .	-0.42	-0.63 – -0.21	<0.001 ***
End-effector space sum of acceleration-change	0.06	-0.04 – 0.16	0.260	0.15	0.02 – 0.27	0.019 *
End-effector space sum of jerk-change	-0.10	-0.20 – -0.01	0.036 *	0.03	-0.08 – 0.15	0.555

[1] EC: estimate coefficient.

[2] *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, · $p < 0.1$)

3.3 Study 2: Functional Trajectories with RRT Methods - Reaching to 9 Targets

Study 2 investigated whether the effects of trajectory characteristics in Study 1 were limited by the number of possible targets or not. Study 2 was conducted in the scenario where the participants must predict the robot's target when there are 9 possible targets. This point was listed as one of the limitations in a study by Dragan et al. (2013), in which 2 objects were presented in the scene.

3.3.1 Methods

Participants

24 adults participated in the study (9 female, 15 male). Mean age = 22.25 years old, *SD* = 2.80. They are undergraduate and graduate students in Engineering at the University of Michigan. Participants were recruited via departmental mailing lists or bi-weekly college of engineering announcements.

Experimental Task

Study 2 used the same trajectories as Study 1, both in the training and in the data collection session. The difference is that instead of displaying three objects at a time in Study 1, all nine objects were displayed during the data collection session in Study 2. All 27 trajectories were performed while nine objects were on the shelf. The robot can reach any of the 9 objects in a randomized order (within-subject design). Study 2 focused on a scenario where the objects are arranged in 3x3 grids rather than three objects in a row for Study 1.

Statistical Analysis

The results were analyzed using the same analysis statistical methods outlined in Subsection 3.2.1.

3.3.2 Results

Table 3-4 shows the statistical analysis results. Participants made a prediction earlier when the trajectory was lower in end-effector distance ($p = 0.001$), lower in energy expenditure ($p = 0.001$), lower in joint-space sum of speed-change ($p = 0.038$), or lower in end-effector space sum of acceleration-change ($p = 0.002$). The participants made a prediction earlier when the trajectory was higher in joint-space sum of acceleration-change ($p < 0.001$), end-effector space sum of speed-change ($p = 0.014$), or end-effector space sum of jerk-change ($p < 0.001$).

Perceived Anthropomorphism

Table 3-4 shows the statistical analysis results. The robot scored higher on perceived anthropomorphism when the trajectory was relatively higher in the end-effector space sum of jerk-change ($p = 0.040$).

Table 3-4: (Study 2) Effects of trajectory characteristics on early prediction time and perceived anthropomorphism.

Predictors	**Early prediction time**			**Perceived anthropomorphism**		
	EC[1]	95% CI	*p*-value[2]	EC[1]	95% CI	*p*-value[2]
End-effector distance	-0.46	-0.73 – -0.19	0.001 **	-0.03	-0.15 – 0.10	0.676
Energy expenditure	-0.27	-0.43 – -0.10	0.001 **	0.02	-0.05 – 0.10	0.556
Duration	0.04	-0.88 – 0.96	0.932	0.21	-0.22 – 0.64	0.334
Joint-space distance	-0.11	-0.27 – 0.06	0.199	0.05	-0.03 – 0.13	0.228
Sum of torque-change	-1.29	-2.51 – -0.07	0.038 *	-0.51	-1.08 – 0.06	0.077 .
Joint-space sum of speed-change	1.33	0.81 – 1.84	<0.001 ***	0.13	-0.11 – 0.37	0.282
Joint-space sum of acceleration-change	-0.22	-0.62 – 0.17	0.261	0.03	-0.16 – 0.22	0.757
Joint-space sum of jerk-change	0.37	0.08 – 0.67	0.014 *	-0.03	-0.17 – 0.10	0.637
End-effector space sum of speed-change	-0.77	-1.25 – -0.29	0.002 **	-0.21	-0.44 – 0.02	0.073 .
End-effector space sum of acceleration-change	0.59	0.31 – 0.87	<0.001 ***	0.14	0.01 – 0.27	0.040 *
End-effector space sum of jerk-change	-0.46	-0.73 – -0.19	0.001 **	-0.03	-0.15 – 0.10	0.676

[1] EC: estimate coefficient.

[2] *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, · $p < 0.1$)

Table 3-5: (Study 2) Effects of trajectory characteristics on perceived animacy and perceived safety.

Predictors	Perceived animacy			Perceived safety		
	EC[1]	95% CI	*p*-value[2]	EC[1]	95% CI	*p*-value[2]
End-effector distance	-0.08	-0.18 – 0.02	0.101	0.17	0.06 – 0.28	0.003 **
Energy expenditure	0.01	-0.05 – 0.07	0.662	-0.01	-0.08 – 0.06	0.843
Duration	0.17	-0.16 – 0.51	0.308	-0.14	-0.51 – 0.24	0.475
Joint-space distance	0.02	-0.04 – 0.08	0.546	-0.02	-0.09 – 0.05	0.555
Sum of torque-change	-0.29	-0.74 – 0.16	0.205	-0.21	-0.71 – 0.29	0.417
Joint-space sum of speed-change	0.04	-0.15 – 0.23	0.686	0.31	0.10 – 0.52	0.004 **
Joint-space sum of acceleration-change	0.03	-0.12 – 0.17	0.704	-0.11	-0.27 – 0.06	0.200
Joint-space sum of jerk-change	0.09	-0.02 – 0.19	0.096 .	-0.19	-0.30 – -0.07	0.002 **
End-effector space sum of speed-change	-0.08	-0.26 – 0.09	0.351	-0.20	-0.40 – 0.00	0.054 .
End-effector space sum of acceleration-change	0.07	-0.03 – 0.18	0.165	0.12	-0.00 – 0.24	0.051 .
End-effector space sum of jerk-change	-0.08	-0.18 – 0.02	0.101	0.17	0.06 – 0.28	0.003 **

[1] EC: estimate coefficient.

[2] *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, · $p < 0.1$)

Perceived Animacy

Table 3-5 shows the statistical analysis results. The effects of trajectory characteristics on perceived animacy were not significant at $\alpha = 0.05$.

Perceived Safety

Table 3-5 shows the statistical analysis results. The robot scored higher perceived safety when the trajectory was relatively lower in end-effector space sum of speed-change ($p = 0.002$), higher in end-effector distance ($p = 0.003$), or higher in joint-space sum of acceleration-change ($p = 0.004$).

3.3.3 Comparison with Study 1, which findings are coherent? Which are not?

Some findings on response time were coherent with Study 1. Trajectories with shorter hand paths or energy expenditure in the hand trajectories were predicted earlier compared to others. Similarly, Study 1 and 2 found that trajectories with higher hand speed were perceived as more animated, and lower hand acceleration was perceived as more human-like. The robot was perceived as safer when it performed a trajectory with lower acceleration and higher jerk in the hand trajectories. The response time of early prediction was not significantly different between Studies 1 and 2 (t = -0.79, *p*-value = 0.43).

3.4 Study 3: Functional Trajectories with Inverse Kinematics Methods - Reaching to 9 Targets

The goal of Study 3 is to investigate the effects of trajectory characteristics under the constraint that the robot does not change its movement directions. The trajectories in Studies 1 and 2 can contain multiple changes of direction (generated by RRT), which may make the trajectory difficult to predict (Dragan et al., 2015). Dragan et al. (2015) described similarly in their studies that the functional trajectories generated using RRT were trickier to predict. In this condition, their participants reported that the robot tried to trick them. This is because the trajectories generated with RRT can change their direction right in front of other possible targets, which may lead the observers to change their prediction. Hence, this study introduced a different set of trajectories, generated with a different constraint and method. The trajectories in Study 3 were designed such that the end-effector reached the target's position in one curved path without sudden changing in direction. Study 3 showed all nine targets during the data collection, in the same way as Study 2. The robot can reach any of the nine targets in a randomized order (within-subject design). This Study has the same level of uncertainty in targets as Study 2, which is more than Study 1.

3.4.1 Methods

Participants

20 adults participated in the study (8 female, 12 male). Mean age = 24.25 years old, *SD* = 3.02. They are undergraduate and graduate students in Engineering at the University of Michigan. Participants were recruited via departmental mailing lists or bi-weekly college of engineering announcements.

Trajectory Generation

Inverse Kinematics (IK) generated the trajectories for the Study 3 instead of RRT. Similar to the RRT planner, IK planner may put the robotic arm in unnatural poses and configurations (Dragan et al., 2015). However, the trajectories were constrained to not have sudden changes in direction, as appeared in Studies 1 and 2. The trajectory was not planned with regard to the obstacle in the scene. During the trajectory generation, trajectories that would collide with the shelf were removed from the experiment.

Statistical Analysis

The results were analyzed using the same analysis statistical methods outlined in Subsection 3.2.1.

3.4.2 Results

Early Prediction Time

Table 3-6 shows the statistical analysis results. Participants made a prediction earlier when the trajectory was lower in end-effector space sum of jerk-change (p = 0.005) or higher in end-effector space sum of acceleration-change (p = 0.036).

Table 3-6: (Study 3) Effects of trajectory characteristics on early prediction time and perceived anthropomorphism.

Predictors	**Early prediction time**			**Perceived anthropomorphism**		
	EC[1]	95% CI	*p*-value[2]	EC[1]	95% CI	*p*-value[2]
End-effector distance	-0.14	-0.38 – 0.10	0.242	-0.1	-0.27 – 0.07	0.258
Energy expenditure	-0.06	-0.16 – 0.04	0.251	-0.02	-0.09 – 0.05	0.597
Duration	-0.42	-1.09 – 0.25	0.214	-0.19	-0.68 – 0.29	0.434
Joint-space distance	-0.44	-1.14 – 0.27	0.222	-0.19	-0.71 – 0.32	0.462
Sum of torque-change	0.00	-0.14 – 0.13	0.957	0.08	-0.02 – 0.18	0.104
Joint-space sum of speed-change	0.65	-0.51 – 1.81	0.270	0.44	-0.41 – 1.28	0.315
Joint-space sum of acceleration-change	-0.07	-0.41 – 0.26	0.667	-0.23	-0.48 – 0.02	0.071 .
Joint-space sum of jerk-change	-0.05	-0.29 – 0.19	0.688	0.25	0.07 – 0.43	0.005 **
End-effector space sum of speed-change	0.25	-0.01 – 0.51	0.060 .	0.02	-0.18 – 0.22	0.826
End-effector space sum of acceleration-change	0.17	0.01 – 0.34	0.036 *	-0.02	-0.14 – 0.09	0.680
End-effector space sum of jerk-change	-0.26	-0.44 – -0.08	0.005 **	0.02	-0.11 – 0.15	0.807

[1] EC: estimate coefficient.

[2] *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, · $p < 0.1$)

Perceived Anthropomorphism

Table 3-6 shows the statistical analysis results. The robot scored higher on perceived anthropomorphism when the trajectory was relatively higher in joint-space sum of jerk-change (*p* = 0.005).

Perceived Animacy

Table 3-7 shows the statistical analysis results. The robot scored higher perceived animacy when the trajectory was relatively higher in sum of torque-change (*p* = 0.042) or higher in joint-space sum of jerk-change (*p* = 0.006).

Perceived Safety

Table 3-7 shows the statistical analysis results. The robot scored higher perceived safety when the trajectory was relatively lower in end-effector space sum of acceleration-change (*p* = 0.006), lower in end-effector distance (*p* = 0.013), or higher in joint-space sum of jerk-change (*p* = 0.039).

Table 3-7: (Study 3) Effects of trajectory characteristics on perceived animacy and perceived safety.

Predictors	Perceived animacy			Perceived safety		
	EC[1]	95% CI	*p*-value[2]	EC[1]	95% CI	*p*-value[2]
End-effector distance	0.04	-0.11 – 0.19	0.639	-0.19	-0.35 – -0.04	0.013 *
Energy expenditure	0.01	-0.05 – 0.07	0.773	0.01	-0.05 – 0.08	0.746
Duration	0.01	-0.41 – 0.44	0.951	-0.32	-0.75 – 0.11	0.145
Joint-space distance	-0.09	-0.54 – 0.36	0.696	-0.29	-0.75 – 0.16	0.207
Sum of torque-change	0.09	0.00 – 0.17	0.042 *	0.08	-0.01 – 0.16	0.080 .
Joint-space sum of speed-change	0.17	-0.57 – 0.92	0.652	0.63	-0.12 – 1.38	0.102
Joint-space sum of acceleration-change	-0.16	-0.37 – 0.06	0.162	-0.21	-0.43 – 0.01	0.056 .
Joint-space sum of jerk-change	0.22	0.06 – 0.37	0.006 **	0.16	0.01 – 0.32	0.039 *
End-effector space sum of speed-change	-0.08	-0.26 – 0.09	0.352	0.06	-0.11 – 0.24	0.474
End-effector space sum of acceleration-change	0.02	-0.08 – 0.13	0.667	-0.15	-0.25 – -0.04	0.006 **
End-effector space sum of jerk-change	0.00	-0.12 – 0.11	0.971	0.06	-0.05 – 0.18	0.288

[1] EC: estimate coefficient.

[2] *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, · $p < 0.1$)

3.4.3 Comparison with Studies 1 and 2, which findings are coherent? Which are not?

The findings on response time Studies 1 and 2 contradict that of Study 3. In Studies 1 and 2, the lower acceleration-change and higher jerk-change in end-effector space are correlated with participants being able to make a correct prediction earlier. However, results of Study 3 found the opposite findings, the characteristics mentioned above would result in participants making a prediction later.

None of the findings on perceived anthropomorphism and animacy are coherent across Studies 1-3. The trajectory characteristics found to be significant in Study 1 and 2 were not found to be significant in Study 3. For a coherent finding, Studies 1-3 found that lower end-effector space sum of acceleration-change improved perceived safety.

3.5 Discussion

Studies 1 and 2 have many coherent findings on the effect of trajectory characteristics on the response time. The observer responded quicker when the trajectory is lower in energy expenditure and distance, regardless of whether the observer was looking for 3 or 9 possible objects at a time.

The response time was also affected by the characteristics observable in the end-effector. The observers also responded quicker when the trajectory is lower in the sum of acceleration-change and higher in the sum of jerk-change. A trajectory with a lower sum of acceleration change produces a smoother trajectory of the hand (Flash & Hogan, 1985). A trajectory with a higher sum of jerk-change may have changes in the direction that is more distinct. This explanation supported the findings from the related literature using models that optimize the prediction time and accuracy through trial and error. The kinematic model that optimizes for quicker prediction and higher accuracy may add other components outside of minimizing jerk (Stulp et al., 2015) or the sum of squared velocity over time (Dragan et al., 2015) to the kinematic model. Findings on the early prediction time of Study 3 are mostly incoherent with Studies 1 and 2. The participants in Study 3 responded quicker when the end-effector is higher in the sum of acceleration-change and the lower sum of jerk-change. These findings are in the opposite direction compared to that of Studies 1 and 2. These results demonstrated that the participants' prediction speed was affected by the characteristics of the end-effector. For both Studies 1 and 3, the participants also responded quicker to a trajectory with a higher end-effector space sum of speed-change (Study 1: $p < 0.001$, Study 3: $p = 0.060$).

For perceived anthropomorphism and perceived animacy, findings between Study 3 and Studies 1 and 2 were not coherent. In Studies 1 and 2, a trajectory with a lower end-effector sum of acceleration-change is correlated with a higher degree of perceived anthropomorphism (Study 1: $p = 0.014$, Study 2: $p = 0.073$). For perceived animacy, a trajectory with a higher end-effector sum of speed-change is correlated with a higher degree of perceived animacy (Study 1: $p = 0.003$, Study 2: $p = 0.096$). This finding is coherent with that of Tremoulet and Feldman (2000). When an object has larger changes in speed, it is perceived as more alive than those with constant velocity. Smoothness of the trajectory also did not significantly affect the perceived animacy, same as the findings of Castro-Gonza´lez et al. (2016). For Study 3, a trajectory with a higher joint-space sum of jerk-change scored higher in perceived anthropomorphism ($p = 0.005$) and animacy ($p = 0.006$). A possible explanation is that the differences in perceived animacy between Studies 1 and 2 and Study 3 is due to the presence of changes in movement directions. Tremoulet and Feldman (2000) found that when an object has a larger angle of changes in movement direction, it is perceived as more alive than those with a smaller angle of changes in movement direction. Because none of the trajectories in Study 3 have sudden changes in movement directions, the participants may have rated a trajectory with jerkiness as more lifelike.

For perceived safety, participants perceived the robot as safer when the trajectory is lower in the sum of acceleration-change in end-effector space (Study 1: $p < 0.001$, Study 2: $p = 0.054$, Study 3: 0.006). In Studies 1 and 2, the robot scored higher in perceived safety when it performed a trajectory that was higher in the sum of jerk-change in the end-effector space (Study 1: $p = 0.019$, Study 2: $p = 0.051$). In Studies 1 and 3, the participants perceived the robot as safer when it performed a trajectory with the higher joint-space sum of jerk-change (Study 1: $p = 0.076$, Study 3: $p = 0.039$). Only Study 2 showed that a lower speed in the end-effector increased perceived safety. Kulic and Croft (2007) argued similarly that robot's speed may impact the observer's anxiety or surprise, which is part of the perceived safety score.

The result of the studies presents in this study agree with the insight from related literature (Dragan et al., 2015; Stulp et al., 2015) that other features, outside of efficiency, may help an observer to make a correct prediction quicker. However, this work does not provide the exact example of a movement that can achieve quicker prediction and higher accuracy. One suggestion for future research is to perform a time-series analysis and look for features that occur right before the observer makes a prediction or changes from one prediction to another. This will tell us what are the features of the movement that an observer takes as a clue to infer the robot's intention. There is also a limited number of public datasets related to this specific topic of human interpreting robot's target. The authors would like to suggest that the dataset related to this topic should be made available so the comparison between different literature can be studied further.

3.6 Summary

This chapter discusses three studies that investigate the following research question: How do human-like optimal feedback control laws in robotic arms affect (1) human performance in predicting robot's movement intention and (2) perceived human-likeness of the robot? This chapter discusses the experimental design, procedure, trajectory generation, method of analysis, and the results of these three studies.

3.7 References

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, *1*(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Broquere, X., Sidobre, D., & Herrera-Aguilar, I. (2008). Soft motion trajectory planner for service manipulator robot. *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2808–2813. https://doi.org/10.1109/IROS.2008.4650608

Busch, B., Grizou, J., Lopes, M., & Stulp, F. (2017). Learning legible motion from human–robot interactions. *International Journal of Social Robotics*, *9*(5), 765–779. https://doi.org/10.1007/s12369-017-0400-4

Castro-González, Á., Admoni, H., & Scassellati, B. (2016). Effects of form and motion on judgments of social robots' animacy, likability, trustworthiness and unpleasantness. *International Journal of Human-Computer Studies*, *90*, 27–38. https://doi.org/10.1016/j.ijhcs.2016.02.004

Dragan, A. D., Bauman, S., Forlizzi, J., & Srinivasa, S. S. (2015). Effects of robot motion on human-robot collaboration. *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 51–58. http://doi.org/10.1145/2696454.2696473.

Dragan, A. D., Lee, K. C. T., & Srinivasa, S. S. (2013). Legibility and predictability of robot motion. *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 301–308. https://doi.org/10.1109/HRI.2013.6483603

Flash, T., & Hogan, N. (1985). The coordination of arm movements: An experimentally confirmed mathematical model. *Journal of Neuroscience*, *5*(7), 1688–1703. https://doi.org/10.1523/JNEUROSCI.05-07-01688.1985

Kuffner, J. J., & LaValle, S. M. (2000). RRT-connect: An efficient approach to single-query path planning. *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*, *2*, 995–1001. https://doi.org/10.1109/ROBOT.2000.844730

Kulić, D., & Croft, E. (2007). Physiological and subjective responses to articulated robot motion. *Robotica*, *25*(1), 13–27. https://doi.org/10.1017/S0263574706002955

Rozo, L., Silverio, J., Calinon, S., & Caldwell, D. G. (2016). Learning controllers for reactive and proactive behaviors in human–robot collaboration. *Frontiers in Robotics and AI*, *3*, 30. https://doi.org/10.3389/frobt.2016.00030

Stulp, F., Grizou, J., Busch, B., & Lopes, M. (2015). Facilitating intention prediction for humans by optimizing robot motions. *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1249–1255. https://doi.org/10.1109/IROS.2015.7353529

Todorov, E., & Jordan, M. I. (1998). Smoothness maximization along a predefined path accurately predicts the speed profiles of complex arm movements. *Journal of Neurophysiology*, *80*(2), 696–714. https://doi.org/10.1152/jn.1998.80.2.696

Tremoulet, P. D., & Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception*, *29*(8), 943–951. https://doi.org/10.1068/p3101

Uno, Y., Kawato, M., & Suzuki, R. (1989). Formation and control of optimal trajectory in human multijoint arm movement. *Biological Cybernetics*, *61*(2), 89–101. https://doi.org/10.1007/BF00204593

Chapter 4
Effects of Interaction Formats on Perceived Human-Likeness of the Robot Movement

4.1 Introduction

In this chapter discusses a study investigating the following questions: how do interaction formats affect perceived human-likeness of the robot? The study described in this chapter focuses on the issue of information presentation format by comparing participant's judgment of the robotic movement patterns, by either observing the physical robotic system or watching video recordings of the same robotic movements. Participants watching videos were intended to be a "simulated" online task, but the participant pool was controlled to be the same in both conditions.

The hypothesized results were that there would be no notable differences between observing a robot's physical movement and watching their videos. Kose-Bagci et al. (2009) found that there was no difference between perceived general appearance of a robot between observing a robot's physical movements and watching a real-time image of the robot projected on a wall.

4.2 Methods

4.2.1 Participants

48 adults participated in the study (19 female, 29 male). Mean age = 22.38 years old, *SD* = 2.89. They are undergraduate and graduate students in Engineering at the University of Michigan. Participants were recruited via departmental mailing lists or bi-weekly college of engineering announcements. See Table 4-1 for participants' demographic in each condition.

Table 4-1: Participant demographic information.

Conditions	**N**	**Male (%)**	**Mean age - *SD***
Physical Robot	24	62.50	22.25 - 2.80
Videos	24	58.33	22.32 - 3.30

4.2.2 Design of Experiment

Independent Variable

The only independent variable investigated in this chapter is formats of interaction which have two levels: physical robot and video. The physical robot condition simulates a human working side-by-side with a robot and watching the robot perform reach movements. This condition uses the same trajectory generation method and experimental task as Study 2 outlined in Chapter 3. The task utilized in this study was an HRI judgment task. The participants observed a robot performing a reach movement toward an item in a fixed position.

In the video condition, short video clips of a robot's movements were presented in a Qualtrics survey instead of by a physical robot. The reach movements in the video were recorded from the same perspective that participants in the physical robot condition would sit. The videos were cropped to show only the robot and the region of interest. The video shows a still image of the first frame from the start of the video to t = 1 second, followed by the rest of the movement. All videos have a resolution of 1280×720 pixels with colors but do not have sounds. Videos have a mean length of 10.09 seconds (*SD* = 0.17). The videos do not contain sounds to ensure that participants in this condition experienced the video in the same way, as opposed to letting participants control whether to mute or turn on the sounds.

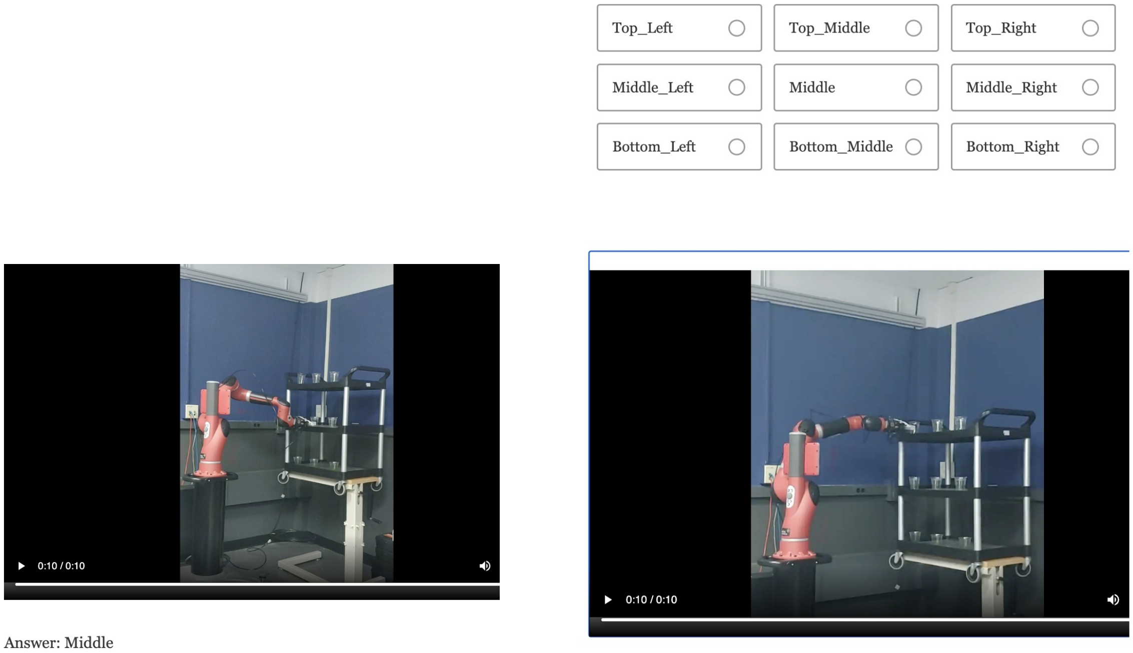
Dependent Variables

Subjective evaluation of human-likeness (perceived animacy, perceived anthropomorphism, and perceived safety) using a subset of Godspeed questionnaire (Bartneck et al., 2009) in 5-point scale.

4.2.3 Experimental Task

The physical robot condition follows the same experimental task procedures described in Chapter 3.2.1, under Experimental Task subsection.

The video condition utilized the same HRI judgment task in the physical robot condition, but they watched short video clips of a robot's movements presented in the Qualtrics survey instead of interacting face-to-face with the actual robot. To familiarize the participants with the task, participants went through a practice session which has two pages. In the first training page, the movement's target was shown in the Qualtrics survey next to the corresponded video (see Figure 4-1). The participants were permitted to rewatch the training videos. The second training page showed the same nine videos in randomized order and asked the participant to use the interface to select the correct label for each video. If they answered incorrectly, the survey system would ask the participant to answer the incorrect ones again until they get all videos labeled correctly. Note the timing of showing the instruction was slightly different between physical robot and video conditions. In the video version, participants saw the instructions for the tasks on the Qualtrics survey after the second training. In the physical robot condition, the researcher provided directions before the second training session.



Answer: Middle

(a) Training: a video with its corresponding label

(b) Main task: a video with an interface to input the robot's target

Figure 4-1: A Qualtrics survey showing (a) a training page and (b) a main task session.

The main experimental session asked the participant to observe the robot's reach movement and judge the target, i.e., the item that the robot was reaching toward, before the robot stopped moving. A total of 27 movements were shown to each participant in a randomized order (within-subject design). After watching each movement, participants filled in 13 five-point questionnaires to evaluate the quality of the robot's movement. These 13 questions were derived from the Godspeed questionnaire (Bartneck et al., 2009), from the section focusing on perceived safety, perceived anthropomorphism, and perceived animacy. This study has been approved by the University of Michigan Institutional Review Board (HUM00188416 and HUM00207720).

4.2.4 Statistical Analysis

Welch's two sample t-tests were conducted to analyze the perceived anthropomorphism, perceived animacy, and perceived safety with formats of interactions as an independent variable.

4.3 Results

For each condition, 24 participants were enrolled. Each participant went through 27 trials. A total of 648 observations were collected for each condition. However, one observation from the physical robot condition was removed because the participants did not input the prediction, hence the observation for this trial was assumed to be invalid. Thus, a total of 647 observations in the physical robot condition and 648 observations in the video condition were included in the analysis.

4.3.1 Perceived Anthropomorphism

Figure 4-2 shows the violin plots for perceived anthropomorphism. Welch's two sample t-test showed that formats of interactions significantly affected perceived anthropomorphism ($t(1283.3) = 5.6744$, $p < 0.001$, 95% CI = [0.206, 0.425]). Participants observing a robot's physical movements (*M* = 2.647*, SD* = 0.901) rated a robot's movement as less human-like than participants watching the same movements in video (*M* = 2.944*, SD* = 0.985). Effect size value (Cohen's d = 0.315) suggested small practical impact.

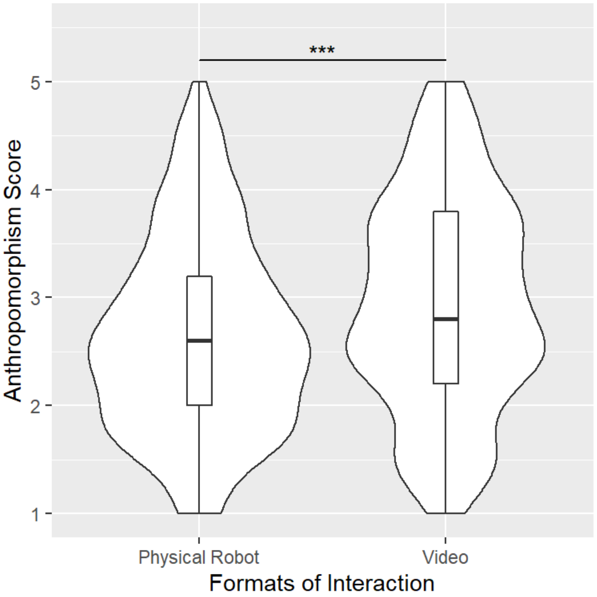


Figure 4-2: Violin plots of perceived anthropomorphism score between physical robot and video conditions.

4.3.2 Perceived Animacy

Figure 4-3 shows the violin plots for perceived animacy. Welch's two sample t-test showed that formats of interactions significantly affected perceived animacy (*t*(1293) = 10.934, *p* < 0.001, 95% CI = [0.496, 0.719]). Participants observing a robot's physical movements (*M* = 2.602*, SD* = 0.859) rated a robot's movement as less life-like than participants watching the same movements in video (*M* = 3.125*, SD* = 0.862). Effect size value (Cohen's d = 0.608) suggested medium practical impact.

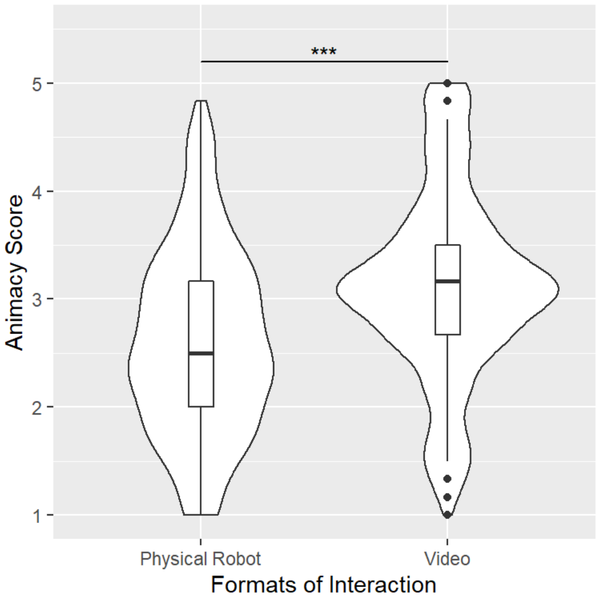


Figure 4-3: Violin plots of perceived animacy score between physical robot and video conditions.

4.3.3 Perceived Safety

Figure 4-4 shows the violin plots for perceived safety. Welch's two sample t-test showed that formats of interactions significantly affected perceived safety (*t*(1249.9) = 4.255, *p* < 0.001, 95% CI = [-0.346, -0.127]). Participants observing a robot's physical movements (*M* = 3.941*, SD* = 0.774) felt safer than participants watching the same movements in video (*M* = 3.738*, SD* = 0.935) during the experiment. Effect size value (Cohen's d = 0.236) suggested small practical impact.

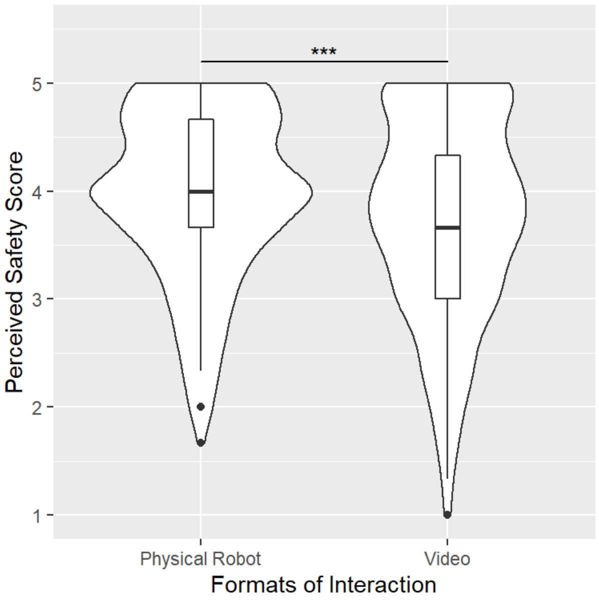


Figure 4-4: Violin plots of perceived safety score between physical robot and video conditions.

4.4 Discussion

This analysis focused on comparing the two formats of interaction: a robot's physical movement and videos. The comparison between participants observing a robot's physical movements and participants watching videos showed that participants would rate the robot face-to-face as less human-like and less life-like compared to watching videos. However, there are several differences between these two formats of interaction which require further investigation to explain which of the factors are affecting the findings. See Table 4-2 for the summary of differences between the two formats of interaction, categorized by sensory systems. Participants who observed the robot's physical movements could visually perceive depth information, as well as higher image resolution and a wider field of view compared to participants who watched videos. Depth information may play a role in the judgment of the robot's target as the experimental task requires predicting the robot's target in 3-dimensional space. The video recordings of the movements took away the depth information that may have been used to differentiate the robot’s target on the left side vs. the right side of the shelf (see Figure 4-1 [b]). Image resolution was also limited to 1280 × 720 pixels in the videos condition compared to a full resolution of visual information perceived by human eyesight in the physical movements condition. Human eyes also provided a wider field of view than videos, which were cropped to show only the robot, the area it passed through during the movement, and the targets on the shelf. Note all participants recruited for these studies have normal or corrected-to-normal vision.

The robot’s movement also generates sound with varying intensity and frequency depending on whether the robot is accelerating, decelerating, or maintaining speed. Similarly, the vibration from the robot’s movement is also transferred to the ground. These changes in sound intensity and frequency and vibration may indicate inflection points (Cabrera Ubaldi, 2018) in the movements. These inflection points provide cues to the participants that the robot is about to change movement direction or stop. The videos do not contain sound to make the condition more comparable to the version completed by online crowdworkers. Online participants may not listen to the robot’s sound or have an audio output device. Thus, audio information was removed from the videos to control potential confounding variables such as the presence of sound. Similarly, vibration from the robot’s movement transferred to the environment can be perceived in the physical movement condition but is not present in the videos. Further study should investigate how the factors mentioned above affect the findings of human’s judgment of robot’s movement quality.

Table 4-2: Summary of the differences between a robot's physical movements and videos condition.

Sensory systems	**Robot's physical movements**	**Videos**
Visual system	Depth information from binocular vision.	No depth information.
	The resolution is as perceived in real-life situation.	Limited resolution of 1280 × 720 pixels.
	Wider field of view.	The field of view limited in the video recordings as shown in Figure 4-1 (b). The videos were cropped to show only the robot and targets on the shelf.
Auditory system	The robot's movement generates noises.	No sounds. Auditory information was removed in the video condition.
Somatosensory system	Vibration from robot's motion.	No vibration.

Participants also rated their emotional state as feeling safer when they interacted face-to-face with the robotic system rather than through videos. A possible explanation is that participants may feel more positively toward a physical robot due to higher excitement and sympathy compared to the video recordings (Bainbridge et al., 2011). This finding may also come from the sense of safety that participants in the physical robot condition sat outside the robot's workspace, but the participants watching the videos without depth information could not tell whether they were within or outside the workspace. Further work should explore underlying factors that contribute to the difference in perceived safety between these two formats of interaction.

Researchers utilizing online platforms should be aware that the depth and sound information (or lack thereof) may affect the findings because this information might be restricted in online experimental studies. Even though the data is collected from the same pool of participants, different formats of interaction can lead to significantly different findings, as well as other aspects. Chapter 5 described additional analysis that investigated whether differences in formats of interaction also impact task quality, subjective experience, and perceived payment fairness.

4.5 Summary

This Chapter presents studies that investigate whether the formats of interaction in HRI judgment task affect the robot’s perceived human-likeness and perceived safety. The design of experiment, experimental task, and the results of the studies is discussed.

4.6 References

Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, *3*, 41–52. https://doi.org/10.1007/s12369-010-0082-7

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, *1*(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Kose-Bagci, H., Ferrari, E., Dautenhahn, K., Syrdal, D. S., & Nehaniv, C. L. (2009). Effects of embodiment and gestures on social interaction in drumming games with a humanoid robot. *Advanced Robotics*, *23*(14), 1951–1996. https://doi.org/10.1163/016918609X12518783330360

Chapter 5
Comparisons of Online and Laboratory Experimental Studies of an HRI Task

5.1 Introduction

In this chapter, two comparative studies in the context of an HRI judgment task in which human participants judge a robot's target were reported. This Chapter focuses on task quality, subjective experience, motivation, and perceived payment fairness. Chapter 5.2 focuses on the information presentation format between participants observing a robot's physical movement and those watching video recordings. Participant pool (college students) and study location (in-lab) were held constant across both conditions. Chapter 5.3 focuses on the participant pool between online crowdworkers from the AMT platform and an in-lab experiment with college students.

5.2 In-laboratory Participants Observing a Robot’s Physical Movements vs. Their Videos

This section discusses additional analysis that compared two groups of students, one directly observed a physical robot in-lab performing the reach movements and the other watched the videos of the same movements, on their task quality, subjective experience, motivation, and perceived payment fairness. All participants were US-based college students. The analysis used the dataset collected as described in Chapter 4.

5.2.1 Methods

Participants

This section utilized the same dataset described in Chapter 4. A total of 48 college students were included in the analysis. 24 participants observed physical movement, and the rest of the participants were assigned to the video condition. The reach movements in both physical robot and video conditions were from the same set of movements.

5.2.2 Design of Experiment

This section compared two different formats of interaction: physical robot and video. Both conditions used the same dataset collected as described in Chapter 4.

Dependent variables

1. Task quality metrics
    I. Judgment accuracy (%) is defined as the proportion of trials the participants judged correctly.
    II. Desired response time rate (%) is defined as the proportion of trials in which the participants follow the instruction given in the study to make their final answers within the allocated timeframe.
2. Subjective experiences: Perceived frustration, perceived effort, and perceived performance using the NASA TLX 7-point scale (Hart & Staveland, 1988).
3. Motivation: Participants were asked to rate the following two statements: (1) I am motivated to do HITs on Mechanical Turk to make money, and (2) I am motivated to do

HITs on Mechanical Turk for fun (Antin & Shaw, 2012). Each question has a 7-point Likert scale with options from *strongly disagree* to *strongly agree*.

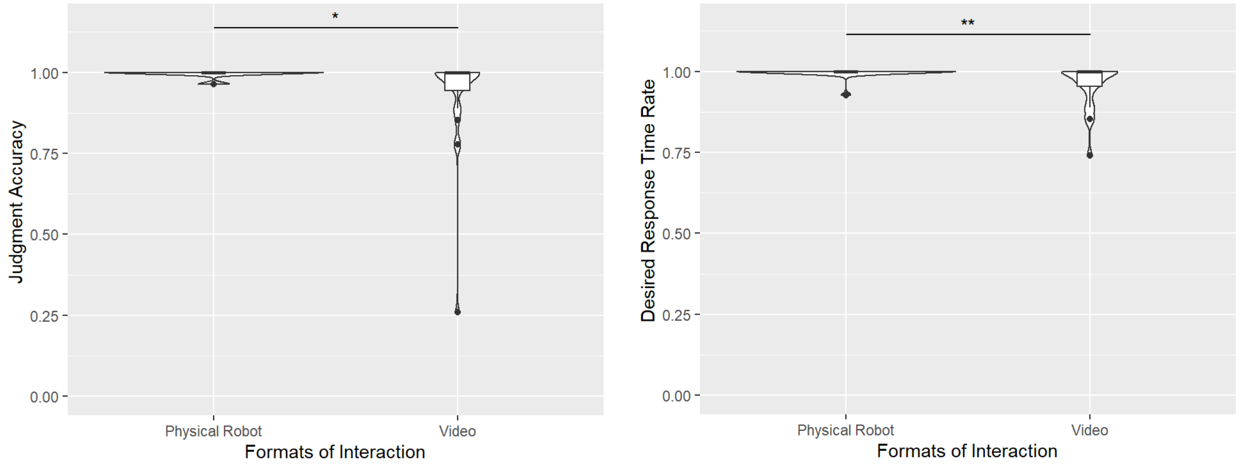
4. Perceived payment fairness: Participants rated the payment fairness with a categorical answer of “Fair”, “Neutral”, or “Unfair” (d’Eon et al., 2019).

Statistical Analysis

Welch’s two sample t-test were used to analyze the task quality (judgment accuracy and desired response time rate), subjective experience, motivation with formats of interactions as an independent variable. Pearson’s Chi-squared test was applied to test for differences in perceived payment fairness ratings among different formats of interaction. Welch’s two sample t-test and Pearson’s Chi-squared test were conducted using *stats* package in r 4.0.3.

5.2.3 Results

Figure 5-1 shows the violin plots for judgment accuracy and desired response time rate. Welch’s two sample t-test showed that the format of interaction significantly affected judgment accuracy ($t(23.288) = 2.098$, $p = 0.047$, 95% CI = [0.011, 1.200]) and desired response time rate ($t(25.335) = 2.861$, $p = 0.008$, 95% CI = [0.220, 1.431]). Participants in the physical robot condition perform better in both judgment accuracy (physical robot: $M = 0.995$, $SD = 0.013$ vs. video: $M = 0.927$, $SD = 0.158$) and desired response time rate (physical robot: $M = 0.997$, $SD = 0.015$ vs. video: $M = 0.957$, $SD = 0.067$) compared to participants in video condition. Effect size value (Cohen’s d = 0.606 for judgment accuracy and Cohen’s d = 0.826 for desired response time rate) suggested medium and large practical impact for judgment accuracy and desired response time rate, respectively.



(a) Judgment accuracy

(b) Desired response time rate

Figure 5-1: Violin plots of task quality metrics: judgment accuracy and desired response time rate.

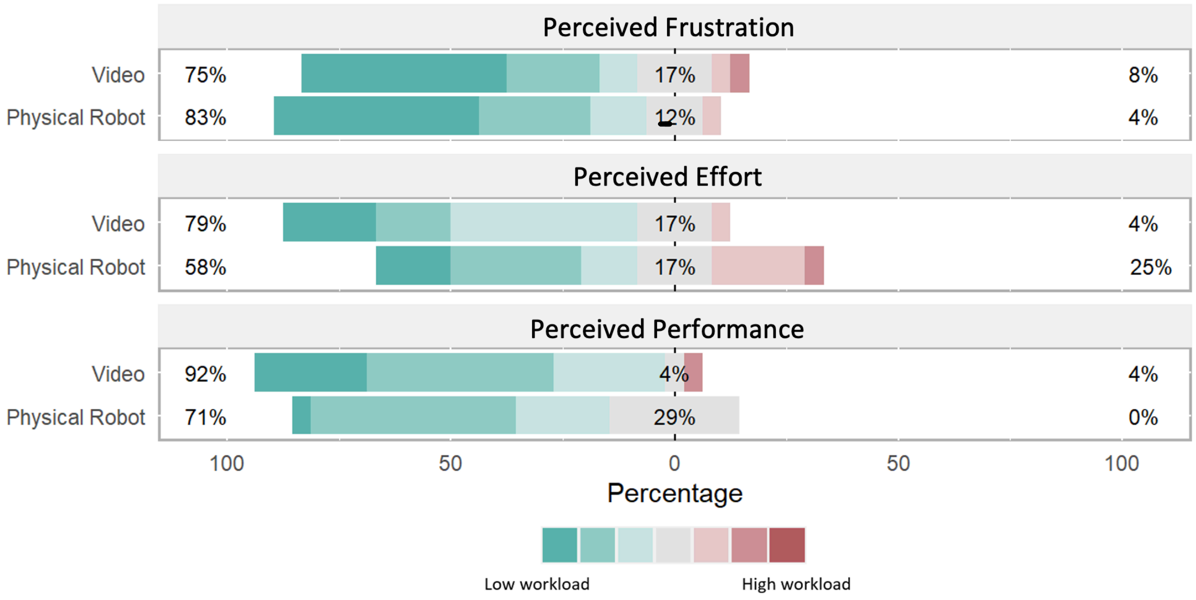


Figure 5-2: Bar charts of subjective experience metrics: Perceived frustration, effort, and performance.

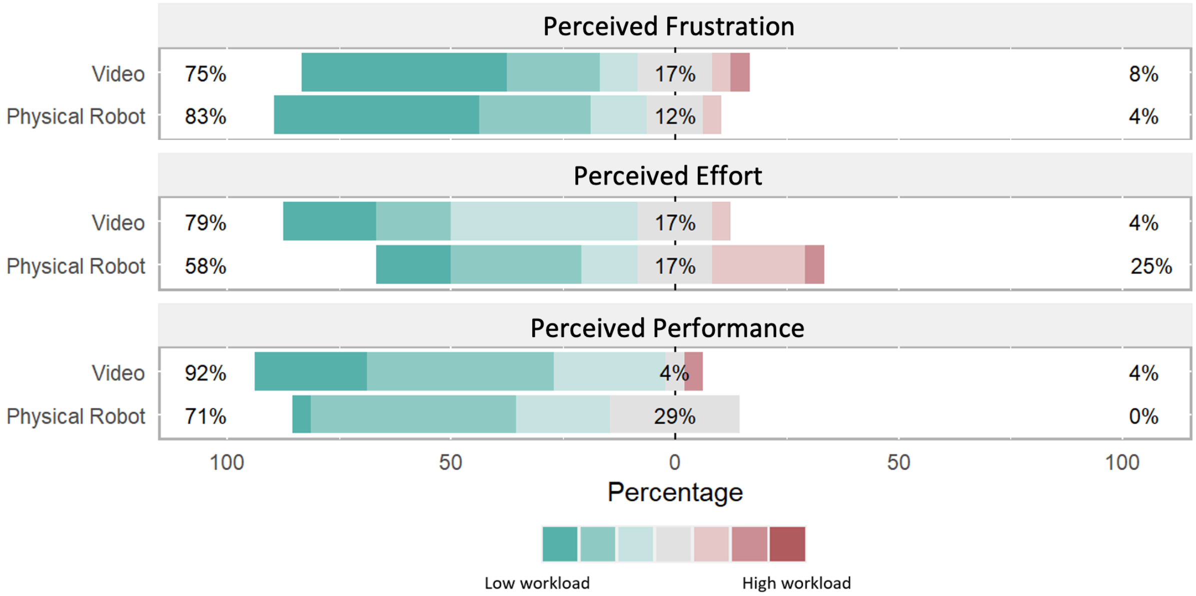


Figure 5-2 shows the bar charts for subjective experience metrics. Welch's two sample t-test did not reveal a significant difference in participants' perceived frustration (*t*(44.221) = 0.523*, p* = 0.603), perceived effort (*t*(41.94) = 1.061*, p* = 0.295), and perceived performance (*t*(44.297) = 1.645*, p* = 0.107) between different formats of interaction.

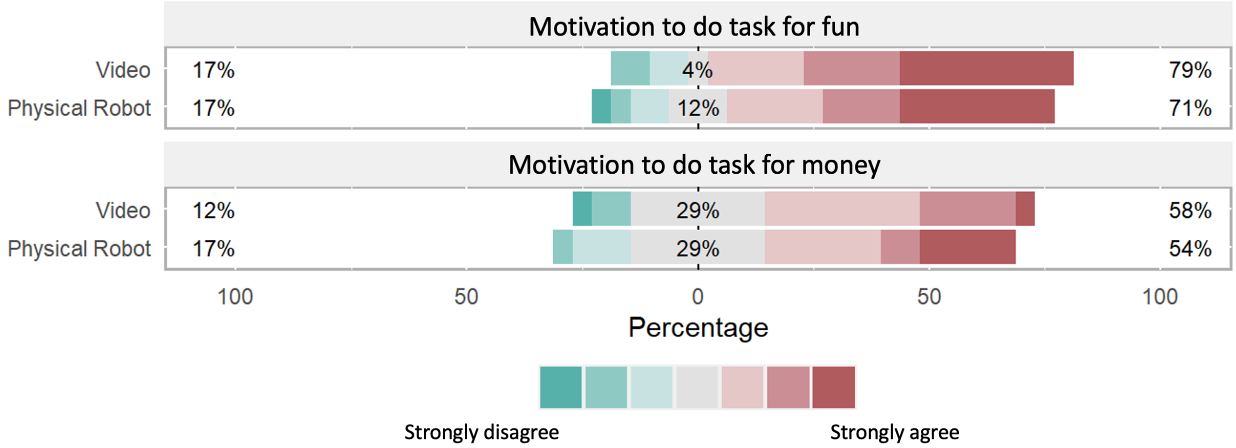


Figure 5-3: Bar charts of participants' motivation: For fun and for money.

Figure 5-3 shows the bar charts for motivation metrics. Welch's two sample t-test did not reveal a significant difference in participants' motivation to complete a task for fun (*t*(45.811) = 0.510, *p* = 0.612) and motivation to complete a task to make money (*t*(45.938) = 0.602, *p* = 0.550) between different formats of interaction.

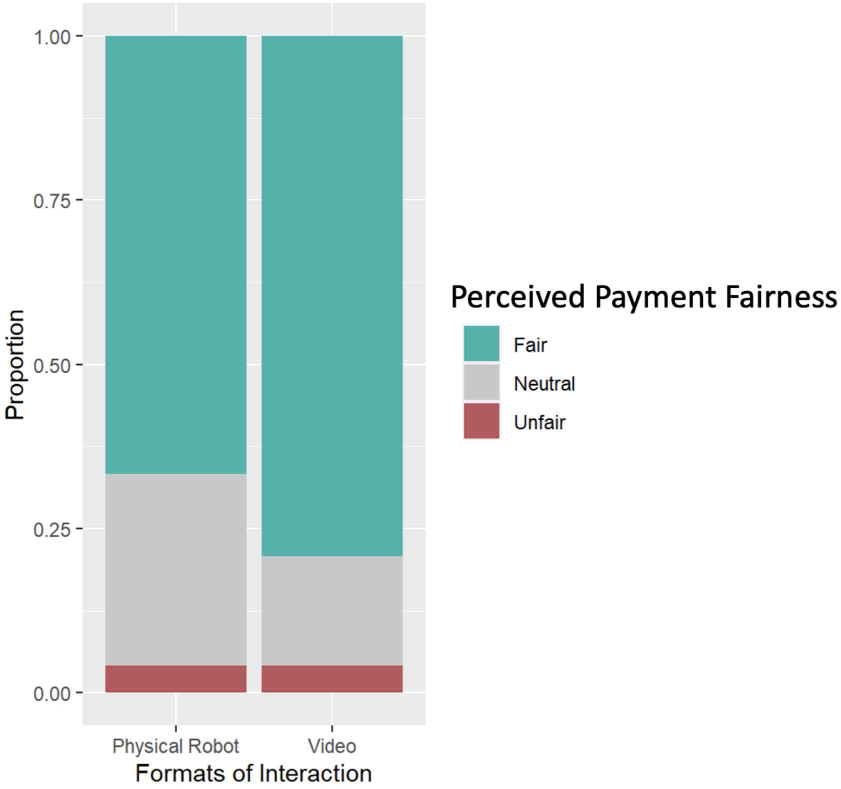


Figure 5-4: Stacked bar charts of perceived payment fairness.

Figure 5-4 shows stacked bar charts of perceived payment fairness by format of interaction. Pearson's Chi-squared test did not reveal a significant difference in perceived payment fairness ratings among different formats of interaction (χ^2(*DOF* = 2, *N* = 48) = 1.075, *p* = 0.584).

The analysis described in Chapter 5.2 compared data collected from participants evaluating a robot's physical movements in a face-to-face interaction in-lab with a "simulated" online study in-lab. In both conditions, participants did not represent the sample group that are generally recruited for online tasks such as online crowdworkers. The next section discusses studies investigates similarities and differences between in-lab participants vs. online crowdworkers.

5.3 In-laboratory Participants vs. Online Crowdworkers

The main objective of Chapter 5.3 is to discuss studies comparing task quality, subjective experience, motivation, and perceived payment fairness between in-lab students and online crowdworker participants, as well as the effects of payment rates on online crowdworker performance and behavior. The research questions and hypotheses are as follows:

Research question 2 (RQ2): Do in-lab and online participants provide data of the same quality?

H2: In-lab participants would differ from online participants in terms of the data quality they provide.

5.3.1 Methods

Participants

84 participants were included in the analysis: 60 AMT participants located in the US and 24 students enrolled in the University of Michigan. AMT participants were recruited over three

different tasks posted on the AMT platform with the only difference being the payment rate of $3, $9, and $15, each group has 20 participants. An AMT participant cannot enroll in this series of study more than once with the same worker ID. All 24 students were enrolled in the in-lab condition described in Chapter 4. Note the effects of payment rate were only studied in the crowdworkers but not students because it is out of the norm to recruit students with a rate of payment of $3/hour or $9/hour. To equate the number of participants in all conditions, 20 out of 24 participants were randomly selected for the analysis. See Table 5-1 for participants' demographic in each condition. The demographic information revealed that AMT participants were older than college students on average, consistent with related literature (Buchheit et al., 2018; Kees et al., 2017; Lewis et al., 2009).

Table 5-1: Studies' conditions and participant demographic information.

Participants	Payment	N	Male (%)	Mean age - *SD*	Recruitment Period
AMT-US	$9	20	70.00	34.60 - 9.27	Sept 27th - Oct 30th, 2020
AMT-US	$3	20	55.00	42.15 - 12.89	Jan 8th - Jan 19th, 2021
AMT-US	$15	20	60.00	41.79 - 13.10	Nov 13th - Dec 16th, 2021
Students	$15	20	65.00	22.84 – 3.13	Mar 14th - May 14th, 2022

5.3.2 Design of Experiment

Independent variables

Online participants were enrolled in one of three payment rates: $3, $9, and $15. Hence, this study compared 4 groups of participants: US-based college students enrolled in an in-lab study with a payment rate of $15 and three groups of US-based AMT participants enrolled online with three different payment rates ($3, $9, and $15).

The in-lab condition simulates a human monitoring a robot's movement by observing it through a computer screen, as described in Chapter 4. Only students were recruited as participants in this condition. The experiment session was no longer than 90 minutes. A researcher was seated nearby throughout the study session. The consent form was signed in ink. Participants received payment immediately after their participation in the study was over.

The online condition followed the same procedure as the in-lab condition with the following differences: First, all of the participants were recruited via the AMT platform with restrictions set to recruit only US-based participants. Second, the participants can participate in this study at the location and time of their choosing but they must complete the study within 24 hours after they enrolled in this study via the AMT platform. They did not have to complete the study in one sitting. Third, the consent form was displayed in a link and the participant only needs to click the button "I consent" to proceed to the study. Fourth, participants received payment within 72 hours after they submitted a completion code via the AMT platform. Fifth, the online condition used a different set of movements from the in-lab condition. The online condition used a set of movements with a mean length of 7.50 seconds (*SD* = 0.86), while the in-lab condition used a different set of movements with a mean length of 9.09 seconds (*SD* = 0.17). The effects of video length were assumed to have no impact on the findings because this aspect is unrelated to whether the participant followed the instruction to answer before the video ended or not.

The task was estimated to take around one hour to complete, thus $3, $9, and $15 are equivalent to an hourly wage of $3/hour, $9/hour, and $15/hour respectively. According to

Guideline for Academic Requester (Dynamo, 2014) compiled by Dynamo users (Salehi et al., 2015), $6/hour is the recommended minimum payment. In the case of this study, $3/hour, $9/hour, and $15/hour represent below minimum-pay, above-minimum-pay, and high-pay respectively.

Dependent variables

The dependent variables of interest are as described in Chapter 5.2.

Experimental Task

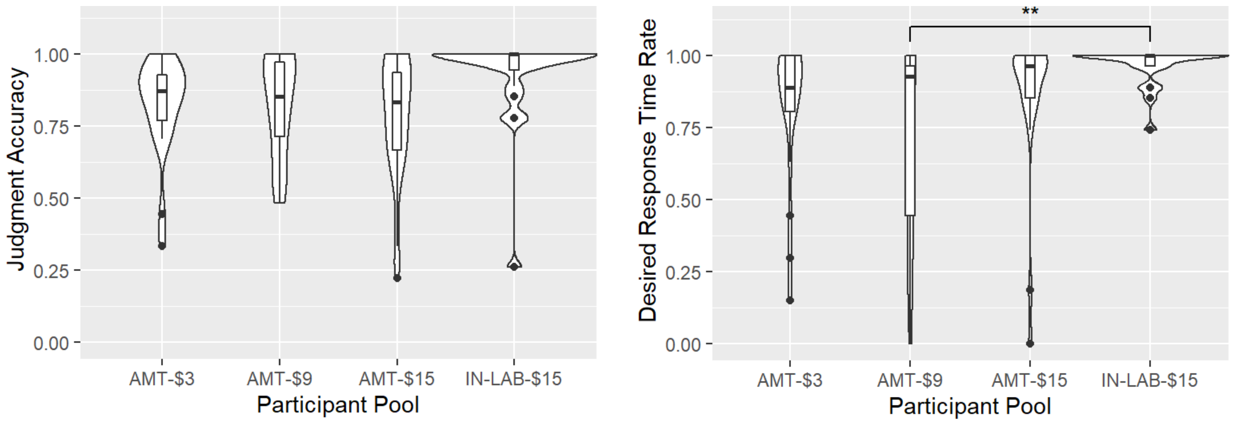
There was no restriction on rewatching the video again before filling in the questionnaire. Once the participants completed 27 trials, they were directed to a post-task questionnaire that asked about the participant's subjective workload using a subset of the NASA TLX questionnaire (Hart & Staveland, 1988). They were also asked about the payment fairness of this task. All studies described in Chapter 5 was approved by the University of Michigan Institutional Review Board (HUM00188416 and HUM00207720).

Statistical Analysis

One-way ANOVA with post-hoc Bonferroni adjustment was applied to find the differences in task quality, subjective experience, motivation, and perceived payment fairness between the four participant groups. Pearson's Chi-squared method was applied to test differences in perceived payment fairness ratings among the four participant groups.

5.3.3 Results

Figure 5-5 shows the violin plots for judgment accuracy and desired response time rate. One-way ANOVA showed that the participant group significantly affected desired response time rate (*F* (3, 76) = 3.882, *p* = 0.012). Main effect of participant group on judgment accuracy was not statistically significant (*F* (3, 76) = 1.996, *p* = 0.122). For the desired response time rate, post-hoc analysis with Bonferroni adjustment revealed that in-lab participants (*M* = 0.963, *SD* = 0.069) were more likely to make their final answer within the desired time as instructed by the study direction than the AMT participants in the $9 group (*M* = 0.693, *SD* = 0.340, *p* = 0.006). The other pairs were not significantly different from each other.



(a) Judgment accuracy

(b) Desired response time rate

Figure 5-5: Violin plots of task quality metrics:

judgment accuracy and desired response time rate.

Figure 5-6 shows the bar charts for subjective experience metrics. One-way ANOVA showed that the main effect of participant group on perceived effort was significant (*F* (3, 76) = 17.17, *p* < 0.001). Post-hoc analysis with Bonferroni adjustment revealed that students (*M* =

2.750, $SD = 1.118$) perceived less workload in terms of effort compared to online AMT participants in the \$3 group ($M = 4.900$, $SD = 1.373$, $p < 0.001$), \$9 group ($M = 5.000$, $SD = 1.298$, $p < 0.001$), and \$15 group ($M = 5.200$, $SD = 1.152$, $p < 0.001$). There were no significant differences in perceived effort between AMT participants in the different payment levels. Perceived frustration ($F\ (3, 76) = 0.448$, $p = 0.719$) and perceived performance ($F\ (3, 75) = 1.552$, $p = 0.208$) were not found to be significantly different between participant groups.

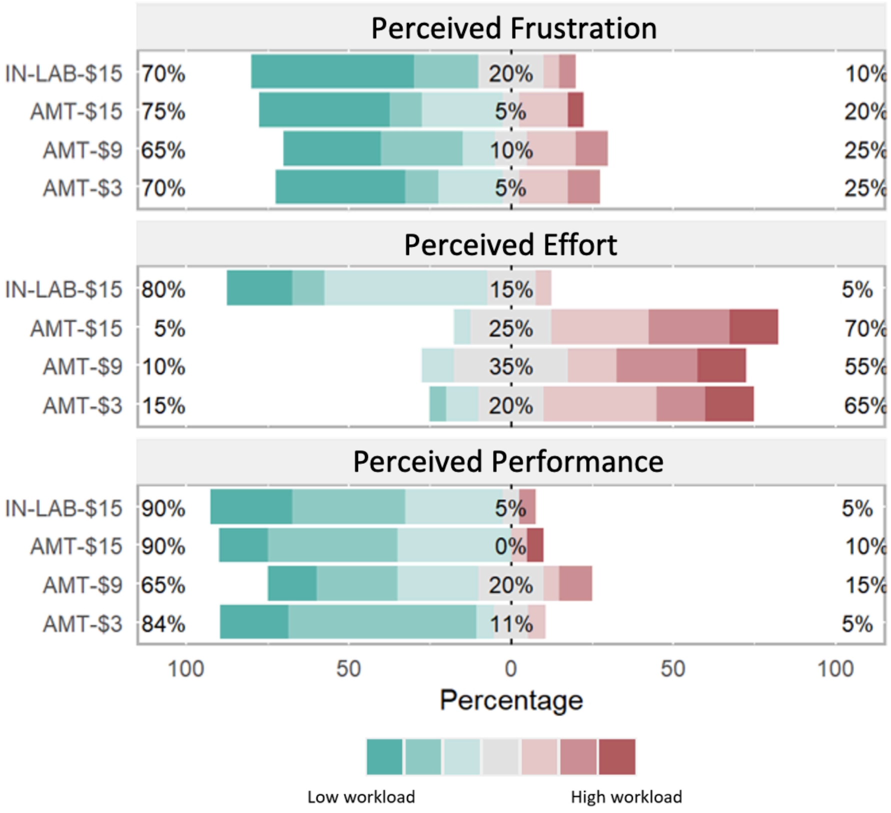


Figure 5-6: Bar charts of subjective experience metrics: perceived frustration, effort, and performance.

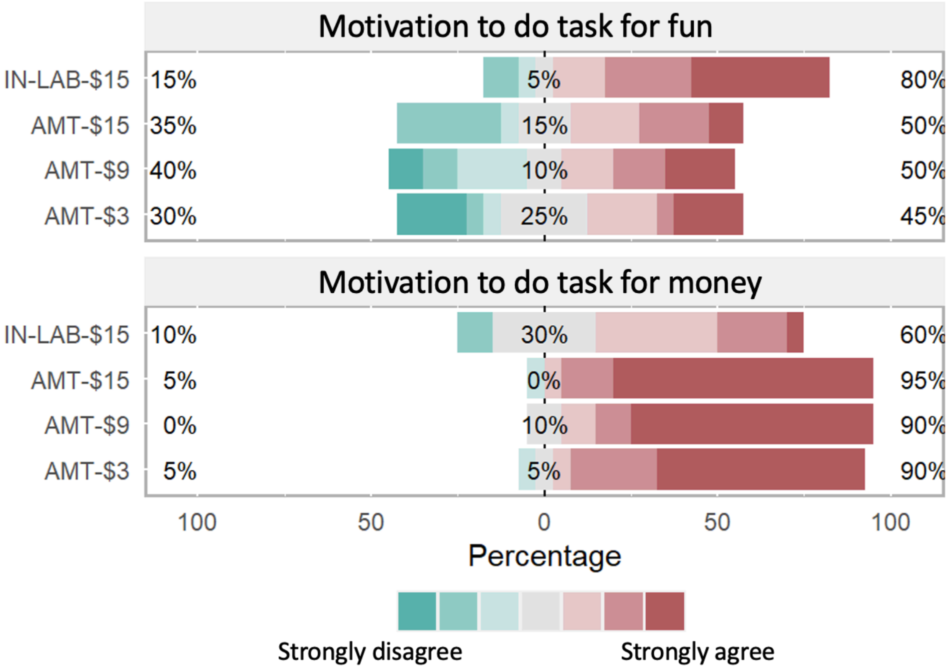


Figure 5-7: Bar charts of participants' motivation: For fun and for money.

Figure *5-7* shows the bar charts for motivation metrics and the bar chart for perceived payment fairness. For the motivation metrics, one-way ANOVA showed that participant group significantly affected participants' motivation to participate in the study for money (*F* (3, 76) = 12.06, $p < 0.001$) and for fun (*F* (3, 76) = 2.531, $p = 0.063$) (marginally significant). Students rated their motivation to complete a task for money lower than online AMT participants in the $3 group, $9 group, and $15 group. For the motivation to complete a task for money, post-hoc analysis with Bonferroni adjustment revealed that students (*M* = 4.700, *SD* = 1.261) rated the motivation to complete a task to make money lower than AMT participants in the $3 group (*M* = 6.300, *SD* = 1.129, $p < 0.001$), $9 group (*M* = 6.400, *SD* = 1.046, $p < 0.001$), and $15 group (*M* = 6.550, *SD* = 0.999, $p < 0.001$). There were no significant differences in motivation to complete a task to make money between online AMT participants in the different payment levels. For the motivation to complete a task for fun, post-hoc analysis with Bonferroni adjustment, there were no significant differences in motivation to complete a task for fun between students and online participants in the different payment levels.

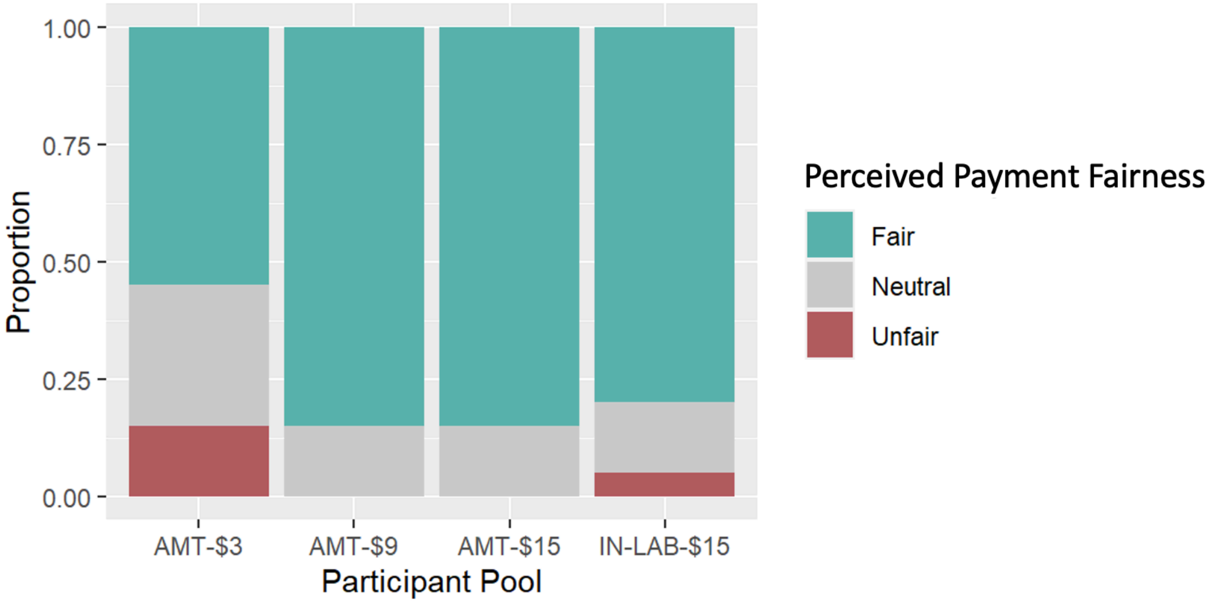


Figure 5-8: Stacked bar charts of perceived payment fairness.

Figure 5-8 shows stacked bar charts for perceived payment fairness. Pearson's Chi-squared test revealed no statistically significant differences in perceived payment fairness ratings among participant groups ($\chi^2(DOF = 6, N = 80) = 9.423, p = 0.151$). This study has shown that there were differences between the students watching video in-lab and AMT participants watching video online in terms of task quality, subjective experience, motivation, and perceived payment fairness.

5.4 Discussion

In terms of task quality, the results discussed in Chapter 5.3 showed that in-lab participants were able to follow the instruction more closely than one of the AMT participant groups. This was an expected result as AMT participants are diverse in survey-taking experience and skill set, which vary the rate they pay attention to the task instruction (AMT-$3: *M* = 0.815, *SD* = 0.247, AMT-$9: *M* = 0.693, *SD* = 0.340, AMT-$15: *M* = 0.848, *SD* = 0.270). Results discussed in Chapter 5.2 showed that students observing a physical robot scored higher than those watching videos in both judgment accuracy and answering within the specified time. Note that the difference between the two formats of interaction was small in both desired response time rate (physical robot: *M* = 0.997, *SD* = 0.015 vs. video: *M* = 0.957, *SD* = 0.067) and judgment accuracy (physical robot: *M* = 0.995, *SD* = 0.013 vs. video: *M* = 0.927, *SD* = 0.158), although the t-test has shown that the difference is marginally significant. A possible explanation is that the difference in accuracy came from the presence (or lack thereof) of depth and sound information between the two formats of interaction. The instructions in the physical robot were also verbally read out to the participants which might be more effective in conveying the task's instruction compared to the text passage in the video study. This finding does not support H2; participants observing a robot's physical movements were more accurate in their judgment compared to participants watching videos.

In terms of subjective experiences, the results discussed in Chapter 5.3 findings showed that all three AMT participant groups rated their effort required to complete a session higher than students completing a study in-lab. The results discussed in Chapter 5.2 finding added that there were no significant differences in perceived effort between video or physical robot conditions. Hence, the differences in subjective experiences found in Chapter 5.3 can be attributed to the participant pool or other factors such as a lack of a controlled environment or monitoring from the researchers. In terms of motivation and perceived payment fairness, in-lab participants displayed a similar trend that they were more likely to participate out of interest rather than financial incentives compared to AMT participants. There was no difference in the perception of payment fairness between AMT participants who received $9 and $15. However, AMT participants who received $3 perceived the payment as less fair compared to AMT participants who received $9 and $15. Note the length of the study was approximately 1 hour, but this can vary depending on participants' experience in taking surveys (Kees et al., 2017; Smith et al., 2016; Weigold & Weigold, 2022). These findings are expected as Guideline for Academic Requester (Dynamo, 2014) made a recommendation that the minimum payment on AMT should be at least $6/hour. Payment of $3 for 1 hour of work would equate to $3/hour, which is lower than the recommended $3/hour and would be perceived as unfair, while $9/hour and $15/hour payment are more acceptable.

Many AMT participants in the $3/hour group perceived this payment rate as fair (55%) or neutral (30%) even though $3/hour was significantly below the recommended minimum payment per hour of $6/hour as well as the US minimum wage of $7.25/hour (rate of the time of data collection in 2020-2021). A possible explanation is the power of the requester to set the wage of their task and workers' preference to perform specific types of work, which manifested as monopsony power in the online crowdwork platform (Dubes et al., 2020). Monopsony refers to a market structure where only one buyer (requesters in this context) is available to purchase goods and services from many sellers (online crowdworkers in this context). Cantarella & Strozzi (2022) discussed that online crowdworkers were inclined to work more hours than they wished to reach their earning goal because the search efforts to find high-pay tasks may negatively impact their earning per hour. This line of rationales may also explain why participants chose to complete a low-pay task in this series of studies and in Buhrmester et al. (2011). Participants were not inclined to complete low-pay tasks for fun, as illustrated in this study that there were no significant differences in motivation to complete a task (both for money and for fun) between different levels of payment rate.

Chapter 5 discussed an analysis compared data collected from participants evaluating a robot's physical movements in a face-to-face interaction in-lab with a "simulated" online study in-lab, and another analysis compared a "simulated" online study in-lab with AMT participants. However, AMT participants discussed in Chapter 5 are only limited to those located in the US. Chapter 6 focused on AMT participants with an additional level of location and payment method included in the analysis to get a clearer picture of online participants' characteristics and their impact on research studies.

5.5 Summary

Chapter 5 discusses two studies investigating the similarities and differences in task quality, subjective experience, motivation, and perceived payment fairness between observing a physical movement and watching videos, and between traditional in-lab participants and online crowdworkers in the context of HRI judgment tasks.

5.6 References

Antin, J., & Shaw, A. (2012). Social desirability bias and self-reports of motivation: A study of amazon mechanical turk in the US and India. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2925–2934. https://doi.org/10.1145/2207676.2208699

Buchheit, S., Dalton, D. W., Pollard, T. J., & Stinson, S. R. (2018). Crowdsourcing Intelligent Research Participants: A Student versus MTurk Comparison. *Behavioral Research in Accounting*, *31*(2), 93–106. https://doi.org/10.2308/bria-52340

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *In A. E. Kazdin (Ed.), Methodological Issues and Strategies in Clinical Research (p. 133–139). American Psychological Association*. https://doi.org/10.1037/14805-009

Cantarella, M., & Strozzi, C. (2022). Piecework and Job Search in the Platform Economy. *IZA Discussion Paper* No. 15775. http://doi.org/10.2139/ssrn.4296733

d'Eon, G., Goh, J., Larson, K., & Law, E. (2019). Paying Crowd Workers for Collaborative Work. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–24. https://doi.org/10.1145/3359227

Dube, A., Jacobs, J., Naidu, S., & Suri, S. (2020). Monopsony in online labor markets. American Economic Review: Insights, 2(1), 33-46. https://doi.org/10.1257/aeri.20180150

Dynamo. (2014). *Guidelines for Academic Requesters (Version 1.1)*. https://irb.northwestern.edu/docs/guidelinesforacademicrequesters-1.pdf

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier. https://doi.org/10.1016/S0166-4115(08)62386-9

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of Advertising*, *46*(1), 141–155. https://doi.org/10.1080/00913367.2016.1269304

Lewis, I., Watson, B., & White, K. M. (2009). Internet versus paper-and-pencil survey methods in psychological experiments: Equivalence testing of participant responses to health-related messages. *Australian Journal of Psychology*, *61*(2), 107–116. https://doi.org/10.1080/00049530802105865

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Salehi, N., Irani, L. C., Bernstein, M. S., Alkhatib, A., Ogbe, E., & Milland, K. (2015). We are dynamo: Overcoming stalling and friction in collective action for crowd workers. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1621–1630. https://doi.org/10.1145/2702123.2702508

Smith, S. M., Roster, C. A., Golden, L. L., & Albaum, G. S. (2016). A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples. *Journal of Business Research*, *69*(8), 3139–3148. https://doi.org/10.1016/j.jbusres.2015.12.002

Weigold, A., & Weigold, I. K. (2022). Traditional and modern convenience samples: An investigation of college student, Mechanical Turk, and Mechanical Turk college student samples. *Social Science Computer Review*, *40*(5), 1302–1322. https://doi.org/10.1177/08944393211006847

Chapter 6
Effects of Payment Scheme on Online Crowdwork Performance and Experience

6.1 Introduction

In this chapter, the results of four studies that have been conducted to explore the effects of participant location, payment method, and payment rate on online crowdwork task quality, subjective experience, motivation, and perceived payment fairness were reported. The four studies were conducted online via the AMT platform using an HRI judgment task. 160 participants participated in the four studies; 80 of them were from the United States (US) and the rest were from India (IN). Task performance measurements include the participant's judgment accuracy on the HRI judgment task and the desired response time rate, which measured the rate at which the participant followed task instructions to perform time-sensitive tasks. The task performance data were analyzed using 3-Way Analysis of Covariance (ANCOVA). The payment fairness was analyzed using the chi-square test of independence.

Chapter 6 describes four studies conducted on AMT to address the goal above. Participants were enrolled in four online studies that asked them to observe videos of a robot's movements and answer questionnaires. The four studies differ in the combination of payment rates and payment methods. Studies 1 and 2 employed the quota payment-method conditions. Studies 3 and 4 used the piece-rate payment-method conditions. Studies 1 and 3 paid $9/hour, while Studies 2 and 4 paid $3/hour. Studies 1-4 recruited participants from both US and India.

6.2 Methods

6.2.1 Participants

All participants were recruited from AMT. See Table 6-1 for the overview of the four studies described in this chapter. 160 participants were included in the analysis. The recruitment period for the AMT participants was not overlapped with each other to prevent the potential participants from participating more than once. When a participant enrolled in one of the studies, they cannot participate in another study in the same series again using the same worker ID.

Table 6-1: The conditions and participant demographic information of the four studies.

Payment Method	Rate (USD/hour)	Participants	N	Male (%)	Mean age - *SD*	Recruitment Period
Quota	$9	AMT-US	20	70.00	34.60 - 9.27	Sept 27th - Nov 2nd, 2020
		AMT-IN	20	65.00	30.75 - 5.55	
Quota	$3	AMT-US	20	55.00	42.15 - 12.89	Jan 8th - Jan 19th, 2021
		AMT-IN	20	80.00	34.25 - 6.66	
Piece-rate	$9	AMT-US	20	60.00	36.90 - 12.59	Feb 8th - Feb 14th, 2021
		AMT-IN	20	80.00	32.05 - 6.06	
Piece-rate	$3	AMT-US	20	35.00	32.20 - 7.59	Feb 15th - Mar 20th, 2021
		AMT-IN	20	80.00	33.10 - 6.90	

* AMT-IN = India-based participants recruited from AMT

6.2.2 Design of Experiment

Independent variables.

1. Payment rate: Low rate ($3/hour) vs. high rate ($9/hour)
2. Participant Location: US-based participants vs. India-based participants
3. Payment method: Quota vs. piece-rate
    - Quota: Participants are instructed in advance that they must complete all the 27 trials to be eligible for payment.
    - Piece-rate: Participants are instructed in advance that they may stop at any point after completing at least 1 trial, and the payment is computed based on the number of tasks completed before the participant decides to stop.

Low rate ($3/hour) and high rate ($9/hour) decided from a recommendation by Guideline for Academic Requester (Dynamo, 2014) that the minimum payment on AMT should be at least $6/hour. Low rate and high rate are $3/hour below and above the recommended minimum payment, respectively.

US-based and India-based participant were chosen because (1) they are the largest and second largest group of participants on AMT, respectively (Difallah et al., 2018), and (2) both are English-speaking countries.

Participants enrolled in the quota payment method (Studies 1 and 2) were told that they had to perform all the 27 trials to be eligible for payment. Participants enrolled in the piece-rate payment method (Studies 3 and 4) were told that, after completing at least one trial, they had the option to end the data collection session and skip to the post-task questionnaire. If the participants choose to end the data collection, the survey would ask for their confirmation and warn that they would not be able to come back to the data collection session. The participants were paid at the rate of $(payment rate/30) per trial completed in addition to $(payment rate/10) participation base rate. Note that if two participants enrolled in the same level of payment rate but in a different payment method condition, both participants would receive the same amount of payment if the participant in the piece-rate condition completed all 27 trials.

Dependent variables

1. Task quality metrics
    - Judgment accuracy (%) is defined as the proportion of trials the participants judged correctly.
    - Desired response time rate (%) is defined as the proportion of trials in which the participants follow the instruction given in the study to make their final answers within the allocated timeframe.
2. Number of participants rejected due to low quality work (fraudulent respondents) (Kennedy et al., 2020). This metric tallies the number of participants that (1) failed the attention check questions - questions for which correct answers exist, (2) participated in the same task from different accounts with copy-paste answers for open-ended questionnaires, or (3) spoofed their location to participate in this study.
3. Subjective experiences: Perceived frustration, perceived effort, and perceived performance using the NASA TLX 7-point scale (Hart & Staveland, 1988).
4. Motivation: Participants were asked to rate the following two statements: (1) I am motivated to do HITs on Mechanical Turk to make money, and (2) I am motivated to do

HITs on Mechanical Turk for fun (Antin & Shaw, 2012). Each question has a 7-point Likert scale with options from *strongly disagree* to *strongly agree*.

5. Perceived payment fairness: Participants rated the payment fairness with a categorical answer of "Fair", "Neutral", or "Unfair" (d'Eon et al., 2019).

6.2.3 Experimental Task

The participants performed an online task titled "Prediction of Automation's Intention in Object Handling Tasks." The task instruction informed participants that this task only involved observing an industrial robot performing tasks and answering questionnaires. The main task of these four studies involved watching videos of a robot performing a reach movement with 9 possible reach destinations on three shelves. During the experiments, participants were provided a link to a Qualtrics survey and asked to predict the robot's reach destination using the interface provided in the online survey. Before they proceed, each participant was provided an informed consent form, which they have to click "consent" to proceed. Next, the participants filled in a demographic questionnaire. Then, they went through two online training pages. The first training page showed nine videos with ground truth labels - the correct answers corresponding to each training video. The second training page showed the same nine videos in randomized order, but without ground truth labels. In the data collection session, the survey instructed the participants to predict the robot's reach destination as quickly and accurately as they can and then fill out a questionnaire to evaluate the robot's movement quality. During the data collection, the survey would occasionally show a multiple-choice attention-check question such as "What was the color of the robot?", "How many plastic cups were on the shelf?," or "What was the color of the wall behind the robot?." The answers to these questions can be obtained by reviewing the video again on the same page. These questions served as screening methods to detect participants who paid too little attention or used automated software to complete the task.

Once the participants finished the data collection session, they were directed to a post-task survey, which asked the participants to rate their perceived effort, frustration, and performance using the NASA TLX questionnaire (Hart & Staveland, 1988). The post-task survey also asked them to rate their motivations (Antin & Shaw, 2012) and perceived payment fairness (d'Eon et al., 2019). At the end of the study, the survey engine provided a debrief form, which revealed that the study also investigates the effects of payment rate and payment method on the participants' task quality, subjective experience, motivation, and perceived payment fairness. The survey did not reveal these research objectives before they complete the task so as to make participants' response to the task as natural as possible. All four studies described in Chapter 6 were approved by the University of Michigan Institutional Review Board (HUM00188416).

6.2.4 Statistical Analysis

Three-way ANCOVA models were used to analyze the task quality (judgment accuracy and desired response time rate), subjective experience, motivation, and perceived payment fairness with participant location, payment rate, and payment method as independent variables and gender, age, robot familiarity, and education level as covariates. Pearson's Chi-squared test was applied to test for differences in perceived payment fairness ratings among different locations, payment rates, and payment methods.

6.3 Results

The participants were asked to rate their familiarity with a robotic arm using a single-select categorical question with 4 options: None, a little (have seen one before), some (have interacted with one before), or a lot (have worked with one before). These options were coded into 4 levels: 0 to 3 respectively. Of the 160 participants, 29 had no familiarity with a robotic arm, 69 had seen a robotic arm before, 35 had interacted with a robotic arm before, 24 had worked with a robotic arm before, and 3 did not disclose. The participants were also asked about the highest level of school education they have completed or the highest degree they have received. Of the 160 participants, 10 answered high school or associate degree, 120 answered bachelor’s degree, 21 answered Master’s degree, 2 answered Ph.D., and 7 did not disclose.

6.3.1 Task Quality

Table 6-2 and Figure 6-1 show the three-way ANCOVA results and boxplots for desired response time rate and judgment accuracy respectively. The results showed that participant location significantly affected the desired response time rate, (F (1, 132) = 6.320, p = 0.013) with US-based participants (M = 0.749, SD = 0.335) scoring higher desired response time rates than India-based participants (M = 0.589, SD = 0.334). A US-based participant is more likely to pay more attention to the task’s instructions and perform a time-sensitive task than an India-based participant. A possible explanation is language barrier. English is one of India’s official languages, but India-based participants may not have the same proficiency in English language as US-based participants. The three-way ANCOVA revealed no statistically significant difference between payment rate, payment method, or participant location on judgment accuracy.

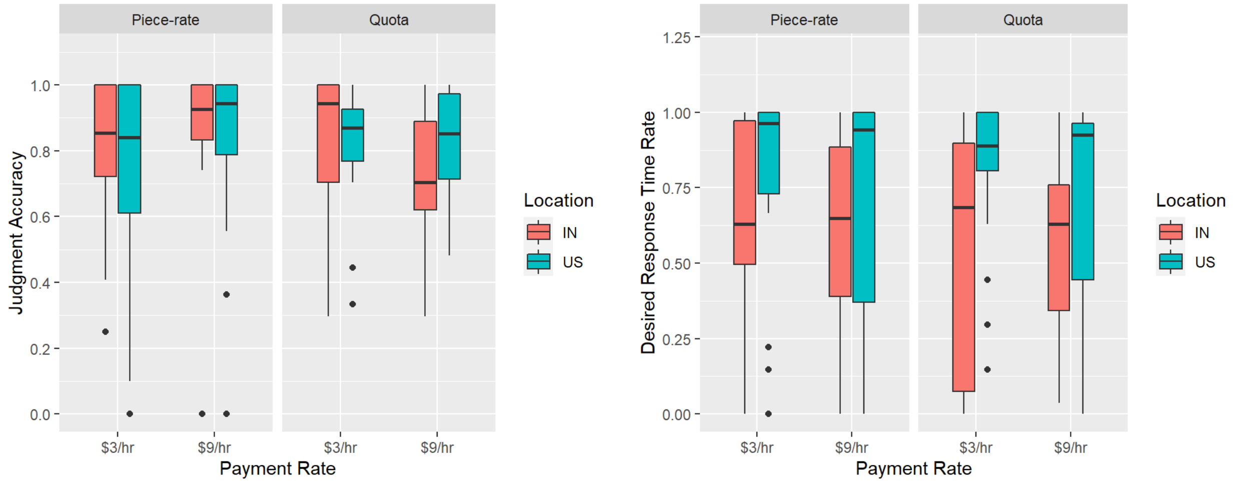
No significant two-way or three-way interaction ($p > 0.10$) was found for desired response time rate and judgment accuracy.

Table 6-2: Three-way ANCOVA results for task quality metrics:

desired response time rate and judgment accuracy.

	Desired response time				Accuracy			
Outcome Variables	df	F	*p*		df	F	*p*	
Gender	1	0.519	0.472		1	0.020	0.889	
Age	1	4.181	0.043	*	1	2.866	0.093	.
Robot Familiarity	3	3.950	0.010	*	3	1.205	0.310	
Education Level	3	0.169	0.917		3	0.454	0.715	
Location	1	6.320	0.013	*	1	0.183	0.670	
Payment Method	1	0.054	0.817		1	0.043	0.836	
Payment Rate	1	0.231	0.632		1	0.001	0.982	
Location : Payment Method	1	1.388	0.241		1	0.816	0.368	
Location : Payment Rate	1	0.046	0.831		1	1.323	0.252	
Payment Rate : Payment Method	1	0.445	0.506		1	1.313	0.254	
Location : Payment Rate : Payment Method	1	0.189	0.664		1	0.475	0.492	
Error	132				132			

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, · $p < 0.1$)



(a) Judgment accuracy

(b) Desired response time rate

Figure 6-1: Boxplots of task quality metrics: judgment accuracy and desired response time rate.

In this series of studies, 33 participants were rejected because they were suspected to be fraudulent participants. A fraudulent participant meets at least one of the following criteria: (1) provided random answers to attention check questions, (2) multiple accounts with copy-paste answers for open-ended questionnaires, or (3) mismatch between participant location constrained by AMT system and participants' reported location in the demographic questionnaire. For example, 5 different accounts answered in the open-ended answer that "I sit and watch the robot." This answer is unlikely to be from 5 different people. Thus, these participants were rejected on the ground that they may have come from the same person. Note that these participants also input different age and gender, presumably to make it convincing that the submissions came from different participants. Some participants with registered account in India (AMT platform allow participants registered in India to see the task) reported that their location is in the US were also rejected, as the reported location is not consistent with the registered location in AMT platform. Note that the rejected participants were not included in the data analysis of other dependent variables discussed in this dissertation and new participants were recruited to replace the rejected participants.

Table 6-3: Number of rejected participants by condition and participant location.

Payment scheme	Base payment*	US-based	India-based
Quota-$9	$9.00	0	6
Quota-$3	$3.00	0	10
Piece-rate-$9	$0.90	0	16
Piece-rate-$3	$0.30	0	1

* Base payment refers to the amount of payment of the task as observed by participants on AMT platform.

Table 6-3 shows the number of rejected participants by payment scheme. All the rejected participants were India-based. Although the quota payment method pays the same amount of compensation as the piece-rate method at the same payment rate, the number of fraudulent participants enrolled in the piece-rate payment method at $9/hour was higher than the quota payment method at $9/hour and $3/hour. The interaction effect between payment rate and payment method on the number of rejected participants was observed. In the quota payment method, $3/hour payment rate attracted more fraudulent participants than $9/hour. However, in the piece-rate payment method condition, $9/hour payment rate attracted more fraudulent participants than $3/hour. A possible explanation is that the fraudulent participants chose the task to perform based on the amount of base pay, which is the minimum amount of payment for the task that was displayed to the prospective participants. However, a piece-rate task with a payment rate of $3/hour has a base payment of $0.3, which might appear that the reward was low and consequentially was not targeted by fraudulent participants. Similarly, Chandler and Kapelner (2013) found that India-based participants produce lower quality work in an image labeling task. This finding may be attributed to either the design of the task or the rate of payment. Our study provided \$3/hour or \$9/hour while Chandler and Kapelner (2013) provided \$1.5/hour on average, which may attract participants who particularly provide low-quality work or those using automated software.

6.3.2 Subjective Experience

Table 6-4 and Figure 6-2 shows the ANCOVA output and boxplots, respectively, for the subjective experience metrics. The results showed that participant location significantly affected participants' perceived frustration, perceived effort, and perceived performance. India-based participants reported higher frustration than US-based participants (IN: $M = 3.64$, $SD = 1.98$ vs. US: $M = 3.01$, $SD = 1.68$; $F\,(1,\ 134) = 4.325,\ p = 0.039$), higher effort (IN: $M = 5.60$, $SD = 1.20$ vs. US: $M = 4.84$, $SD = 1.50$; $F\,(1,\ 134) = 6.338,\ p = 0.013$), and rated their work closer to perfect (IN: $M = 2.34$, $SD = 1.60$ vs US: $M = 2.92$, $SD = 1.43$; $F\,(1,\ 133) = 4.898,\ p = 0.029$). The interaction effect between participant location and payment method on perceived frustration was marginally significant $F\,(1,\ 134) = 3.272,\ p = 0.073$. India-based participants enrolled in the quota payment method reported higher frustration than US-based participants enrolled in the quota payment method (IN: $M = 4.03$, $SD = 2.03$ vs. US: $M = 2.80$, $SD = 1.79$; $p = 0.018$ after Bonferroni adjustment).

Table 6-4: Three-way ANCOVA results for subjective experience metrics:

Perceived frustration, effort, and performance.

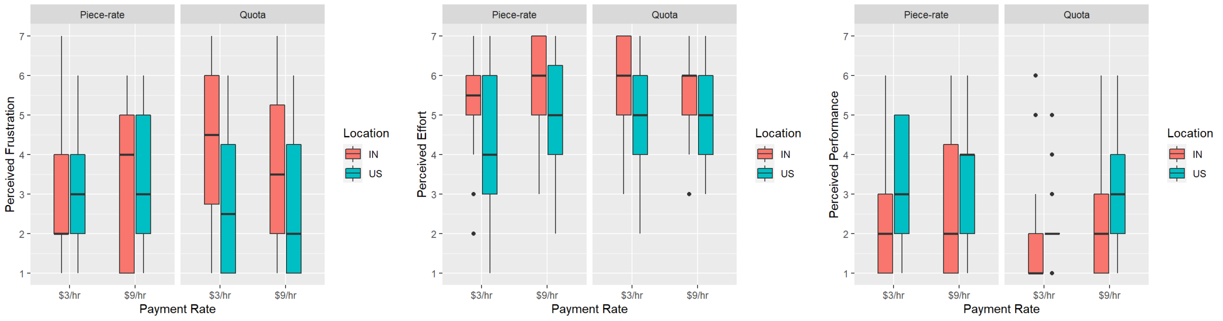
	Perceived frustration				Perceived effort				Perceived performance			
Outcome variables	df	F	*p*		df	F	*p*		df	F	*p*	
Gender	1	1.430	0.234		1	0.049	0.825		1	2.066	0.153	
Age	1	0.027	0.869		1	0.241	0.624		1	0.000	0.988	
Robot Familiarity	3	2.054	0.109		3	0.376	0.770		3	2.083	0.106	
Education Level	3	1.714	0.167		3	3.511	0.017	*	3	0.469	0.705	
Location	1	4.325	0.039	*	1	6.338	0.013	*	1	4.898	0.029	*
Payment Method	1	0.171	0.680		1	1.414	0.237		1	4.001	0.048	*
Payment Rate	1	0.005	0.944		1	2.701	0.103		1	6.404	0.013	*
L : PM	1	3.272	0.073	.	1	0.040	0.841		1	0.001	0.969	
L : PR	1	0.131	0.718		1	0.218	0.641		1	0.000	0.985	
PM : PR	1	0.655	0.420		1	1.565	0.213		1	1.106	0.295	
L : PM : PR	1	0.107	0.744		1	0.013	0.908		1	0.025	0.875	
Error	134				134				133			

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, · $p < 0.1$)

L = Location; PM = Payment Method; PR = Payment Rate

The main effects of payment rate *F* (1*,* 133) = 6.404*, p* = 0.013 and payment method *F* (1*,* 133) = 4.001*, p* = 0.048 on perceived performance were significant. Participants enrolled in the \$9/hour payment rate (*M* = 2.35, *SD* = 1.43) rated their task as more successful than participants enrolled in the \$3/hour payment rate (*M* = 2.90, *SD* = 1.60). Participants in the quota group (*M* = 2.41, *SD* = 1.52) rated their performance higher than those in the piece-rate group (*M* = 2.85, *SD*

= 1.54). There was no significant three-way interaction ($p > 0.10$) for perceived frustration, effort, and performance.



(a) Perceived frustration (b) Perceived effort (c) Perceived performance

Figure 6-2: Boxplots of subjective experience metrics: Perceived frustration, effort, and performance. Higher value indicates higher perceived workload.

6.3.3 *Motivation*

Table 6-5 and Figure 6-3 shows the ANCOVA outputs and boxplots, respectively, for the motivation metrics. When considering each motivation metric separately, the main effect of payment rate on the motivation to perform tasks for fun was nearly significant F (1, 134) = 3.614, p = 0.059. Participants who received $3/hour ($M$ = 4.86, SD = 1.90) rated motivation to do tasks for fun higher than participants who received $9/hour ($M$ = 4.31, SD = 1.92).

For the motivation to complete a task for fun, the interaction between participant location and payment rate (F (1, 134) = 2.959, p = 0.088) was marginally significant, and between participant location and payment method (F (1, 134) = 5.105, p = 0.025) was significant. The post-hoc test with Bonferroni adjustment method showed that the $3/hour-IN group ($M$ = 5.43, SD = 1.62) rated working for fun higher than India-based participants who were paid $9/hour ($M$ = 4.18, SD

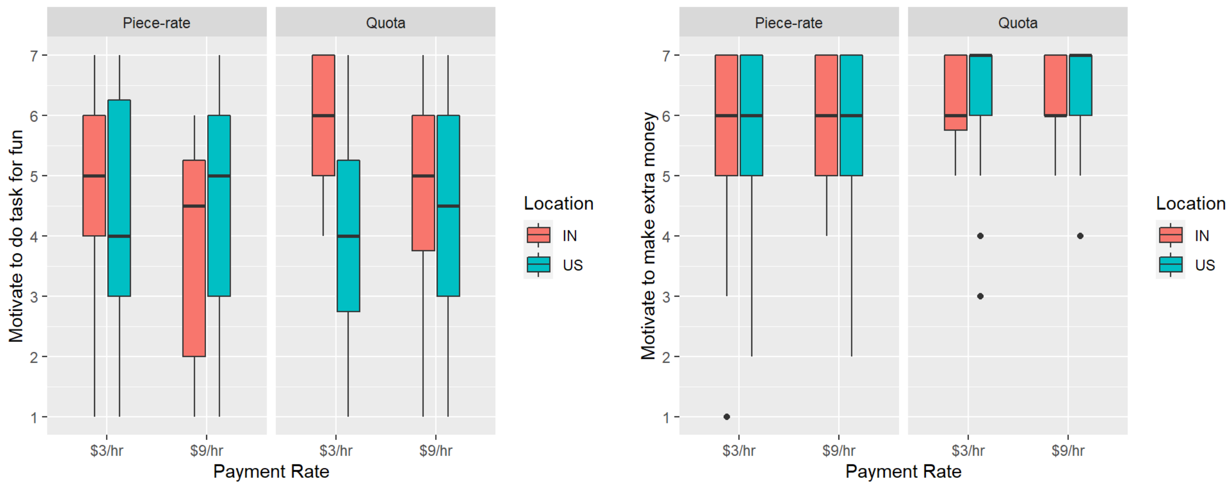
= 1.97, *p* = 0.021) and US-based participants who were paid $3/hour (*M* = 4.30, *SD* = 2.03, *p* = 0.050). India-based participants enrolled in the quota payment method group (*M* = 5.33, *SD* = 1.62) rated working for fun higher than did the India-based participants enrolled in the piece-rate payment method (*M* = 4.28, *SD* = 2.03, *p* = 0.086) and US-based participants enrolled in the quota payment method (*M* = 4.25, *SD* = 2.05, *p* = 0.073).

Table 6-5: Three-way ANCOVA results for participants' motivation: For fun, for money, and differences between the two.

	Motivation: for fun				Motivation: for money				Differences			
Outcome variables	df	F	*p*		df	F	*p*		df	F	*p*	
Gender	1	1.332	0.250		1	0.220	0.948		1	0.013	0.359	
Age	1	0.181	0.671		1	1.946	0.640		1	4.547	0.910	
Robot Familiarity	3	3.179	0.026	*	3	2.160	0.125		3	2.601	0.005	*
Education Level	3	1.118	0.344		3	0.122	0.096	.	3	1.553	0.055	.
Location	1	1.671	0.198		1	4.526	0.727		1	0.210	0.215	
Payment Method	1	0.571	0.451		1	0.076	0.035	*	1	2.927	0.647	
Payment Rate	1	3.614	0.059	.	1	0.343	0.783		1	4.686	0.089	.
L : PM	1	5.105	0.025	*	1	0.209	0.559		1	2.734	0.032	*
L : PR	1	2.959	0.088	.	1	0.381	0.649		1	0.468	0.101	
PM : PR	1	0.201	0.655		1	0.027	0.538		1	0.120	0.495	
L : PM : PR	1	0.101	0.751		1	0.220	0.869		1	0.025	0.729	
Error	134				134				134			

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, · $p < 0.1$)

L = Location; PM = Payment Method; PR = Payment Rate



(a) Motivation: For fun

(b) Motivation: For money

Figure 6-3: Boxplots of participants' motivation: For fun and for money.

The main effect of payment method on the motivation to make extra money was significant (*F* (1, 134) = 4.526, *p* = 0.035). Participants in the piece-rate payment method (*M* = 5.91, *SD* = 1.31) were motivated to complete tasks for money less than the participants in the quota payment method (*M* = 6.26, *SD* = 0.94). These results can be explained by the base payment, which was shown on the AMT website when the task was posted. The base payment of the piece-rate payment method would appear as a lower-paid task compared to the quota payment method and hence may have attracted participants who were focused less on making money compared to the quota payment method. There was no significant three-way interaction ($p > 0.10$) for motivation to complete tasks for fun or for money.

A paired t-test showed that participants rated their motivation to earn money (*M* = 6.09, *SD* = 1.15) higher than their motivation to complete tasks for fun (*M* = 4.59, *SD* = 1.93; *t*(159) = 8.02, $p < 0.001$). This result indicates that the average participants are more motivated to complete tasks for money than for pleasure, which supported findings from Antin and Shaw

(2012); Kaufmann et al. (2011); Litman et al. (2015). However, differences in motivation were not found between US-based and India-based participants that were found by Litman et al. (2015). An explanation could be that this trend has changed since then.

When considering the differences between motivation to make money vs. motivation to complete tasks for fun, the ANCOVA results revealed that the main effect of payment rate on the difference between motivation to make money vs. motivation to complete tasks for fun was marginally significant (F (1, 134) = 2.927, p = 0.089). Participants enrolled in \$3/hour condition (M = 1.16, SD = 2.30) had smaller differences in motivation ratings compared to participants enrolled in \$9/hour condition (M = 1.84, SD = 2.39). The model also found that the interaction effect between participant location and payment rate (F (1, 134) = 4.686, p = 0.032) was significant. After a post-hoc test with Bonferroni adjustment method, none of the groups was found to be significantly different from each other ($p > 0.10$). There was no significant three-way interaction ($p > 0.10$) for the differences between motivations. This difference in motivation metric ratings inferred that participants in \$9/hour condition feel more strongly about completing tasks for money than participants enrolled in \$3/hour condition.

6.3.4 Payment Fairness

Table 6-6 shows the output from Chi-square test of independence on perceived payment fairness. The chi-square test of independence showed a significant effect of payment rate ($\chi^2(DOF = 2, N = 160) = 6.769, p = 0.034$) on perceived payment fairness. \$9/hour rate was perceived as more fair compared to \$3/hour. This is an expected result; the payment rate of \$3/hour and \$9/hour were set to be \$3/hour less or more than the suggested amount of payment on the online platform (Martin et al., 2014; Salehi et al., 2015), respectively.

Table 6-6: Chi-square test of independence on perceived payment fairness.

Outcome variables	χ^2	df	p	
Location	4.016	2	0.134	
Payment Rate	6.769	2	0.034	*
Payment Method	0.301	2	0.860	

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, · $p < 0.1$

Then, if participants perceived the task as more fair, would it encourage participants to complete more trials in the piece-rate payment method? To answer this question, a two-way ANOVA was conducted. This analysis only includes Studies 3 and 4, in which the participants were paid using the piece-rate payment method. Surprisingly, the analysis did not show any significant effects of participant location and payment rate on the number of answers in the piece-rate payment method at $\alpha = 0.10$. See Table 6-7 for the two-way ANOVA output. The correlation was low between payment fairness and the number of tasks completed ($r(78) = 0.23$, $p = 0.32$). In a related work, Ikeda and Bernstein (2016) compared the task completion rates of the piece-rate payment method between two forms of payment, one that paid with Amazon credits and another that paid with coupons that discount the participant's phone bill. The coupon condition has a higher monetary value than Amazon's credit for the same amount of work. They found that participants who received Amazon credits per task had a higher number of tasks completed per person than the coupon condition. These results may have been affected by the utility of the reward and the participants that each condition attracts.

Table 6-7: Two-way ANOVA results: effects of payment rate and participant location on number of answers in the piece-rate payment method.

	Number of answers		
Outcome variables	df	F	*p*
Location	1	0.118	0.733
Payment Rate	1	0.174	0.678
Location : Payment Rate	1	0.483	0.489
Error	76		

6.4 Discussion

The results showed that US-based participants had a higher rate of following the written task description than India-based participants. Task accuracy was comparable between the quota and piece-rate payment methods. However, because a piece-rate task appeared as a task with a much lower payment than a quota task, a higher-pay piece-rate task attracted more fraudulent participants. This is an issue that researchers should be aware of and actively filter out. A higher payment rate was not correlated with the number of tasks completed by the participants.

Some of the findings of this study are also different from the related literature. Our results for participants' motivation showed that participants were motivated to complete tasks for money more than for pleasure, similar to the findings of Antin and Shaw (2012); Kaufmann et al. (2011); Litman et al. (2015), but there was no statistically significant difference between US-based and India-based participants, in contrary to the findings of Litman et al. (2015). Our result did not show statistically significant difference between different payment rates on the number of tasks completed in the piece-rate condition, which was observed by Ikeda and Bernstein (2016). Participants' motivation on AMT may also change over time. Future studies on AMT should report the time period of data collection, estimated task length, and payment rate so that participants' motivation and subjective experience can be compared between different studies and years.

There are some limitations of the current study, which point to the need and questions for future research. First, the results of the present studies were obtained using AMT participants enrolled in this study. Different base payment rates may attract participants with different motivations to complete the task. Hence, these results may only apply to the payment levels that were studied and additional levels of payment rate may need to be studied to apply the current findings more broadly. Second, the result of this work may not be suitable for task situations that require each participant to complete a fixed number of trials or conditions. This work is more suitable for tasks that allow flexibility on the researcher's side to accept an unequal number of observations from each participant, such as image/video labeling.

6.5 Summary

Chapter 6 discusses a series of studies that investigated the effects of participant location, payment method, and payment rate on online crowdwork participants' task quality, subjective experience, motivation, and perceived payment fairness.

6.6 References

Antin, J., & Shaw, A. (2012). Social desirability bias and self-reports of motivation: A study of amazon mechanical turk in the US and India. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2925–2934. https://doi.org/10.1145/2207676.2208699

Chandler, D., & Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. Journal of Economic Behavior & Organization, 90, 123–133. https://doi.org/10.1016/j.jebo.2013.03.003

d'Eon, G., Goh, J., Larson, K., & Law, E. (2019). Paying Crowd Workers for Collaborative Work. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–24. https://doi.org/10.1145/3359227

Difallah, D., Filatova, E., & Ipeirotis, P. (2018). Demographics and dynamics of mechanical Turk workers. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 135–143. https://doi.org/10.1145/3159652.3159661

Dynamo. (2014). Guidelines for Academic Requesters (Version 1.1). https://irb.northwestern.edu/docs/guidelinesforacademicrequesters-1.pdf

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In Advances in psychology (Vol. 52, pp. 139–183). Elsevier. https://doi.org/10.1016/S0166-4115(08)62386-9

Ikeda, K., & Bernstein, M. S. (2016). Pay it backward: Per-task payments on crowdsourcing platforms reduce productivity. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 4111–4121. https://doi.org/10.1145/2858036.2858327

Kaufmann, N., Schulze, T., & Veit, D. (2011). More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk. Amcis, 11(2011), 1–11.

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. (2020). The shape of and solutions to the MTurk quality crisis. Political Science Research and Methods, 8(4), 614–629. https://doi.org/10.1017/psrm.2020.6

Litman, L., Robinson, J., & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US-and India-based workers on Mechanical Turk. Behavior Research Methods, 47(2), 519–528. https://doi.org/10.3758/s13428-014-0483-x

Martin, D., Hanrahan, B. V., O'Neill, J., & Gupta, N. (2014). Being a turker. Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, 224–235. https://doi.org/10.1145/2531602.2531663

Salehi, N., Irani, L. C., Bernstein, M. S., Alkhatib, A., Ogbe, E., & Milland, K. (2015). We are dynamo: Overcoming stalling and friction in collective action for crowd workers. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 1621–1630. https://doi.org/10.1145/2702123.2702508

Chapter 7
Conclusions

7.1 Dissertation Summary

In collaborative work, understanding the intention of the working partners should enhance the fluency of collaboration. However, this is still a work in progress in human-machine teaming. One of the problems is how to make a robot's target easier to predict for humans working collaboratively with a robot. This dissertation describes a series of studies investigating robot's predictability and perceived human-likeness as a function of trajectory characteristics, in search of a method to enhance human-robot collaboration by guiding the design of trajectory to improve robot's action predictability and human-likeness. Related literature suggested that some of the optimal feedback control laws should be included in the investigation to design a human-centric robot's trajectory that supports human prediction of robot's intentions.

The findings discussed in Chapter 3 point out that a trajectory that used similar path planning strategies to humans, such as minimizing energy expenditure, may enable the observer to predict the robot's target quicker. The end-effector of the robot might be where the observers focus when observing a robot's motion more than the other parts of the arm. A trajectory with a lower change in acceleration in the end-effector was perceived as safer. The result of the studies presents in this study agree with the insight from related work that other features, outside of efficiency, may help an observer to make a correct prediction quicker.

In parallel with the first research question, an additional study was conducted to investigate whether introducing changes to the format of interaction would affect participants' perception of the robot's human-likeness and its perceived safety. The results of a study described in Chapter 4 showed that participants rated the robot's physical movements as less human-like and less life-like compared to watching the robot through the video. However, participants also rated their emotional state as feeling safer when they interacted face-to-face with the robotic system rather than through video. Chapter 4 offers empirical evidence that there are differences between the perception of the robot between two different interaction formats, observing a robot's physical movements vs. their videos. However, the latter were generally adopted for online research conducted on online crowdwork platforms. Hence, the second research objective described in this dissertation work is to examine the effects of (1) formats of interaction and (2) participant pools on task quality, subjective experience, and perceived payment fairness.

Chapter 5 reports two studies that investigated the similarities and differences in task task quality, subjective experience, motivation, and perceived payment fairness between observing a physical movement and watching videos in the context of HRI judgment tasks. The findings of two studies revealed that students may produce higher-quality data than crowdworkers. Observing a robot's physical movements yielded better task quality than their video recordings. Students were also motivated to complete a study out of interest while the crowdworkers were more motivated by the task's compensation. While the differences between students and online crowdworkers were significant, the differences in task quality, subjective experience, motivation, and perceived payment fairness between different groups of online crowdworkers were less clear.

Chapter 6 reports a series of studies that investigated the effects of participant location, payment method, and payment rate on online crowdwork participants' task quality, subjective experience, motivation, and perceived payment fairness. Four studies were conducted on an online crowdwork platform. Location was found to significantly affect the likelihood in which participants follow written task descriptions for an online task. A piece-rate task attracted more fraudulent participants compared to a quota task on the same payment rate. This might be because a piece-rate task appeared on the AMT as a lower-pay task than a quota task that pays at the same rate. The number of tasks completed by the participants was not found to be positively correlated with payment rate. The participants were not motivated to complete more tasks even though they perceived the rate of payment as more fair. These dissociations between payment rate, perceived payment fairness, and task performance are important findings that online work researchers and practitioners should be aware of. Participants on AMT were motivated to complete tasks for money more than for pleasure, similar to the findings of related literature, but there was no statistically significant difference between US-based and India-based participants. The result did not show a statistically significant difference between different payment rates on the number of tasks completed in the piece-rate payment method.

Scientific Merits

The work described in this dissertation focuses on understanding how to design a robot's movements to exhibit intent in reach movement. Although this work does not provide the exact example of a movement that can achieve quicker prediction and higher accuracy, it can serve as a starting point for future work on trajectory characteristics that should be incorporated when designing a robot's movements. The results can inform future researchers on how to produce a predictable and human-like movement as perceived by those interacting with a robotic arm, both directly and through a digital screen.

The comparability of results between online and in-lab studies in an HRI context was also investigated. The results of this work illustrated that there is a difference between findings, task quality, and subjective experience between in-lab and online experimental studies. In-lab and online experimental studies are different from each other in terms of participant pools, stimuli that can be present to participants, and other environmental factors such as surrounding noises or interruptions. Equating both online and in-lab experimental studies is challenging, as both have different trade-offs. To make the in-lab participants more comparable to the general population being investigated, researchers should consider recruiting participants from their local community to participate in the same study and comparing the findings with that of college students to investigate the effects of different age ranges and education levels on the research outcome. To control potential confounding factors that occur from the differences in formats of interaction, both in-lab and online versions should utilize the same survey engine for training and data collection. For example, in the context of an HRI judgment task on a robot's movements, researchers should opt for a two-dimensional task context as it is more comparable between observing physical movements and videos. If the data collection requires time-sensitive input, alternative methods of data collection such as asking online participants to install software that can run natively on the participant’s machine may have better control of the interface as well as the quality of stimuli being presented.

Compensation issues on online crowdsourcing platforms were also investigated. The results complement and enrich the related literature on the effects of payment methods, rates, and participants’ locations on task quality, subjective experience, motivation, and perceived payment fairness in the context of an online crowdwork platform. The piece-rate payment method was systematically compared to the default quota payment method in the work described in this dissertation.

Practical Impacts

This work can help to guide the design of movements to enhance the interaction between humans and robotic arms, such as making the robot's intent more predictable and making the workers feel safer, thereby improving the acceptance of technology. The context being studied in this work can apply to manufacturing tasks as well as robotics arms deployed in other sectors, such as healthcare or services. Human-likeness and life-likeness have been one of the focus in the social robotics field. The findings from this study may be able to inform the decision to design movements of the robot intended for service industries.

Moreover, these findings benefit the online crowdwork platform owners, requesters, and workers. Online crowdwork platform owners can use the findings of this dissertation to improve payment policies for the benefit of the requesters and workers. Providing proper payment to online crowdworkers based on the workers' motivation to complete a task may induce a more positive relationship between requesters and workers.

The online crowdwork platform users (both requesters and workers) may become more informed of the fairness issue on the platforms, which may provoke more discussion to find a better solution that makes online crowdwork platforms a more accessible workplace. Further investigations of this work will deepen our understanding of workers in online crowdwork platforms, which will guide human behavioral researchers to design better methods for distributing payments to participants from various backgrounds and socioeconomic levels. The findings may help fill the gaps in the research literature regarding payment fairness in an online market, where requesters could set a fair price for online participants.

7.2 Limitations and Future Work

The findings of this work show trends of trajectory characteristics on prediction time and accuracy. However, they do not provide the exact example of a movement that can achieve quicker prediction and higher accuracy. Future work should study the effects of specific trajectory characteristics found to correlate with prediction performance to produce specific examples of trajectory. The experimental design may need to be constrained to similar paths for all targets. Independent variables can be the paths' energy expenditure, speed, and jerkiness, each with several levels (e.g., low, medium, and high). Note the challenges will be in controlling the other trajectory characteristics that correlate with the independent variables. For example, an increase in speed can lead to shorter path length, which can be difficult to control across different conditions. Additionally, future research should perform a time-series analysis to look for features that occur right before the observer makes a prediction or changes from one prediction to another. This will tell us what are the features of the movement that an observer takes as a clue to infer the robot's intention. There is also a limited number of public datasets related to this specific topic of human interpreting robot's target. Future datasets related to this topic should be made available so the comparison between different literature can be studied further.

Further work should incorporate physiological measurements of mental stress and compare them with Godspeed questionnaire's (Bartneck et al., 2009) perceived safety score. The perceived safety score described in this study was measured using three 5-gradation semantic-differential scales from Godspeed questionnaire asking participants to "rate [the participant] emotional state on these scales": Anxious-Relaxed, Agitated-Calm, and Quiescent-Surprise. These subjective measures can be collected in parallel with other physiological indicators of mental stress measurements such as higher galvanic skin response duration (Kim et al., 2020), higher blinking frequency (Giannakakis et al., 2017), higher heart rate (Giannakakis et al., 2017), lower inter-beat interval measured in EKG (Sloan et al., 1994), and lower EEG power spectral density in the alpha band (Al-Shargie et al., 2015).Physiological indicators of stress can also be utilized to analyze temporal aspect of the movements that correlate with an increase in users' mental stress.

Studies described in this dissertation work controlled the perspective from which the participants observed the robot, which may lack realism. All participants observed the scene where they see the robot and all objects on the shelf and stay outside of the robot's reachable workspace. This is to ensure that the viewing angle is controlled and does not confound the findings. However, in a realistic setting, the observers can move around to adjust their position so that they can observe the robot's movement better. Similarly, studies with video conditions described in this dissertation do not contain sounds to ensure that the presence of the robot's sounds is a controlled variable for online studies, as opposed to allowing participants to control whether to mute or turn on the sounds. However, this introduces uncertainty whether the difference observed in the findings was from the lack of auditory information or the depth of information. Both pieces of information may help increase the task quality as the sound of the robot's motor can indicate whether the robot is accelerating or slowing down, and depth information helps judge the robot's movement quality in three-dimensional space. Future work should be conducted to investigate further which information affects judgment task quality.

There is also a limitation on the response collection method. In this series of studies, the participant used a computer-based interface to input predictions. Once the participant had a prediction, they would have to look at the screen, move the mouse to the correct button, and click. Busch et al. (2017) used a push of a physical button to input predictions. This requires a participant to look down from the robot, search for the button, and reach for the button. Dragan et al. (2015) used the time that the participant declared aloud or started to reach for a correct object in preparation for an impending collaborative task, whichever comes first. This talk-aloud method may have the shortest time between the prediction to recorded time. These different ways of inputting participants' responses may affect the time between the actual predictions and the time the research team recorded the prediction time.

The studies described in this dissertation only used an HRI judgment task of a robot's movement in three-dimensional space as the context. The results showed that formats of interaction between a robot's physical movements and their videos produced different ratings in perceived anthropomorphism, perceived animacy, and perceived safety. Future work could explore other contexts such as two-dimensional movements of an autonomous ground vehicle (both holonomic and non-holonomic).

The results of the online studies were obtained using the AMT platform, which is a platform among many online crowdwork platforms available (e.g., Qualtrics, CrowdFlower, Prolific, etc.). This platform, although widely used in research studies, may not be the best representation for other online crowdwork platforms. Different base payment rates as well as platforms may attract participants with different motivations to complete online tasks. Hence, these results may only apply to the payment levels that were studied and only on the AMT platform. Additional payment rates and online crowdwork platforms should be studied to test the current findings more broadly. Future studies on online crowdwork platforms should report the time period of data collection, estimated task length, and payment rate so that participants' motivation and subjective experience can be compared between different studies and years.

Finally, the results on the piece-rate payment method demonstrated in this dissertation work may not be suitable for task contexts that require each participant to complete a fixed number of trials or conditions. The piece-rate payment method is more suitable for tasks that allow flexibility on the researcher's side to accept an unequal number of observations from each participant, such as image/video labeling. Future work should explore additional payment methods such as pay-by-performance or pay-by-hour to the quota and piece-rate payment methods.

7.3 Summary

This section offers an overview of the research questions being investigated, the findings of the study on these research questions, and potential contribution offered by this dissertation work. Finally, Chapter 7 discusses limitations and future research directions.

7.4 References

Al-Shargie, F., Tang, T. B., Badruddin, N., & Kiguchi, M. (2015). Simultaneous measurement of EEG-fNIRS in classifying and localizing brain activation to mental stress. 2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 282–286. https://doi.org/10.1109/ICSIPA.2015.7412205

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. International Journal of Social Robotics, 1(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Busch, B., Grizou, J., Lopes, M., & Stulp, F. (2017). Learning legible motion from human–robot interactions. International Journal of Social Robotics, 9(5), 765–779. https://doi.org/10.1007/s12369-017-0400-4

Dragan, A. D., Bauman, S., Forlizzi, J., & Srinivasa, S. S. (2015). Effects of robot motion on human-robot collaboration. 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 51–58. http://doi.org/10.1145/2696454.2696473.

Giannakakis, G., Pediaditis, M., Manousos, D., Kazantzaki, E., Chiarugi, F., Simos, P. G., Marias, K., & Tsiknakis, M. (2017). Stress and anxiety detection using facial cues from videos. Biomedical Signal Processing and Control, 31, 89–101. https://doi.org/10.1016/j.bspc.2016.06.020

Kim, J., Park, J., & Park, J. (2020). Development of a statistical model to classify driving stress levels using galvanic skin responses. Human Factors and Ergonomics in Manufacturing & Service Industries, 30(5), 321–328. https://doi.org/10.1002/hfm.20843

Sloan, R., Shapiro, P., Bagiella, E., Boni, S., Paik, M., Bigger Jr, J., Steinman, R., & Gorman, J. (1994). Effect of mental stress throughout the day on cardiac autonomic control. Biological Psychology, 37(2), 89–99. https://doi.org/10.1016/0301-0511(94)90024-8