Gesture-Based Nonverbal Interaction for Exercise Robots

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Mayumi Mohan
aus Kerala/Indien

Tübingen
2023

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	09.10.2023
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Katherine J. Kuchenbecker
2. Berichterstatter/-in:	Prof. Dr. Martin A. Giese

Gesture-Based Nonverbal Interaction for Exercise Robots

Abstract

The emergence of standalone exercise-coach robotic systems is an exciting prospect in the pursuit of reliable ways to help people learn new physical skills and maintain fitness. While socially assistive robots have demonstrated potential as exercise coaches, real-world deployment remains a challenge. Central to this issue is the topic of gesture-based interactions. When teaching or coaching, humans augment their words with carefully timed hand gestures, head and body movements, and facial expressions to provide feedback to their students. Robots, however, rarely utilize these complementary cues. A minimally supervised social robot equipped with these abilities could support people in exercising, physical therapy, and learning new activities. This thesis poses the question of how the intuitive power of human gestures can be harnessed to enhance human-robot interaction. In an effort to answer this query, the presented research explores gesture-based interactions to expand the capabilities of a socially assistive robotic exercise coach, investigating the perspectives of both novice users and exercise-therapy experts.

This thesis begins by concentrating on the user's engagement with the robot, analyzing the feasibility of minimally supervised gesture-based interactions. This exploration seeks to establish a framework in which robots can interact with users in a more intuitive and responsive manner. The investigation then shifts its focus toward the professionals who are integral to the success of these innovative technologies: the exercise-therapy experts. Roboticists face the challenge of translating the knowledge of these experts into robotic interactions. We address this challenge by developing an algorithm that can enable exercise therapists to create customized gesture-based interactions for a robot, thereby bridging the knowledge gap. This thesis is thus divided into four parts, as follows.

The initial part of this thesis focuses on the development of the Robot Interaction Studio, a cutting-edge, minimally supervised environment designed to facilitate social-physical interactions with an exercise robot. By leveraging markerless motion-capture technology, the system is capable of obtaining extensive quantitative information from a user as they interact with the robot. A within-subjects study with seven users was conducted to evaluate this setup, employing three distinct gesture-based interactions rooted in commonly observed social behaviors. Traditional exercise coaching systems often require heavy supervision by technical experts and can be cumbersome and time-consuming to set up, lacking the ability to provide sufficient information about user movement. This innovative system overcomes these limitations by generating rich, quantitative data that could enable more personalized interactions.

The second part of this thesis shifts the focus to the robot's ability to provide meaning-

ful feedback to the user. This phase also demonstrates how the Robot Interaction Studio can foster dynamic gesture-based interactions with a robot. Drawing inspiration from time-tested techniques found in the educational literature, we designed two distinct feedback paradigms for the robot, formative feedback and summative feedback, which are evaluated in the Robot Interaction Studio using the gesture-based cues developed in the first part of this thesis. Through a mixed-methods-design study with 28 participants, we demonstrate that robots can effectively utilize gesture-based feedback to enhance user performance and understanding, offering a novel approach to increase user engagement during physical activities with robots.

The third section focuses on the professionals who are responsible for crafting exercise techniques for patients. A significant challenge is enabling the customizability of exercise-coach robots for use by exercise-therapy experts. Existing robots, often hard-programmed, cannot be easily modified by non-roboticists. We propose that motion-capture-based teleoperation could be used to resolve this issue by allowing an exercise expert to move their own body to create motions for a robot's arms and head. Thus, we utilize semi-structured interviews and qualitative surveys to understand the perspectives of eight experts on exercise robots and their teleoperation. We illustrate how infusing customizability through teleoperation can notably enhance experts' perception of robots, aligning the technology more closely with their needs.

Lastly, the thesis turns to the development of an intuitive system that therapy experts could readily employ to create robot arm motions. Guided by the fact that experts are open to using teleoperation, the system is founded on motion-capture-based kinematic retargeting, an approach shown to be highly intuitive for human operators. An optimization-based algorithm named OCRA was created and thoroughly evaluated; the user can customize its behavior by setting one weight that dictates the relative importance of hand orientation error versus the shape of the arm in space, which we term arm skeleton error. It exhibits versatility across robots with varying limb lengths and degrees of freedom, provided they have revolute joints. A detailed study in which 70 users watched videos of a robot executing four operator motions each at four weight values demonstrates the algorithm's ability to replicate human-like motions. Rigorous benchmarking and validation further emphasize OCRA's robustness, making it a promising tool for creating robotic gestures.

Thus, this thesis lays the groundwork for dynamic gesture-based interactions in minimally supervised environments, with implications for not only exercise-coach robots but also broader applications in human-robot interaction.

Zusammenfassung

Das Aufkommen eigenständiger Trainingsrobotersysteme ist eine aufregende Entwicklung, um Menschen zuverlässige Möglichkeiten beim Erlernen neuer motorischer Fähigkeiten und bei der Erhaltung ihrer Fitness zu bieten. Sozial unterstützende Roboter haben bereits ihr Potenzial als Fitnesstrainer unter Beweis gestellt, doch der Einsatz in der realen Welt bleibt eine Herausforderung. Im Mittelpunkt steht dabei das Thema der gestenbasierten Interaktion. Wenn Menschen unterrichten oder trainieren, ergänzen sie ihre Worte mit sorgfältig getimten Handgesten, Kopf- und Körperbewegungen und Gesichtsausdrücken, um ihren Schülern Feedback zu geben. Roboter hingegen nutzen diese ergänzenden Hinweise nur selten. Ein minimal überwachter unterstützender Roboter, der mit diesen Fähigkeiten ausgestattet ist, könnte Menschen beim Sport, bei der Physiotherapie und beim Erlernen neuer Aktivitäten unterstützen. Diese Arbeit untersucht, wie die intuitive Wirkung menschlicher Gesten genutzt werden kann, um die Mensch-Roboter Interaktion zu verbessern. Dafür thematisiert diese Dissertation gestenbasierte Interaktionen, um die Fähigkeiten eines sozial unterstützenden robotischen Bewegungstrainers zu erweitern. In diesem Zusammenhang werden die Perspektiven sowohl von unerfahrenen Nutzern als auch von Experten für Bewegungstherapie eingebracht.

Diese Dissertation setzt den ersten Schwerpunkt auf die Interaktion des Benutzers mit dem Roboter und analysiert die Machbarkeit von minimal überwachten gestenbasierten Interaktionen. Diese Untersuchung zielt darauf ab, einen Rahmen zu schaffen, in dem Roboter auf eine intuitivere und reaktionsfreudigere Weise mit dem Benutzer interagieren können. Die Untersuchung verlagert dann ihren Schwerpunkt auf die Fachleute, die für den Erfolg dieser innovativen Technologien unerlässlich sind: die Experten für die Bewegungstherapie. Robotiker stehen vor der Herausforderung, das Wissen dieser Experten in Roboterinteraktionen zu übersetzen. Wir gehen dieser Herausforderung nach, indem wir einen Algorithmus entwickeln, der es der Bewegungstherapie ermöglicht, maßgeschneiderte gestenbasierte Interaktionen für einen Roboter zu erstellen und so die Wissenslücke zu schließen. Diese Arbeit ist daher wie folgt in vier Teile gegliedert.

Der erste Teil dieser Arbeit konzentriert sich auf die Entwicklung des „Robot Interaction Studio“, einer hochmodernen, minimal überwachten Umgebung, die sozial-physische Interaktionen mit einem Trainingsroboter ermöglicht. Durch den Einsatz von markerloser Motion-Capture-Technologie ist das System in der Lage, umfangreiche quantitative Informationen von einem Benutzer zu erhalten, während dieser mit dem Roboter interagiert. Zur Evaluierung dieses Systems wurde eine Nutzerstudie mit sieben Teilnehmern und drei unterschiedlichen gestenbasierten Interaktionen durchgeführt, die auf

allgemein beobachteten sozialen Verhaltensweisen beruhen. Herkömmliche Trainings-Coaching-Systeme erfordern oft eine intensive Betreuung durch technische Experten, sind umständlich und zeitaufwändig in der Einrichtung und können keine ausreichenden Informationen über die Bewegungen der Benutzer liefern. Dieses innovative System überwindet diese Einschränkungen, indem es reichhaltige, quantitative Daten generiert, die eine persönlichere Interaktion ermöglichen könnten.

Der zweite Teil dieser Arbeit verlagert den Schwerpunkt auf die Fähigkeit des Roboters, dem Benutzer sinnvolles Feedback zu geben. In dieser Phase wird auch demonstriert, wie das „Robot Interaction Studio“ dynamische gestenbasierte Interaktionen mit einem Roboter fördern kann. Wir haben uns von bewährten Techniken aus der pädagogischen Literatur inspirieren lassen und zwei verschiedene Feedback-Paradigmen für den Roboter entwickelt: formatives Feedback und summatives Feedback, die beide im „Robot Interaction Studio“ anhand der im ersten Teil dieser Arbeit entwickelten gestenbasierten Hinweise ausgewertet werden. Anhand einer Studie mit kombinierten Methoden und 28 Teilnehmern zeigen wir, dass Roboter gestenbasiertes Feedback effektiv nutzen können, um die Leistung und das Verständnis des Benutzers zu verbessern. Dies bietet einen neuartigen Ansatz zur Steigerung des Engagements des Benutzers bei körperlichen Aktivitäten mit Robotern.

Der dritte Teil konzentriert sich auf die Fachleute, die für die Ausarbeitung von Übungstechniken für Patienten verantwortlich sind. Eine große Herausforderung ist die Anpassbarkeit von Trainingsrobotern, um diese für Experten der Bewegungstherapie zugänglich zu machen. Vorhandene Roboter sind oft eingeschränkt programmiert und können von Nicht-Robotikern nicht leicht verändert werden. Wir schlagen vor, dass dieses Problem durch Motion-Capture-unterstützte Teleoperation gelöst werden kann. Daher nutzen wir halbstrukturierte Interviews und qualitative Umfragen, um die Sichtweise von acht Experten auf Trainingsroboter und deren Teleoperation zu verstehen. Wir zeigen auf, wie die Anpassbarkeit durch Teleoperation die Wahrnehmung von Robotern durch Experten deutlich verbessern kann, indem die Technologie besser auf ihre Bedürfnisse abgestimmt wird.

Zuletzt widmet sich diese Dissertation der Entwicklung eines intuitiven Systems, das Therapieexperten zur Erstellung von Roboterarmbewegungen verwenden könnten. Ausgehend von der Tatsache, dass Experten der Teleoperation gegenüber aufgeschlossen sind, basiert das System auf dem Motion-Capture-basierten kinematischen Retargeting. Dieser Ansatz dient dazu, Bewegungen von einem Modell auf ein anderes zu übertragen, welches sich für menschliche Bediener als äußerst intuitiv erwiesen hat. Es wurde ein optimierungsbasierter Algorithmus mit dem Namen OCRA entwickelt und gründlich evaluiert. Der Benutzer kann sein Verhalten anpassen, indem er eine Gewichtung festlegt, dass die relative Relevanz des Handorientierungsfehlers im Vergleich zur Form des Arms im Raum bestimmt, welches wir als Arm-Skelett-Fehler bezeichnen. Er ist vielseitig einsetzbar bei Robotern mit unterschiedlichen Gliedmaßenlängen und Freiheitsgraden, vorausgesetzt, sie verfügen über Drehgelenke. Eine detaillierte Studie, bei der 70 Teilnehmer Videos von einem Roboter sahen, der vier Bewegungen jeweils mit vier Gewichtungen

ausführte, zeigt die Fähigkeit des Algorithmus, menschenähnliche Bewegungen nachzubilden. Rigoroses Benchmarking und Validierung unterstreichen die Robustheit von OCRA und machen ihn zu einem vielversprechenden Werkzeug für die Erstellung von Robotergesten.

Somit legt diese Arbeit den Grundstein für dynamische gestenbasierte Interaktionen in minimal überwachten Umgebungen, was nicht nur für Trainingsroboter, sondern auch für breitere Anwendungen in der Mensch-Roboter-Interaktion von Bedeutung ist.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my Ph.D. advisor, Katherine J. Kuchenbecker. Your guidance has not only shaped me as a researcher but also served as a wonderful role model. Your teachings on how to question and think critically, your encouragement towards independence, and your challenges that pushed me to reach new heights have made me a better researcher. I am truly fortunate to have had the privilege of working with you. No amount of words can express my appreciation for your support through all the hurdles during my Ph.D.

My special appreciation goes to Martin A. Giese and Michael J. Black, who have both been integral parts of my IMPRS-IS Thesis Advisory Committee (TAC). Thank you both for taking the time to join my TAC meetings and provide thoughtful advice. Your insights have been invaluable in shaping my research and thinking.

To my family, who have endured long periods of my absence and for whom I have become a virtual rather than physical presence. Your sacrifices have not gone unnoticed, and I often wished that my research on teleoperation could miraculously transform into teleportation. To my husband, Chitvan Killawala: you have been my lifeline. Your understanding and support when I made the decision to move to Germany, even though it meant an ocean would separate us, along with your engaging technical discussions on mathematical concepts, and your unwavering companionship during our parallel Ph.D. journeys, have been the backbone of my experience, and they mean more to me than words can express. To my father, Mohan Das Babu P and mother, Sapna V Roshni: your continual encouragement to aim higher has been the foundation of my strength. Your belief in my abilities has always been the wind in my sails. Thank you both for being my biggest supporters and the incredible parents that you are; your impact on my life is immeasurable. I also want to extend my heartfelt thanks to my grandmother, Govinda Meera Bai, whose endless love and consistent concern for my well-being have been a constant source of comfort and encouragement. I must also thank my in-laws, Viplav Killawala and Paraginee Killawala. Your love and support, and your role as another set of parents, have been comforting and uplifting.

I would like to extend a heartfelt thank you to Haliza Mat Husin, who has been much more than a colleague. Haliza, you have been a friend and a big sister to me. Your endless patience and dedication in spending countless hours teaching me statistics, meticulously discussing study designs, and dissecting data with me have been invaluable. Thank you for always being ready to lend a supportive ear when I needed it most. I feel absolutely blessed to have met you.

A few other colleagues deserve my thanks for their exceptional support. Bernard Javot

and Joey Burns, your assistance with moving a massive motion-capture setup four and a half times helped me create the Robot Interaction Studio. Joey, thank you for patiently spending hours debugging network issues with me. I also want to thank Michaela Wieland. Your support with my many lab moves has been invaluable.

This adventure has bestowed upon me several cherished friendships that I hold dear. My special thanks go to Andria Smith, Farimah Fazlollahi, Natalia Egana Rosa, Saekwang Nam and Behnam Khojasteh, who have each contributed uniquely to my experience. Andria, I fondly remember all the wonderful dinners, filled with laughter and conversations, and I'm grateful for those shared moments. Farimah, how I wish you were in Stuttgart so that we could share more cups of coffee and engaging conversations. Saekwang, thank you for your amazing and supportive friendship. Nati, your support in the lab has been invaluable, but it's your wonderful friendship, filled with understanding and encouragement, that I treasure most. Behnam, thank you for all the conversations over cups of coffee. Thank you, additionally, for editing the German version of the abstract of this thesis.

I express my gratitude to the Max Planck Society, the International Max Planck Research School for Intelligent Systems (IMPRS-IS) and IntCDC for their financial support. MPI-IS is truly a haven for research. The facilities, people, and freedom have been unparalleled, and I am deeply thankful. My particular thanks go to Leila Masri and Sara Sorce for being the backbones of the IMPRS-IS Ph.D. program, with special mention to Leila for her encouragement and support. I would also like to thank Maike Trost and Dorothee Klink for their immense support in making my stay in Germany seamless.

Finally, I wish to acknowledge many other current and former members of the Haptic Intelligence Department. Thank you Maria Paola Forte, Yijie Gong, Ilona Jacobi, Rachael L'Orsa, Rachael Burns and Ravali Gourishetti. A special thank you to Yijie for her invaluable assistance with setting up and testing my motion-capture devices, and for her wonderful friendship. I also want to thank Naomi T. Fitter, who has been a wonderful collaborator. I enjoy doing research and organizing workshops with you.

The cumulative effort, love, and support of all these individuals have shaped my academic journey. I remain indebted to each one of you for making this not only an intellectual pursuit but also an amazing experience.

And lastly, this dissertation could not have been finished without the cooperation of several other close colleagues. The following are the acknowledgments for each of the research chapters in this dissertation:

- **Chapter 2**: I appreciate the significant contributions of Cara M. Nunez and Katherine J. Kuchenbecker for helping design the user study, guiding the creation of the figures, writing short sections of text, and extensively editing the manuscript. Additionally, Katherine J. Kuchenbecker provided feedback on the data analysis, design of the system and user interaction. I also thank Bernard Javot, Joey Burns, and Nils Hasler for their invaluable technical support on this project and also express gratitude to Michaela Wieland, Naomi Tashiro, Behnam Khojasteh, Eric Young,

Farimah Fazlollahi, Hojin Lee, and Metin Sitti. I also thank the National Science Foundation Graduate Research Fellowship Program (NSF GRFP) for supporting Cara M. Nunez.

•**Chapter 3**: I appreciate the significant contributions of Katherine J. Kuchenbecker in this project all the way from study design, data interpretation, guidance in developing figures and preparation of text. I also thank Cara M. Nunez for providing input on study design and preparation of the manuscript. I would like to thank Shari L. Young Kuchenbecker for extensive advice regarding modes of assessment in education and evaluation of user attention; Haliza Mat Husin and Phoenix Von Wagoner for invaluable guidance on statistical analysis; Chitvan Killawala for guidance on designing the pose cue error; Joey Burns, Bernard Javot, and Nils Hasler for technical support on this project; Ilona Jacobi for translating the consent form into German; Naomi Tashiro and Maria Paola Forte for support with system testing; Ben Richardson for guidance on the use of the Wasserstein metric; Lijuan Wang for grading user responses; and Andrew Schulz for feedback on figure layouts.

•**Chapter 4**: I appreciate the significant contributions of Haliza Mat Husin and Katherine J. Kuchenbecker for helping identify the problem, helping design the user study and providing input on associated analyses and interpretation of the data. A special thanks to Haliza for co-conducting all the interviews and supporting the logistics of the study.

•**Chapter 5**: I appreciate the significant contributions of Katherine J. Kuchenbecker in algorithm and study design, data analysis and interpretation and figure creation. This work is partially supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2120/1 - 390831618. I would also like to express my gratitude towards Chitvan Killawala, Ivan Oreshnikov, Gokhan Serhat, Maria Paola Forte and Behnam Khojasteh. I would also like to thank Towela Tembo for teleoperating the robot and permitting us to use the videos.

Contents

1 Introduction **1**
- 1.1 Overall Research Questions . . . 3
- 1.2 Outline . . . 3
- 1.3 Relevant Publications . . . 4

2 Robot Interaction Studio: An Environment for Minimally Supervised HRI **7**
- 2.1 Robot Interaction Studio . . . 8
  - 2.1.1 Hardware . . . 9
  - 2.1.2 Software . . . 10
- 2.2 Methods . . . 11
  - 2.2.1 Robot Task Cue Design . . . 11
  - 2.2.2 Study Design . . . 13
  - 2.2.3 Quantitative Metrics . . . 15
  - 2.2.4 Statistical Analysis . . . 16
- 2.3 Results . . . 17
  - 2.3.1 User Exploration . . . 17
  - 2.3.2 User Interaction . . . 18
  - 2.3.3 Survey Data . . . 22
  - 2.3.4 Experimenter Observations . . . 24
- 2.4 Discussion . . . 24
- 2.5 Potential Future Work . . . 26

3 Formative and Summative Feedback in Minimally Supervised HRI **27**
- 3.1 Introduction . . . 27
- 3.2 Materials and Methods . . . 28
  - 3.2.1 Objectives . . . 28
  - 3.2.2 Study Design . . . 29
  - 3.2.3 Cue and Feedback Design . . . 31
  - 3.2.4 Participant Information . . . 37
  - 3.2.5 Statistical Analysis . . . 38
- 3.3 Results . . . 38
  - 3.3.1 Performance . . . 39
  - 3.3.2 Cue Comprehension . . . 39
  - 3.3.3 Confidence in Cue Comprehension . . . 39
  - 3.3.4 Activity . . . 41

3.3.5 Participant Perception of Robot Behavior . . . 47
3.3.6 Participant Self-Reported Cue Exploration Measures . . . 47
3.3.7 Participant Survey Responses . . . 50
3.4 Discussion . . . 55
3.5 Conclusion . . . 61

4 Expert Perception of Exercise Robots 63
4.1 Introduction . . . 63
4.2 Methods . . . 64
4.2.1 Phase 1: General Therapy . . . 65
4.2.2 Phase 2: Exercise Robots . . . 65
4.2.3 Phase 3: Teleoperation . . . 66
4.2.4 Final Debrief . . . 66
4.2.5 Statistical Analysis . . . 67
4.2.6 Participants . . . 67
4.3 Results . . . 67
4.3.1 Expert Views on Social Exercise Robots . . . 67
4.3.2 Expert Views on Teleoperation . . . 69
4.3.3 Expert Opinions on Robots . . . 71
4.4 Discussion . . . 72
4.5 Conclusion . . . 75

5 OCRA: An Optimization-based Customizable Retargeting Algorithm for Teleoperation 77
5.1 Introduction . . . 77
5.2 Algorithm . . . 79
5.2.1 Initial Transformations . . . 80
5.2.2 Retargeting . . . 82
5.2.3 Initialization . . . 83
5.3 Methods . . . 84
5.3.1 Benchmarking Tests . . . 86
5.3.2 Generalizability Experiment . . . 86
5.3.3 User Perception Experiment . . . 86
5.3.4 Participant Information . . . 88
5.3.5 Analysis . . . 89
5.4 Results: Benchmarking Tests . . . 91
5.4.1 Effects of Weighting Parameters . . . 91
5.4.2 Latency . . . 95
5.5 Results: Generalizability Experiment . . . 96
5.6 Results: User Study . . . 98
5.6.1 User perception of motion . . . 99
5.6.2 Insight into user ratings . . . 102

5.6.3 Relation between errors and participant ratings 102
5.7 Discussion 104
5.8 Conclusion 107

6 Conclusion **109**
6.1 Summary of Contributions 109
6.1.1 Robot Interaction Studio: An Environment for Minimally Supervised HRI 109
6.1.2 Formative and Summative Feedback in Minimally Supervised HRI 110
6.1.3 Expert Perception of Exercise Robots 110
6.1.4 OCRA: An Optimization-based Customizable Retargeting Algorithm for Teleoperation 111
6.2 Revisiting Overall Research Questions 111
6.3 Implications and Future Outlook 112
6.3.1 Exercise Robots in Real-World Facilities 112
6.3.2 Designing Interactions 113
6.3.3 Implications for the Field of HRI 113

A Supplementary Materials **115**
A.1 Feedback Study 115
A.1.1 Experimenter Script 115
A.1.2 Administered Surveys 116
A.1.3 Additional Results 120
A.2 Expert Perception Study 126
A.2.1 Administered Surveys 126
A.2.2 Semi-Structured Interview Questions 131
A.2.3 Summary Statistics for RAS 133
A.3 Retargeting Study 134
A.3.1 Study Instructions 134
A.3.2 Surveys Administered 135
A.3.3 Additional Results 138

Symbols **139**

Abbreviations **143**

Bibliography **161**

Chapter 1

Introduction

Regular physical activity is essential for maintaining wellness (Childs and de Wit, 2014; Piercy *et al.*, 2018). Exercise is also known to reduce the risk of developing chronic conditions like cardiovascular disease, cancer, and diabetes (Kujala, 2009). Research has demonstrated that even modest levels of physical activity can yield substantial health benefits (Piercy *et al.*, 2018; Gibson-Moore, 2019). Engaging in regular physical activity has an impact on two major indicators of decreased stress: a reduction in heart rate in healthy adults Childs and de Wit (2014) and an increase in heart rate variability in older adults Soares-Miranda *et al.* (2014). However, despite widespread scientific support for the health benefits of exercise, physical inactivity is on the rise worldwide Ozemek *et al.* (2019). A recent systematic review found that the COVID-19 pandemic reduced physical activity levels by an additional 14–72% Christensen *et al.* (2022). There is undeniably an urgent need for strategies that effectively motivate individuals to maintain consistent exercise routines.

As we search for a reliable means of helping people to keep fit, the prospect of standalone robotic systems as exercise coaches is becoming increasingly appealing. Socially Assistive Robots (SARs), which are assistive robots known to provide aid via social interaction (Feil-Seifer and Matarić, 2005), are garnering increased recognition as robotic coaches that can motivate individuals to exercise regularly (Matarić and Scassellati, 2016). For example, a long-term study using the child-sized humanoid robot NAO to provide encouragement and monitor performance found that robot assistance increased user adherence to the rehabilitation program (Céspedes Gómez *et al.*, 2021). In another instance, a Pepper robot was revealed to be preferable to instructions printed on a computer screen for post-stroke rehabilitation (Feingold-Polak *et al.*, 2021). Several cross-sectional studies have also found that people enjoy interacting with robotic exercise companions (Matsusaka *et al.*, 2009; Guneysu and Arnrich, 2017; Fitter *et al.*, 2020). Moreover, these robots have demonstrated considerable promise in healthcare applications (Dawe *et al.*, 2019; Cifuentes *et al.*, 2020). Such robots could help alleviate caregiver staff shortages (Beard and Bloom, 2015; Fleming *et al.*, 2003; Childs *et al.*, 2022) by reducing healthcare workload and ensuring worker safety during situations like the recent COVID-19 pandemic (Scassellati and Vázquez, 2020; Yoganandhan *et al.*, 2021). These scenarios highlight the need for standalone robotic systems that can

effectively coach human users in exercise, physical therapy, and other movement-based activities without close supervision by roboticists or other human experts.

Current research into minimally supervised systems has primarily focused on home-based therapy for stroke rehabilitation (Lambercy *et al.*, 2021; Ranzani *et al.*, 2021; Zhang *et al.*, 2011; McCabe *et al.*, 2019; Hyakutake *et al.*, 2019; Wolf *et al.*, 2015; Cordo *et al.*, 2009). For example, Ranzani et al. modified the ReHaptic Knob, a desktop device for assessing and improving hand function after stroke, to enable lightly supervised use and found that such therapy is possible with minimal initial training (Ranzani *et al.*, 2021). However, in the realm of social robotics, the robot interacts with users as an independent social agent, rather than as a device or tool Feil-Seifer and Matarić (2005). Here, the explicit idea of minimally supervised interaction is gaining popularity as the focus shifts to robots that can leave the laboratory to help users in everyday settings. Specifically, in-home socially assistive robots for children with autism (Clabaugh *et al.*, 2019; Scassellati *et al.*, 2018; Jain *et al.*, 2020), in-home reading companions (Michaelis and Mutlu, 2018), and care robots for older adults (Bajones *et al.*, 2019) both at home and within care facilities (Frennert *et al.*, 2017) have been extensively studied. Minimally supervised in-the-wild robot deployments have also been widely explored within schools (Kanda *et al.*, 2007), in museums (Shiomi *et al.*, 2006; Nourbakhsh *et al.*, 1999), as comedians (Vilk and Fitter, 2020), and as tour guides (Wang and Christensen, 2018). However, these existing socially assistive robots with minimal supervision have predominantly explored verbal (Davison *et al.*, 2020; Clabaugh *et al.*, 2019; Scassellati *et al.*, 2018; Frennert *et al.*, 2017; Akalin *et al.*, 2019; Guneysu and Arnrich, 2017; Swift-Spong *et al.*, 2015) and affective interactions (Scassellati *et al.*, 2018; Jain *et al.*, 2020; Kanda *et al.*, 2007), with gesture-based exercise interactions restricted to closely supervised settings (Fitter *et al.*, 2020; Guneysu and Arnrich, 2017).

Gesture-based communication is as essential as its verbal counterpart for in-person human-human interaction (Norris, 2011). For exercise and other motor-learning tasks, physical demonstrations and gestures become an even more important part of the exchange. Indeed, nonverbal communication forms the silent backbone of interaction that is omnipresent and assists with understanding (Burgoon *et al.*, 2021), sometimes even transcending language barriers (Bartneck *et al.*, 2020). Thus, we believe that nonverbal robot behavior has the potential not only to influence user perception of a robot, but also to elicit behavioral responses and improve task performance (Saunderson and Nejat, 2019). Although there is extensive research on the detection of user gestures (Liu and Wang, 2018), robot-generated gestures are only rarely studied in human-robot interaction (HRI) and are frequently combined with speech (Wolfert *et al.*, 2022). In particular, gesture-based interactions have been heavily studied in the context of storytelling robots (Ham *et al.*, 2015; Sauppé and Mutlu, 2014), social interactions (Zabala *et al.*, 2021; Zinina *et al.*, 2020), and providing engaging educational activities for children (Michaelis and Di Canio, 2022). However, further investigation is needed to de-

termine how humanoid robots can use nonverbal cues to help guide humans to perform physical tasks with minimal expert supervision. This topic is especially timely given the rapid recent improvements in vision-based perception of humans (Pavlakos *et al.*, 2019; Dang *et al.*, 2020) and the increasing deployment of robots in everyday settings (Lambert *et al.*, 2020; Henschel *et al.*, 2021).

1.1 Overall Research Questions

This dissertation plays a substantial role in bridging the knowledge gap in the realm of dynamic gesture-based human-robot interaction under minimal expert supervision. It does this by delving into two pivotal overall research questions:

[ORQ 1] What strategies can be employed by social exercise robots to effectively leverage gesture-based interactions within minimally supervised settings?

[ORQ 2] What methods can be implemented to streamline the development process of gesture-based human-robot interactions?

To address **ORQ 1**, we scrutinize the notion of gesture-based interactions from the perspective of the user interacting with the robot. In doing so, we create a hardware and software system that facilitates minimally supervised human-robot interaction. We subsequently deploy this system to assess a suggested strategy designed to provide dynamic gesture-based feedback in such scenarios. For **ORQ 2**, we investigate gesture-based interactions through the lens of their creation, introducing a paradigm that involves exercise therapy experts who play a crucial role in the design of these interactions.

1.2 Outline

Through Chapter 2, we introduce a minimally supervised setting designed to quantify detailed gesture-based interactions with an exercise robot. We combine Captury Live, a real-time markerless motion-capture system, with a Rethink Robotics Baxter Research Robot to create the Robot Interaction Studio. This chapter provides an in-depth look into the implementation and evaluation of this platform, demonstrated through a within-subjects user study. We showcase the rich and extensive quantitative data that this system can acquire.

In Chapter 3, we propose a novel methodology to deliver gesture-based feedback in minimally supervised scenarios, inspired by proven techniques from the educational sector. We posit that feedback can be divided into two distinct groups, determined by when and in what manner it is delivered during the task. Leveraging the data gathered through the Robot Interaction Studio, we assess the influence of this dynamic nonverbal feedback, provided by the Baxter robot, on human behavior via a comprehensive user study.

Chapter 4 investigates expert viewpoints on the use of teleoperation for the creation of gesture-based interactions. By incorporating expert insights, we identify the necessary features a system should have to enhance technology acceptance and usability. We conduct semi-structured interviews with exercise therapy experts to better understand their perspectives and learn whether they would be interested in our proposed method for enabling them to create these gesture-based interactions.

Chapter 5 focuses on the creation and evaluation of a system that experts can utilize to create gesture-based behavior for the robot. This objective has been achieved via a real-time optimization-based algorithm for mapping motion between two kinematically dissimilar serial linkages, such as a human arm and a robot arm. We then carry out a series of experiments to benchmark this algorithm, demonstrate its applicability to different robots, and finally conduct a user perception study to evaluate the range of motions generated by this algorithm.

Chapter 6 concludes this thesis with a discussion of the contributions and potential future avenues of research that can leverage our findings.

1.3 Relevant Publications

This dissertation contains research documented in the following publications:

1. Mohan, M. and Kuchenbecker, K. J. (2019). A design tool for therapeutic social-physical human-robot interactions. Workshop paper (3 pages) presented at the HRI Pioneers Workshop
2. Fitter, N. T., Mohan, M., Kuchenbecker, K. J., and Johnson, M. J. (2020). Exercising with Baxter: Preliminary support for assistive social-physical human-robot interaction. *Journal of NeuroEngineering and Rehabilitation*, **17**(19)
3. Mohan, M., Nunez, C. M., and Kuchenbecker, K. J. (2020). How does real-time feedback affect communicative actions in social-physical human-robot interaction? Workshop paper (2 pages) presented at the ROMAN Workshop on Quality of Interaction in Socially Assistive Robots (QISAR)
4. Mohan, M., Nunez, C. M., and Kuchenbecker, K. J. (2021b). Robot interaction studio: A platform for unsupervised HRI. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Xian, China
5. Mohan, M., Mat Husin, H., and Kuchenbecker, K. J. (2021a). Evaluation of a teleoperated robotic exercise coach. Workshop paper (4 pages) presented at the HRI Workshop on Workshop YOUR study design! Participatory critique and refinement of participants' studies
6. Mohan, M. and Kuchenbecker, K. J. (2023). OCRA: An optimization-based customizable retargeting algorithm for teleoperation. Workshop paper (3 pages) presented at the ICRA Workshop Toward Robot Avatars
7. Fitter, N. T., Mohan, M., Preston, R. C., Johnson, M. J., and Kuchenbecker, K. J. (2023). How should robots exercise with people? robot-mediated exergames win with

music, social analogues, and gameplay clarity. *Frontiers in Robotics and AI*. Under review

8. Mohan, M., Nunez, C. M., and Kuchenbecker, K. J. (2023). Formative and summative feedback in minimally supervised human-robot interaction. *Under Review*

Chapter 2

Robot Interaction Studio: An Environment for Minimally Supervised HRI

Note: This chapter is based on a published conference paper (Mohan *et al.*, 2021b). Some paragraphs in this dissertation's Abstract, Acknowledgments, Introduction, Background, and Conclusion are also adapted from this publication (©2021 IEEE. Reprinted, with permission, from Mohan, Nunez, and Kuchenbecker, "Robot Interaction Studio: A Platform for Unsupervised HRI", Proceedings of the *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021).

To enable gesture-based interactions with an exercise robot in real-world scenarios, we need minimally supervised systems that can monitor the users effectively. Thus, this chapter introduces the Robot Interaction Studio – a walk-in minimal-setup environment where users can experience minimally supervised interaction with a robot within a relatively large motion-capture workspace. Section 2.1 explains the system's hardware and software setup. As described in Section 2.2, we conduct a preliminary study to explore the potential of the system as a environment for standalone minimally supervised social-physical HRI. Hence, our experiment is designed to enable participant exploration. We hypothesize that when minimal instructions are provided:

H1. Participants will traverse the interaction space.

H2. Participants will enter the workspace of the robot's arms.

H3. Participants will tend to mimic the robot's arm pose.

Section 2.1 describes the design and implementation of the Robot Interaction Studio. The methodology used to evaluate our hypotheses has been detailed in Section 2.2. The results have been presented in Section 2.3, followed by the discussion and conclusion in Sections 2.4 and 2.5, respectively.

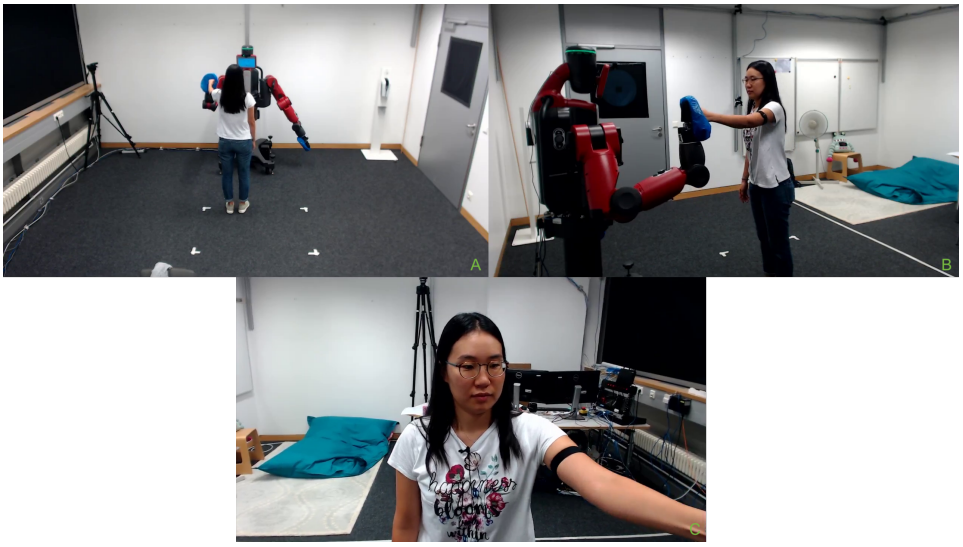


Figure 2.1: Three high-resolution cameras show distinct views of what the minimally supervised user is doing in the Robot Interaction Studio. The cameras for the two upper views are stationary. The lower view comes from the camera attached to Baxter robot's face screen, which continually rotates to face the user based on data from the markerless motion-capture system.

2.1 Robot Interaction Studio

An environment that would enable minimally supervised interaction would need to be easy to setup and able to track user movement with versatility Unfortunately, few suitable human motion-capture solutions exist. Depth sensors such as the Microsoft Kinect and Intel RealSense have a limited workspace and are severely affected by occlusion, especially when physical contact occurs. Other motion-capture systems such as marker-based visual motion capture systems like Vicon Vicon Motion Systems Ltd. or inertial motion-capture suits like Xsens (Movella Inc., 2018) provide high-quality motion tracking but are cumbersome to set up. Motion-capture systems that require worn components do not enable a user to walk into a robot's workspace and begin interaction. Thus, we opted for a markerless solution.

There are several commercial markerless motion-capture systems available such as Theia3D (Theia Markerless, Inc., 2022) and Simi Shape 3D (Simi Reality Motion Systems GmbH, 2023). However, these systems do not provide real-time motion tracking, which is required for the robot to react to user actions. As such, we chose Captury Live (GmbH, 2020), a commercial markerless motion-capture system. Captury's performance has been shown to be comparable to marker-based systems such as Vicon and Qualisys (Manns *et al.*, 2016). Additionally, Captury's Linux compatibility makes it easy to integrate within the ROS environment.

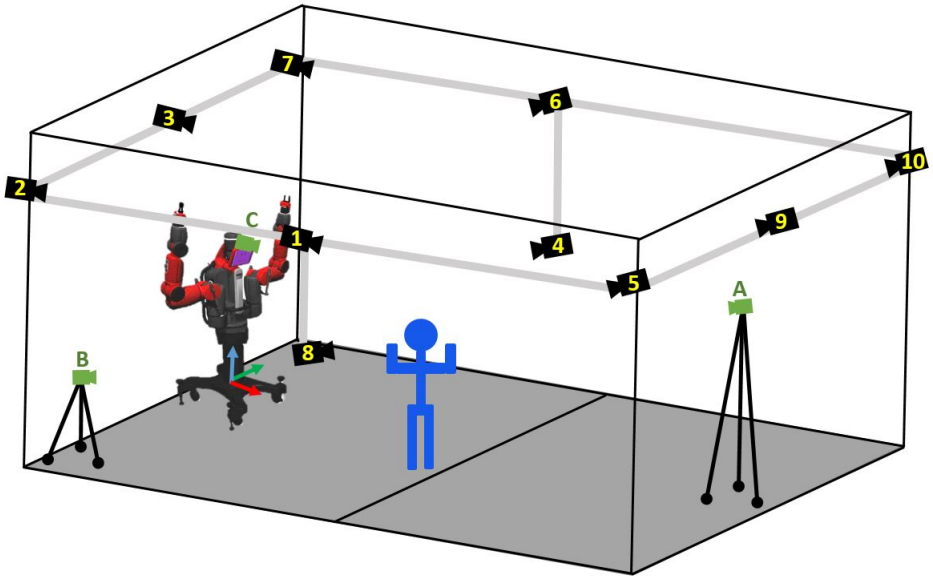


Figure 2.2: The approximate locations of the thirteen cameras used in the Robot Interaction Studio. The three-green camera icons (A, B, C) show the placement of the high-resolution cameras that allow viewing of the interaction after it occurs, as shown in Fig. 2.1. The ten black camera icons (1–10) represent the cameras used for markerless motion capture, which have a lower resolution but a faster frame rate to allow for real-time performance. The coordinate frame shows the origin of the environment.

2.1.1 Hardware

The setup comprises a Rethink Robotics Baxter Research Robot and the Captury Live markerless motion-capture system (GmbH, 2020). Baxter is an upper-body humanoid robot with two 7-degree-of-freedom (7-DoF) arms and a 2-DoF screen that acts as a face. Captury uses ten 0.5 MP BlackFly GigE (Model No.: BFLY-PGE-05S2C-CS) cameras positioned around the room (Fig. 2.2) to perform high-fidelity real-time motion capture of the user's skeleton at 50 frames per second without any markers. The workstation computer is an Alienware Area 51 CPU with an Intel i9-7980XE processor, 64 GB RAM, and two NVIDIA GeForce GTX 1080 Ti GPUs.

There are three 1080p Logitech c925e webcams in the environment, one of which is attached to the top edge of Baxter's face screen, whereas the other two are on tripods. The Captury cameras are attached to the walls of the room using Bosch Rexroth Aluminum strut profiles and can be seen in Fig. 2.2. These cameras are PoE (Power over Ethernet) compatible and are hence connected via a D-Link DGS-1510-28P network switch. The switch is connected to the CPU via an SFP+ (enhanced small-form-factor pluggable transceiver) port on an Intel X710 10GbE Network Adapter. The system also includes a Sennheiser XSW 2-ME2 wireless lapel microphone that can be worn by the user if voice capture is desired, plus a Polar OH1 optical heart-rate monitor (Polar Electro, 2020). The



Figure 2.3: A screenshot of the Captury experimenter dashboard while a user takes part in the reported proof-of-concept study. Each small numbered image shows the view from one of the ten cameras used for real-time markerless motion capture. The lower-right image shows an expanded view from camera 8; note that other the locations of other cameras are annotated with a small reference frame in each camera's view. Captury also draws the human skeleton that it calculates on top of each video frame, demonstrating a good match with the user's actual pose in three dimensions.

entire setup costs approximately 71,000€ (Captury: 23,000€, Baxter robot: 42,000€, workstation computer: 5,000€, webcams and network equipment: 1,000€).

2.1.2 Software

The Captury Live software runs on Baxter's workstation computer. After camera installation and focusing, the intrinsic and extrinsic camera parameters are found once via calibration with a checkerboard. The calibration process for a new user is then very simple. The user walks into the Captury workspace and performs simple motions that involve moving their arms around and flexing their leg joints. This process enables Captury to scale its human body model to the user. Tracking can begin within 20 seconds of a user entering the workspace. We were unable to use Captury's latest tracking algorithm due to stability issues in our software environment, which led to periodic crashes. Hence, we used the legacy algorithm.

The operating system on the workstation computer is Ubuntu 18.04 (Bionic Beaver)

with ROS (Robot Operating System) Melodic Morenia (Quigley *et al.*, 2009; Open Robotics, 2020b). Baxter is programmed via ROS and accesses Captury's real-time tracking information via a ROS-compatible plug-in. Additionally, Captury has a C++ library that can be used to remotely control the software via custom-built code. Examples of features provided by this library include searching for the presence of a user in the Captury workspace, starting and stopping user tracking, and scaling and re-scaling the user being tracked.

The raw data recorded via Captury includes not only data pertaining to the tracked users but also videos from all ten cameras. Recordings from the high-resolution webcams and microphone are collected using OBS Studio (OBS Project, 2020). The data from the Captury ROS plug-in is transmitted to Baxter's ROS environment in the form of tf (transform) messages. A homogeneous transformation is then applied to this data to bring it into Baxter's frame of reference. The origin of our setup coincides with the origin of the Baxter robot, which is directly below the robot. The coordinate frame can be seen in Fig. 2.2, where red, green, and blue represent the X, Y, and Z axes, respectively.

The ROS implementation of the code for the experiment relies heavily on ROS *actionlib* (Open Robotics, 2020a) which enables the creation of preemptable tasks, thus enabling real-time responses from the robot. Once a user enters the Captury workspace, the robot immediately starts tracking this person using its head. As the user walks around in the robot's workspace, the head pan angle continually changes to point toward the user's head. The ROS compatibility of the system could enable robots other than Baxter to be used within the Robot Interaction Studio.

2.2 Methods

To assess our platform and evaluate how minimally supervised users interact with a humanoid robot, we developed specific cues that a Baxter robot can perform and conducted a corresponding user study in our Robot Interaction Studio. The robot's parallel-jaw grippers were equipped with foam boxing pads and sterile covers to make contact safe and comfortable.

2.2.1 Robot Task Cue Design

We created three types of task cues with five variants that a Baxter robot can relay with the goal of trying to get the user to move around in the room and interact with the robot. The three task cues take inspiration from existing gesture-based HRI and the psychology literature. The robot's nonverbal feedback also seeks to engage and challenge participants (McMillan, 2013). Prior work demonstrated that a robot needs to provide only minimal social cues to enhance the persuasiveness of the interaction (Vossen *et al.*, 2010). Thus, we chose simple cues including sounds, emotions displayed on the robot's face screen, and arm-based gestures to attract attention. Fig. 3.1 shows the three cue types,

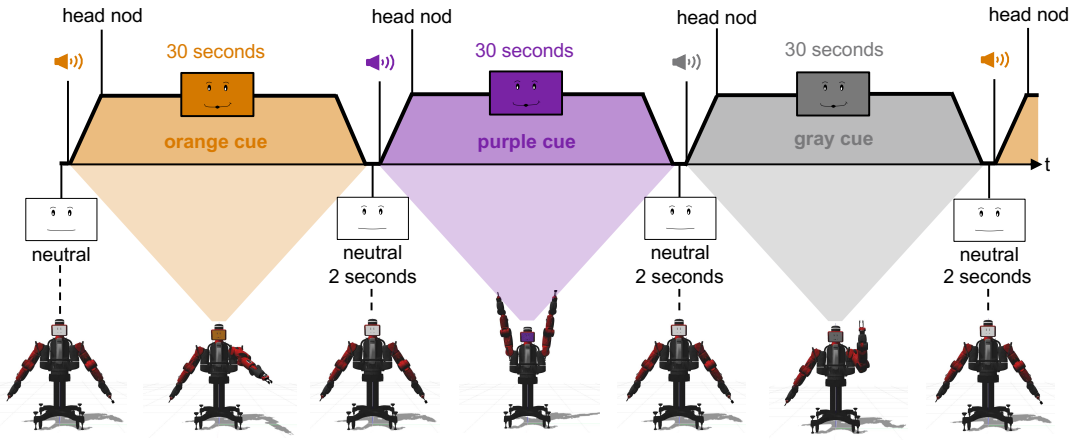


Figure 2.4: Timeline showing the sequential presentation of orange, purple, and gray cues during the experiment. The robot begins with a white neutral face, holding both arms down at its sides. Each cue starts with a characteristic sound; the robot's face takes on the cue's color, and then its arms move to show the cue. When the robot reaches its final position, it nods its head to indicate that the user should respond. After 30 seconds, the arms and face return to neutral, and the next cue begins after a short pause. Cues always cycle through in the depicted color order, with each cue chosen randomly from the five available options within the color category.

their five variants, their associated types of formative feedback, and how summative feedback was conveyed across tasks. The cues are multimodal, blending the different types of nonverbal communication available to robots, which can be classified as gaze, gesture, mimicry, touch, movement, and interaction timing based on how they are implemented (Bartneck *et al.*, 2020). Utilizing multimodal feedback also aligns with our goal of leveraging concepts from the field of education where this idea is supported by the dual-coding theory (Clark and Paivio, 1991) i.e., information is more likely to be remembered when it is presented with both verbal and nonverbal symbolism. Thus, Baxter communicates each cue to the user by providing a multi-modal combination of prompts that include facial expressions, facial colors, sounds, head motions, and arm motions (Fig. 2.4).

Eye gaze is an integral element of human social interaction across all cultures (Argyle *et al.*, 1994) and has thus been utilized by roboticists in many ways, such as indicating a robot's mental state, directing user attention, and increasing fluidity in conversations (Admoni and Scassellati, 2017). Users have also been shown to prefer robots that move their heads (Zinina *et al.*, 2020). Thus, we combined these ideas into a referential robot gaze directed at the user, as in (Admoni and Scassellati, 2017): across all conditions of the study, the robot constantly looked at the user by turning its head to face them as they walked around the room. One high-resolution camera was mounted on top of the robot's head to capture a view of the participant. The robot's eyes periodi-

cally blinked to indicate awareness (Fitter *et al.*, 2020), and the robot displayed verified open-source facial expressions (Fitter and Kuchenbecker, 2016a) and played open-source sound (Freesound, 2020) at key points in the interaction (Fig. 2.4). We used previously validated Baxter faces (Fitter and Kuchenbecker, 2016b) and grouped the cues into three categories based on the facial color of the robot: orange, purple, and gray.

All three cue types harness gesture-based interaction to convey meaning integral to the task. First, the orange is a pointing (deictic) gesture that indicates a particular thing or area (Bartneck *et al.*, 2020; McNeill, 1992) and is one of the first gestures learned by infants (Tomasello *et al.*, 2007). These types of gestures have been well-explored within HRI (Sauppé and Mutlu, 2014). Thus, for these types of task cues, the robot points to a location on the ground.

Second, the purple cue invokes mimicry or imitation-based nonverbal communication (Bartneck *et al.*, 2020). Imitation is a fundamental technique children use to learn physical skills (Wang *et al.*, 2015b) and develop social interaction abilities (Over, 2020), and pose imitation has been extensively used and validated in social robotics (Guneysu and Arnrich, 2017; Shao *et al.*, 2019; Lewis *et al.*, 2016; Görer *et al.*, 2013). The tested variants of the purple cue are based on the Roboga Game by Fitter et al. (Fitter *et al.*, 2020). Our robot held up both of its arms in a symmetric pose.

The third cue type was the gray cues; it highlights touch gestures, which have been shown to evoke positive feelings with users (Gallace and Spence, 2010; Bartneck *et al.*, 2020). Similar to a high-five (Fitter and Kuchenbecker, 2020, 2019), the robot held up one or both of its hands, palm(s) out.

While we expect a large variation in responses between users, we designed each of these categories of cues to attempt to communicate specific actions. During the orange cues, Baxter uses one straight arm to point to a specific location on the floor of the Robot Interaction Studio workspace in the hopes that the user moves to that location. For purple cues, Baxter symmetrically moves its arms into a distinct elevated pose to try to communicate that the user should mimic this arm posture. Finally, in the gray cues, Baxter positions one or both of its hand pads to try to encourage the user to touch its presented palm(s). We are interested in exploring how users interpret these cues when they receive *no explicit instructions or feedback about their actions*.

2.2.2 Study Design

To investigate minimally supervised users in the Robot Interaction Studio, we conducted a user study with 7 participants (4 male, 3 female; aged 23–33). Participants were recruited from the Max Planck Institutes in Stuttgart, the University of Stuttgart, and a local Facebook group for foreigners. Five participants reported their experience interacting with a Baxter robot as “Complete Novice” (that they did not know what a Baxter robot was before this study), one reported that they were a “Beginner” (they had seen and/or read about a Baxter robot before), and one reported that they were an “Expert” (they frequently program and/or work with a Baxter robot); the expert was not a member of

our team or institution. The experimental procedure was approved by the Ethics Council of the Max Planck Society under the Haptic Intelligence Department's framework agreement as protocol F007A. Participants not employed by the Max Planck Society were compensated at a rate of 8€ per hour.

After providing informed consent, participants completed a pre-study robot evaluation survey in which they rated their level of agreement with 21 prompts using continuous sliders. (Weiss *et al.*, 2008) used this survey to assess changes in user perception of robots including participant attitude toward technology, self-efficacy and attachment; it is adapted from the Unified Theory of Technology Acceptance (Venkatesh *et al.*, 2003), and object sociality (Cetina, 1997). All surveys were administered via LimeSurvey. The experimenter then explained that the robot will present the user with a command via arm and head movements, facial expressions on the robot's screen, and sounds. The participants were also informed that they could move around the room or make contact with the robot to complete the task indicated by the robot; they were asked not to touch the robot's joints for safety reasons. The script that the experimenter used to maintain consistency across participants can be seen in Appendix Section A.1.1.

Since this study was the first deployment of the Robot Interaction Studio, the experimenter stayed in the room to ensure safe operation. She sat at a desk at the back of the room to monitor the technical state of the system via several large computer monitors that blocked her direct view of the user. If a participant asked additional questions or requested help in completing the task, the experimenter merely stated that the participant was allowed to move around the room or make contact with the robot in order to complete the task. No additional guidance or answers were provided.

The interactive part of the study followed the timeline shown in Fig. 2.4. To begin, Baxter had a white neutral face, holding both arms down at its sides. Then, the sound associated with the orange cue played, Baxter's face changed color to orange, and Baxter moved its arms into one of the five orange cue variations. When Baxter's moving arm reached its final position, Baxter nodded its head to indicate to the user that the trial had begun. The user had 30 seconds to respond to Baxter's cue before Baxter's face returned to the white neutral image and its arms went back down to the neutral position. After a two-second-long break, the sound associated with the purple cue played, Baxter's face changed color to purple, and Baxter moved its arms into one of the five purple cue variations. Again, once Baxter reached its final pose, it nodded to indicate that the trial had begun. Following the same described cycle as before, Baxter returned to neutral at the end of the trial before beginning the sequence to relay a gray cue.

Participants completed 25 triplets of cues with the trials always following the order of orange cue, purple cue, gray cue, for a total of 75 trials. To determine the presentation order, we first determined all possible combinations of one orange cue, one purple cue, and one gray cue and then permuted the order of the combinations. To provide enough unique combinations for each group of participants, we duplicated the initial list of 125 combinations, resulting in a total of 250 combinations. The trial order for the first participant of each feedback group corresponded to combinations 1–25, the trial order for the

second participant corresponded to combinations 26–50, and so on, such that the participants in each feedback group all experienced the same number and order of randomized trials. A brief break was provided to participants approximately after trials 25 and 50. During each trial, the Robot Interaction Studio collected raw pose and position data and raw video from Captury, video from the three webcams, audio from the microphone, and heart-rate data, as described in Section 2.1. Participants were given a short break approximately every 12 minutes.

After participants had completed all of the trials, they completed a post-study survey (Appendix Section A.1.2) consisting of the same robot evaluation survey from the start of the study, a System Usability Scale (SUS) (Brooke, 1996), and a NASA Task Load Index (TLX) (Hart and Staveland, 1988) for each of the cue colors. This survey also included demographics questions and open-ended questions designed to elucidate the participant's actions, thoughts, and understanding during the study. Users completed all study activities in approximately 1 hour and 15 minutes.

2.2.3 Quantitative Metrics

Since the study was run within the Robot Interaction Studio, we had access to full-body markerless motion-capture data of the user, facial expression recognition from the camera on the robot's head, and an additional Polar OH1 (Polar Electro, 2020) optical heart-rate sensor worn around the user's left bicep. For the purposes of this study, we narrowed these down to the following measures.

Distance Traveled: This metric (D) can be defined as the total length of the path traversed by a participant over the 30 s duration of a trial. Let k be the distance traveled by the user across the floor in time step t. Then, this metric can be calculated as follows:

$$D = \sum_{0\,\mathrm{s}}^{30\,\mathrm{s}} k(t)dt \tag{2.1}$$

Time Spent in the Robot's Workspace: This metric is the percentage of trial time that a user's hips were located within Baxter's reachable workspace, which we defined as a circular region of the floor centered on the robot's torso.

User Activity Heatmaps: We created three types of heatmaps for each of the 15 cues by discretizing the relevant plane into a 20×20 grid and then binning the position data. We created three types of cue-specific heatmaps. The orange heatmaps represented the total time that users spent in a given position along the XY plane in the workspace in response to each of the fifteen cues. The purple and blue heatmaps represent the user's left- and right-hand positions, respectively. All user arms were scaled up to the length of the robot's arms to facilitate comparisons. These heatmaps represent the YZ plane or

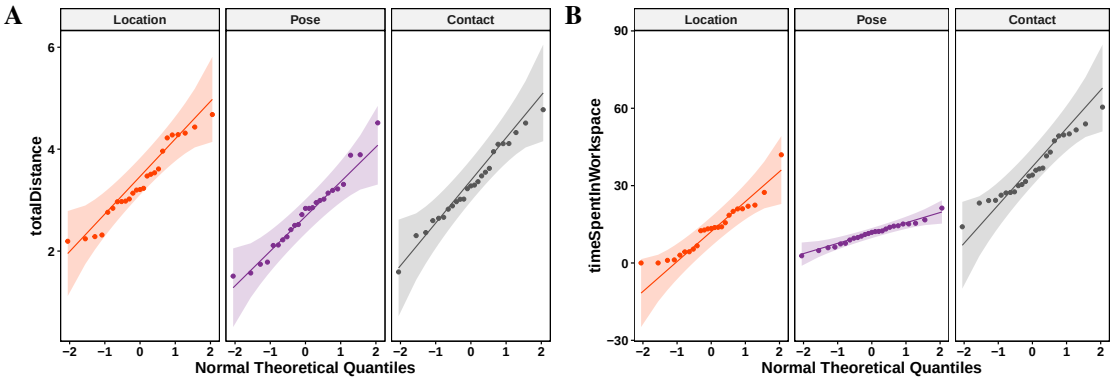


Figure 2.5: **(A)** Q-Q plot for the total distance travelled dependent variable that shows that all points fall along the reference line. **(B)** Q-Q plot for the percent time spent in workspace variable that shows that all points fall along the reference line for this dependent variable as well.

the front view. Finally, the gray heatmaps depict the overhead view of the user's hand position in the XY plane.

Recovered Data

The heatmap visualizations revealed that the hand positions recorded from approximately 20% of all trials (from 0 to 23 trials per participant) were corrupted due to a bug in the data-logging software. However, we were able to successfully recover the lost data by reprocessing the raw motion-capture data, which had been saved separately.

2.2.4 Statistical Analysis

This study has been designed as a within-subjects study with cue type as the only factor with three levels. Since we were interested in participant activity over time across the three cue types, we used the number of trials as our unit of analysis ($n = 25$). We verified if the two dependent variables (distance traveled and time spent in workspace) met the assumptions for a parametric one-way repeated measures analysis of variance (ANOVA). There were no extreme outliers. Additionally, assumptions of normality were not violated, as demonstrated by a non-significant Shapiro-Wilk test ($p > 0.05$) and the Q-Q plot, which had all the points fall approximately along the reference line (Fig. 2.5). Thus, we analyzed these two dependent variables via parametric one-way repeated measures ANOVAs. All post-hoc testing utilized a Bonferroni correction. The data was analyzed in R version 4.2.3. The function utilized for the ANOVA automatically checks if sphericity has been violated via Mauchly's test and applies the Greenhouse-Geisser sphericity correction when required.

2.3 Results

The developed platform enabled us to collect a plethora of data. For this chapter, our goals are to showcase the capabilities of our platform in enabling minimally supervised human-robot interaction and to investigate how users explore within a space when left alone with an active robot. Therefore, our presented results will focus on these elements. Sample trials from the study can be viewed via the supplementary video associated with Mohan *et al.* (2021b).

2.3.1 User Exploration

To test if participants explored the interaction space (**H1**), we examined the horizontal distance that they traveled during each trial, averaged within cue types (see Fig. 2.6A). Participants varied in their movements and responses to each cue type, with some participants traversing several meters during a particular trial and others typically standing still. The repeated measures ANOVA revealed that there was a significant difference between the cue types ($F(2,48) = 10.265, p = 0.000194, \eta_g^2 = 0.13$) in the distance traveled. A post-hoc pairwise paired t-test with Bonferroni correction showed that the distances users traveled during orange cues were significantly higher than during the purple cues ($t(24) = 4.64, p = 0.000309, d = 0.928$). Additionally, users walked significantly more for the gray cues compared to the purple cues ($t(24) = 3.58, p = 0.005, d = 0.716$). This result matches our initial design intent for the cues; we imagined that the user might move to a new point in the workspace when confronted with the single downward-pointing arm of an orange cue, but not when the robot stretched both of its arms into the air during a purple cue, which we hoped would inspire them to stand still and mimic the robot's pose. Finally, the palm-offering gray cues would ideally require the participant to walk to the robot from their current position in the Robot Interaction Studio to touch the palm.

To investigate if participants entered the workspace of the robot's arms (**H2**), we looked at the percentage of time during each trial that the user's hip center point was located within Baxter's workspace (see Fig. 2.6B). The one-way repeated measures ANOVA (with Greenhouse-Geisser sphericity correction) revealed that there were significant differences between the various cue types for this metric ($F(1.45, 34.83) = 56.787, p = 2.61 \times 10^{-21}, \eta_g^2 = 0.616$). Post-hoc pairwise t-tests revealed that participants spent significantly more time in the robot's workspace during the gray cues in comparison to the orange cues ($t(24) = 6.98, p = 9.78 \times 10^{-7}, d = 1.4$) and purple cues ($t(24) = 11.6, p = 7.98 \times 10^{-11}, d = 2.31$). There was no significant difference in the time spent in the robot's workspace between orange and purple cues. Again, this result corresponds with our original design intent for the cues, as we anticipated that participants would move toward Baxter when receiving the gray cues in order to interact with its hands, but that they would likely be moving around the workspace during the orange cues and would stand still to mimic the robot during the purple cues.

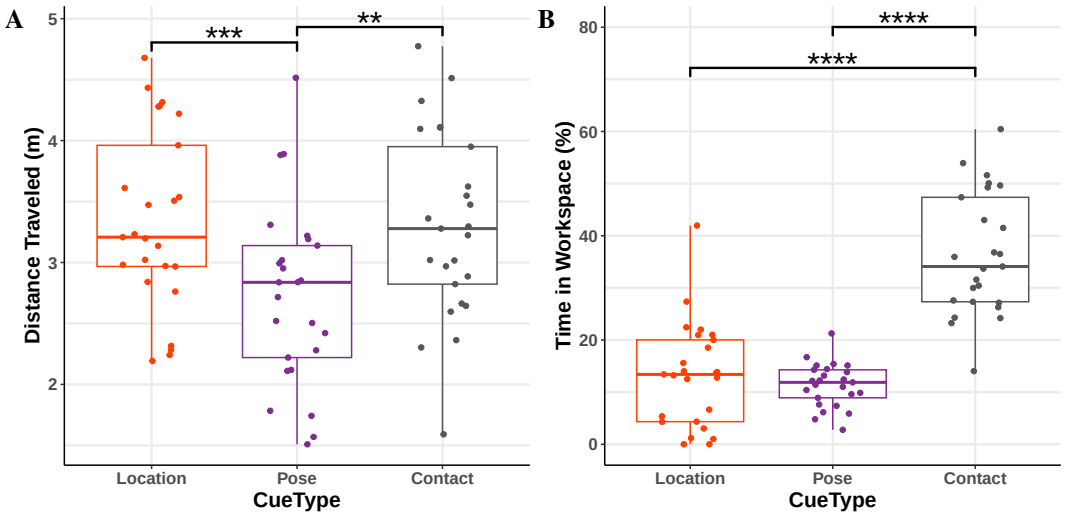


Figure 2.6: **(A)** Average distance participants traveled per trial of the different cue types across the floor of the Robot Interaction Studio during each cue type. **(B)** Average percentage of the trial time that each user spent inside the robot's reachable workspace during trials of each color. A value of 100% corresponds to approximately 30 seconds. Significance notations: $\star\star$ $p < 0.01$, $\star\star\star$ $p < 0.001$, and $\star\star\star\star$ $p < 0.0001$.

2.3.2 User Interaction

We created heatmaps for the three cue types to obtain a birds-eye view of user activities during the 15 different task cues; they can be seen in Figures 2.7, 2.8 and 2.9).

We also used these heatmaps to examine our third hypothesis (**H3**) and understand if participants tended to mimic the robot's arm poses. The blue and purple heatmaps illustrate the two most common locations for the user's hands across all of the cue types: either mirroring the hand positions displayed by the robot or resting at the user's sides. The following subsections use the heatmaps as a lens to examine user activities during the task cues in more detail.

Orange cue: The heatmaps for the orange cues can be seen in Fig 2.7. The orange heatmaps show that the participants often moved away from the center of the workspace in the direction that Baxter was pointing. They also tended to imitate the robot or hold the robot's hand as shown by the hand position heatmaps.

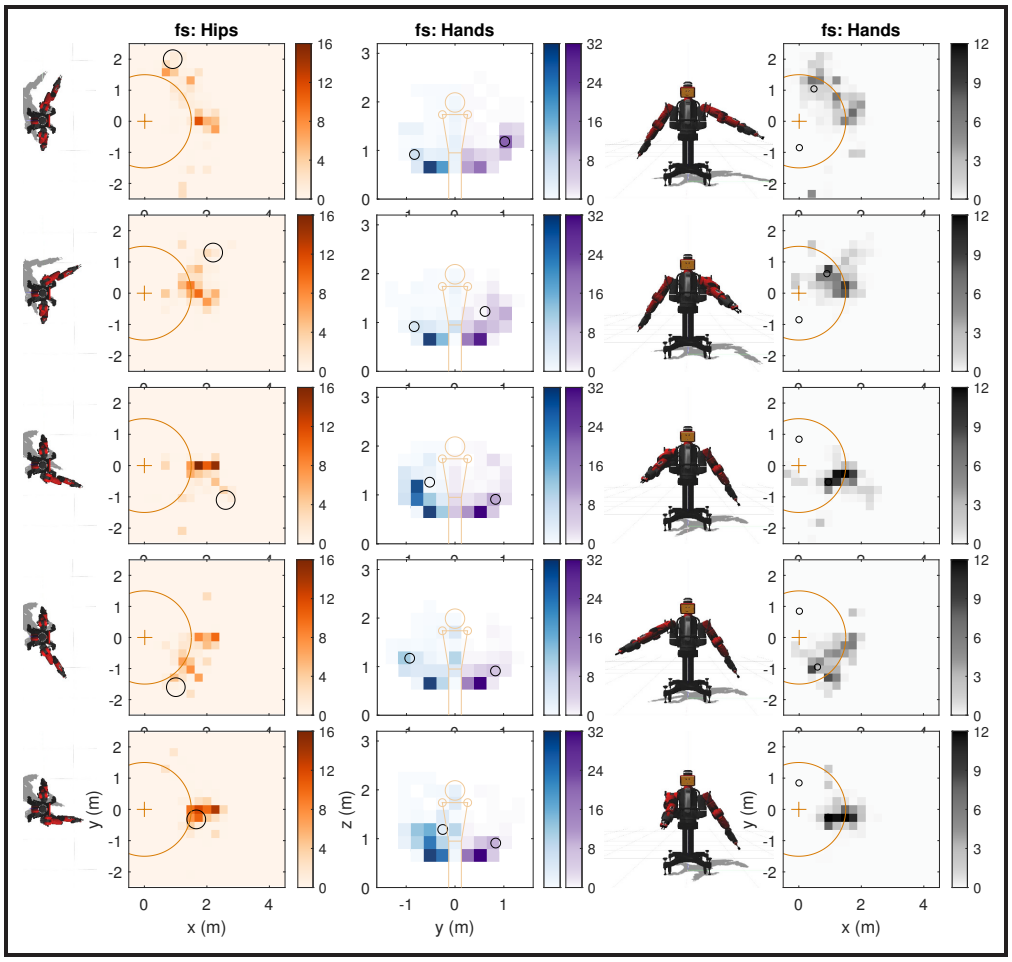


Figure 2.7: Heatmaps showing all five variants of the orange cue. The first column shows the top view of the robot's task cue instruction. The second column shows heatmap visualizations of user position in the workspace of the Robot Interaction Studio. The black circle represents the position that the robot was pointing to. The third column depicts heatmap visualizations of the front view of the user's hands. The black circles represent hand positions corresponding to the robot end-effector positions. These visualizations show that participants often either mimicked the robot or held their hands at their sides. The next column shows front views of the robot's instructions for the variants of the orange cue. The final column shows the top view of the user's hand position. The black circles are top-view positions of the robot's end effectors. This interesting set of plots shows the tendency of users to hold the robot's end-effectors.

Purple cues: Fig 2.8 presents the heatmaps for the purple cues. The orange heatmaps show that the participants spent the majority of time in the center of the rectangular space, opposite the robot. The purple and blue heatmaps show that participants mimicked the robot frequently. Finally, the gray heatmaps show that they rarely came into physical contact with the robot for cues of this type; an exception was that one participant hugged the robot for several cues of this type.

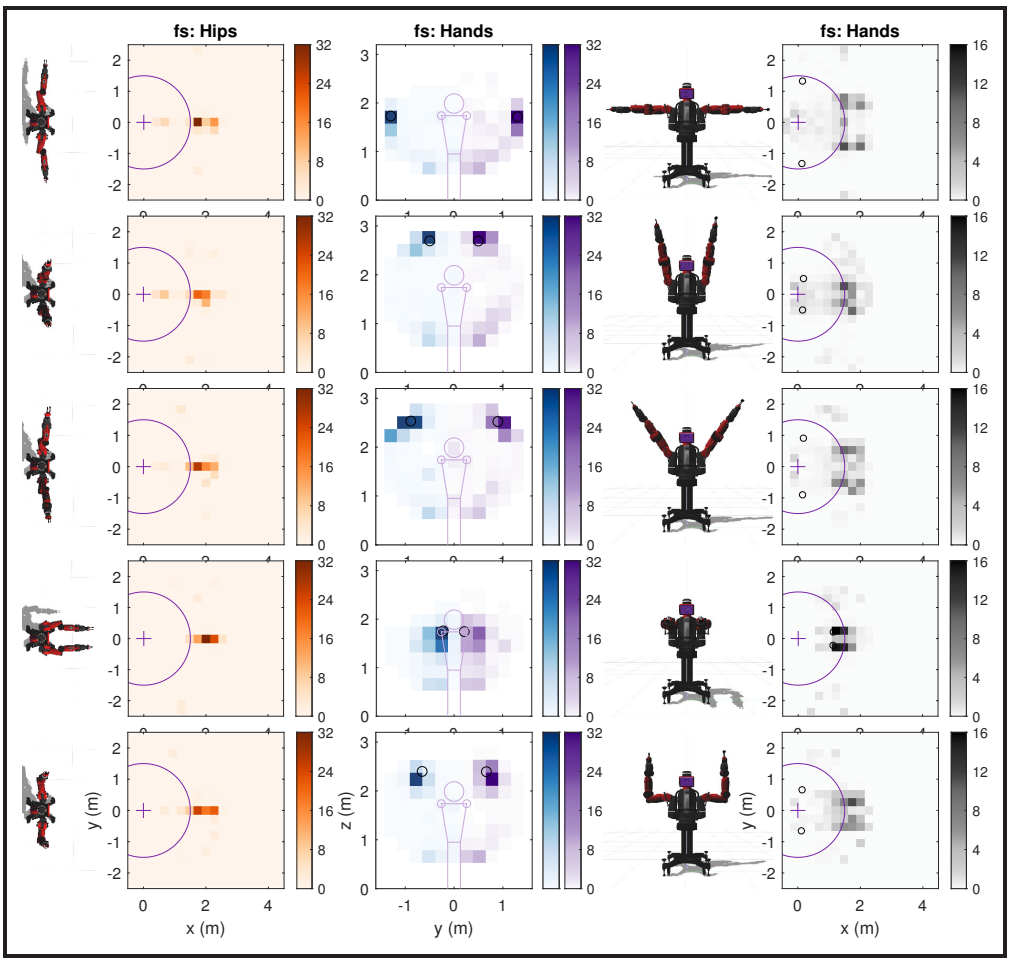


Figure 2.8: Heatmaps showing all five variants of the purple cue. The first column shows the top view of the robot's task cue instruction. The second column shows heatmap visualizations of user position in the workspace of the Robot Interaction Studio; participants did not move much during these cues. The third column depicts heatmap visualizations of the front view of the user's hands. The black circles represent hand positions corresponding to the robot end-effector positions. These visualizations show that almost all participants mimicked the robot over several trials. The next column shows front views of the robot's instructions for the variants of the purple cue. The final column shows the top view of the user's hand positions.

Gray cues: Fig 2.9 shows the heatmaps for the gray cues. One can see from the orange and gray heatmaps that participants tended to move toward the robot's offered hand(s), often stepping inside its reachable workspace. They also mimicked the robot's pose as shown by the purple and blue heatmaps.

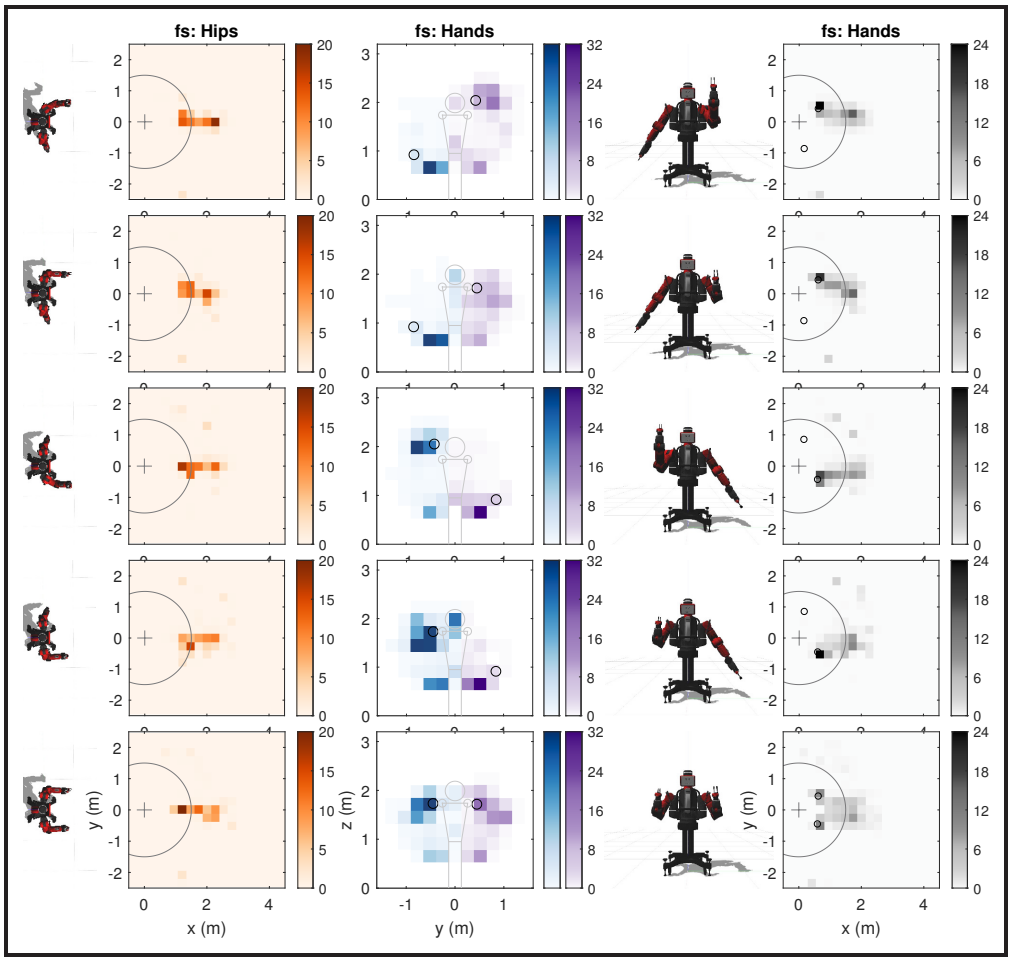


Figure 2.9: Heatmaps showing all five variants of the gray cues. The first column shows the top view of the robot's task cue instruction. The second column shows heatmap visualizations of user position in the workspace of the Robot Interaction Studio. Participants generally walked into the robot's workspace. The third column depicts heatmap visualizations of the front view of the user's hands. The black circles represent hand positions corresponding to the robot end-effector positions. These visualizations show that participant pose is similar to that of the robot. The next column shows front views of the robot's instructions for the variants of the gray cues. The final column shows the top view of the user's hand position(s). These show that participants often made contact with the robot's offered hand, as seen by the darker squares near the annotated end-effector positions.

2.3.3 Survey Data

Since the items of the Robot Acceptance Survey fulfilled the requirements of a parametric test, we conducted paired t-tests to compare the pre- and post-study responses to each of the questions within the Robot Acceptance Survey; there were no significant differences in the user responses to any of the questions (Fig. 2.10 and Fig. 2.11). Many human-robot interaction studies aim to achieve improved robot acceptance responses in the post-study survey compared to the pre-study survey. However, due to the open-endedness of our study and our refusal to provide participants with any additional instructions, we are pleased that there is no significant difference in the reported survey responses before and after the study.

We calculated an average SUS score of 55.7 ± 14.0 out of a maximum of 100, which is below the average score of 68 (Fig. 2.12). Although participants thought the system was well integrated (high score), they also believed they would need assistance to be able to use it (low score), perhaps due to the minimal level of instructions they received in this study.

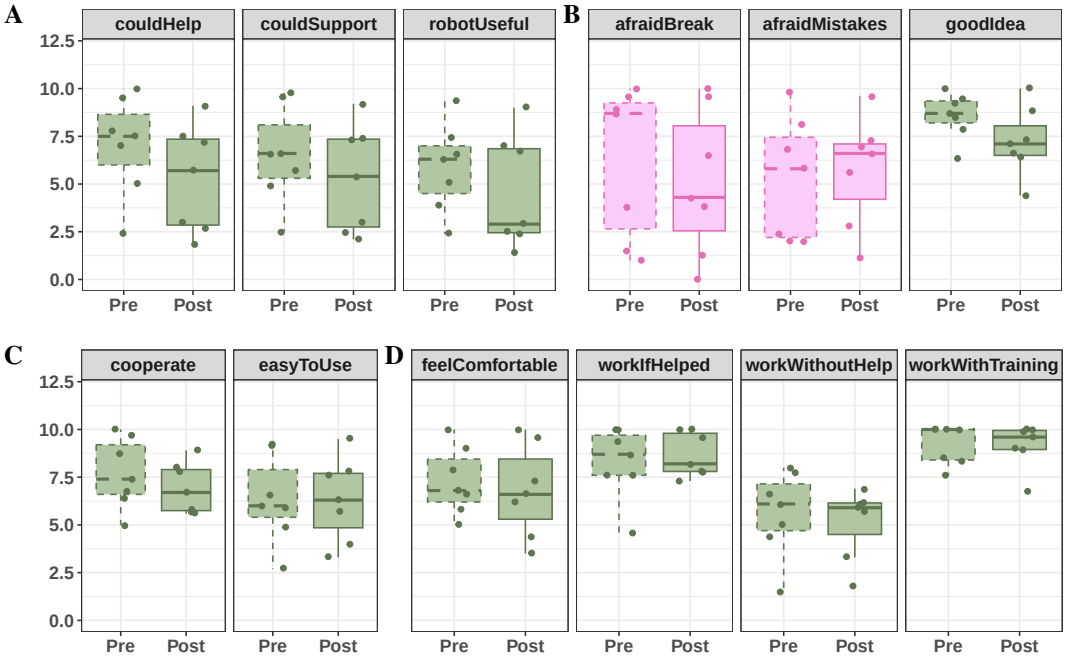


Figure 2.10: Robot Acceptance Survey categories based on the UTAUT model before (pre) and after (post) the study. Positive questions are colored green, and negative questions are colored pink. The paired t-tests revealed no significant results. **(A)** Participant responses regarding performance expectancy of robot. **(B)** Participant responses to questions assessing their attitudes toward technology. **(C)** Participant responses about effort expectancy. **(D)** Participant responses to self-efficacy measures.

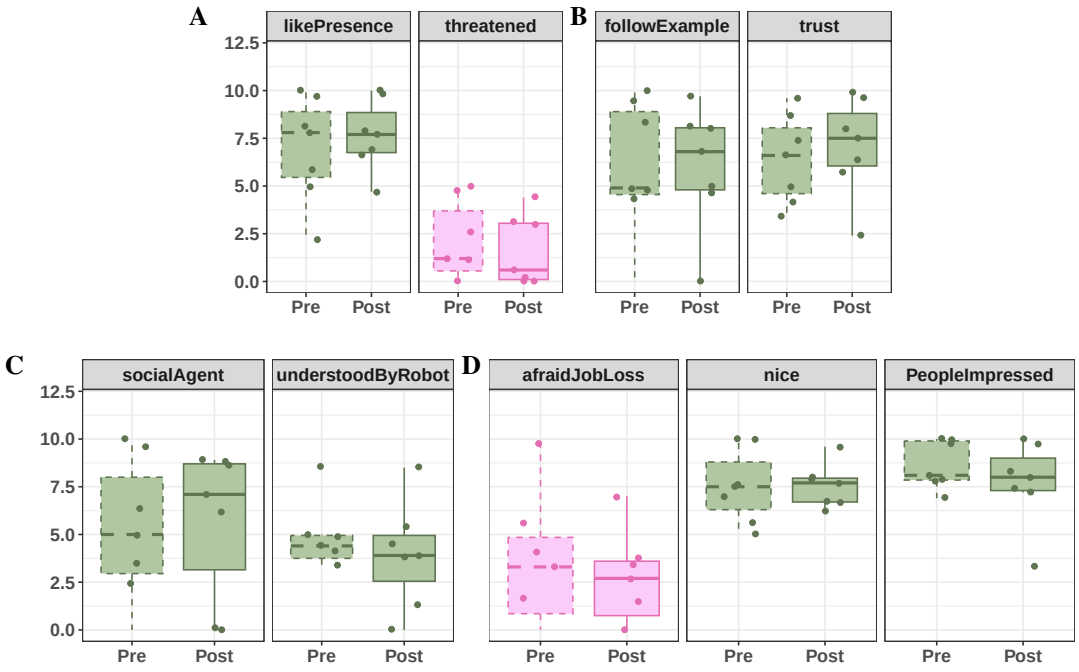


Figure 2.11: Robot Acceptance Survey categories based on the concept of object-centered sociality before (pre) and after (post) the study. Positive questions are colored green, and negative questions are colored pink. Paired t-tests revealed no significant differences. **(A)** Participant responses to forms of grouping category items. **(B)** Participant responses regarding attachment toward robots. **(C)** Feelings of reciprocity reported by participants. **(D)** Participant responses to cultural perception of robots.

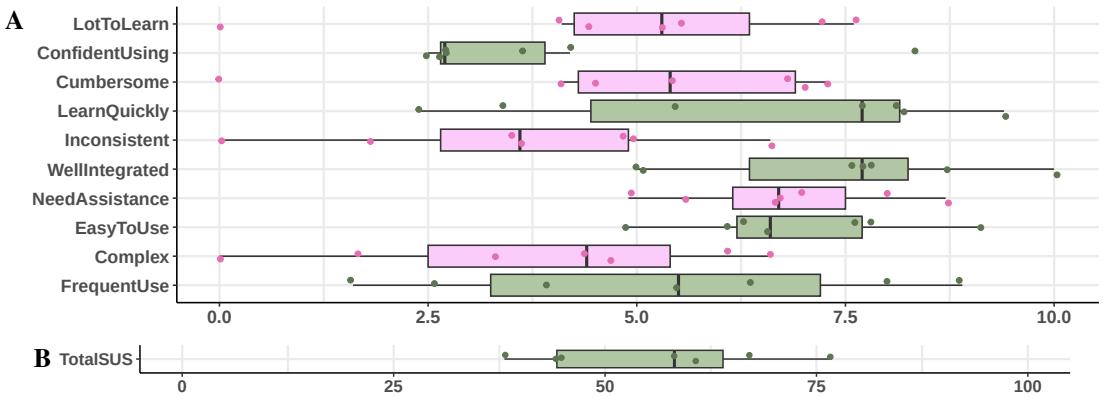


Figure 2.12: Items on the System Usability Scale. **(A)** The ratings on the individual items are quite average. Although participants felt that they would require assistance to use the system, they felt that it was well-integrated. **(B)** The overall SUS score is relatively low as a result of the lack of instructions.

2.3.4 Experimenter Observations

The most common participant reaction was to mimic the robot, especially for the purple cues. Some participants touched, punched, or high-fived the robot's hand for the gray cues. The orange cue elicited diverse reactions from the participants. These included pushing, pulling, or even kicking the outstretched robot arm. One participant asked the robot if it wanted a kiss on the hand and proceeded to kneel on the floor and pretend to kiss the robot's hand. This participant also hugged the robot for several of the purple cues. This participant later mentioned that they thought the orange cues were a marriage proposal scenario. Some participants walked around the workspace, especially during the orange cue as they went to interact with the robot hand. Some participants also became bored after several trials; they then proceeded to perform similar actions for the different cues without performing further exploration. Few participants spoke to the robot.

2.4 Discussion

As discussed in Chapter 1, there is a need for easy-to-use robotic systems that enable minimally supervised human-robot interaction while having the capability to monitor the user, especially for use with socially assistive robots. Our platform addresses many of the issues that have previously impeded achievement of this goal. Most importantly, the user does not need to don any markers or other equipment; they can simply walk into the workspace and begin interacting with the robot and similarly leave at any point without needing to do anything special. While we did have an experimenter in the room to conduct this study (to meet the conditions stipulated for the ethical approval), the platform does not require the presence of an experimenter to help the user start the interaction nor to control the robot. The robot can function fully autonomously, and the robot is inherently quite safe; users could be close to it, touch it, and push its arms around during the experiment without being in danger.

Hypothesis Testing

The heatmaps seen in Figures 2.7, 2.8 and 2.9 showcase the high-quality quantitative record of the user-robot interaction that can be captured by the Robot Interaction Studio. Column one shows patterns of user traversal across the workspace of the Robot Interaction Studio in detail. Combined with Fig. 2.6A, one can clearly see that hypothesis **H1** held true; participants indeed walked around the interaction workspace. This finding implies that the Robot Interaction Studio could also be used in conjunction with robotic gait-training exercise coaches for lower-limb exercise detection (O'Reilly *et al.*, 2018) and to calculate gait-training metrics (Horak *et al.*, 2015). For example, Gosine et al. (Gosine *et al.*, 2015) use motion capture for a stepping game.

Examining Fig. 2.6B along with the heatmaps also shows us that participants entered

the robot's workspace (**H2**). The motion-capture data can be used by a ROS-compatible robot to perceive the user in real time. It can also be programmed to respond to the user's motions (such as with our programmed head tracking). We believe Captury enables the robot to see and interpret a human's movements even better than a typical human observer or the user's own proprioception. Interestingly, one participant held a backpack in front of their face to test whether the robot could continue tracking them under facial occlusion, seemingly delighted by this interactive feature. The system collects and exports all data in a manner that does not require post-processing in order to understand movement. It is simple to register the user's and robot's motions together in the same space to better understand the interaction. This capability removes the need for a human to painstakingly watch and score the user's performance either during the session or afterwards.

As seen from the purple and blue heatmaps for all three cues (Figures 2.7, 2.8 and 2.9), the upper-limb motion-capture data shows that participants had a propensity to mimic the robot (**H3**). Participant hands can be seen to hold the same position as those of the robot. The full-body motion-capture data recorded by the system can be used to detect exercise movements (Ashapkina *et al.*, 2019; Rybarczyk *et al.*, 2017; Gao *et al.*, 2020) using gesture recognition. Thus, the conducted study provided support for all three of our hypotheses about minimally supervised human-robot interaction.

Current Limitations

We were unable to get Captury's latest learning-based algorithm to function properly on our workstation, so we had to use an older algorithm. This choice led to several tracking issues such as sensitivity to light that filters in around the black-out curtains at different times of day and clothing color constraints for participants. In general, all Captury algorithms work best if participants wear pants, rather than skirts, so that their legs are visible. Captury tracking quality decreases when the user is significantly occluded by the robot, as occurs during hugging. Furthermore, our system is not yet self-aware enough to realize when the data it receives from Captury Live should not be trusted.

In addition, Captury cannot see the finer details of the user's pose that are key to expressiveness, such as the fingers and facial expressions; these aspects are clearly important and should be added to future systems. Because they were calm and safe, the tested robot cues were rather boring, simply moving to a pose, holding it, and moving back. Furthermore, users were very frustrated that the person running and monitoring the experiment would not answer their questions about the robot interaction. They wanted instructions or examples so they could know what to do. In tests with a broader audience, the setup might be perceived as deceptive because the robot's perceptual system is spread throughout the room rather than being limited to its own "eyes". Lastly, the small participant pool makes it hard to generalize our results, though they serve as a good validation for the desired functionality of the Robot Interaction Studio.

2.5 Potential Future Work

The features of the Robot Interaction Studio could be extended to perform other tasks such as mimicking user motion, pointing to the user's position within the workspace, reacting when the user enters certain regions of the workspace, and responding to social gestures such as high-fives, hugs, or handing over objects. Furthermore, the feed from the high-resolution camera attached to the robot's face screen could be used to perform facial expression analysis or eye gaze detection, yielding quantitative information about the user's emotion and visual attentiveness as they interact with the robot. It would also be possible to project information onto the floor of the Robot Interaction Studio to create a virtual game board. To make the system even more self-contained and resilient to tracking problems, we believe it would be beneficial for the user to know how their movement and pose are being seen by the robot. One approach would be to display a mirror image of the world as seen from Baxter's viewpoint, with the user's overlaid skeleton on top. This virtual thought bubble over the robot's head would enable the user to "read the robot's mind". It would also have an additional benefit of acting like a mirror, so the user could see their own motion from a third-person perspective.

For long-term deployment and practical use of the system for participant monitoring, we envision a ROS-based software suite that uses motion- and heart-rate-based metrics to quantify participant activity. The software can also record progress over time. Once set up, there is minimal expert knowledge required to run the system. We envision that future versions of the Robot Interaction Studio could be used in community rehabilitation settings such as therapy centers, care facilities for older adults, or even fitness studios. As shown by the promising results of the reported study, such systems have the potential to act as a standalone system where users can walk in and begin interacting with their robotic exercise coach with little to no assistance. This system can also be used to conduct other minimally supervised studies with robots and can reduce the load on video coding and annotation required for such studies.

Chapter 3

Formative and Summative Feedback in Minimally Supervised HRI

3.1 Introduction

With the introduction of a walk-in environment to facilitate minimally supervised human-robot interactions, our focus shifted towards identifying an effective method to provide feedback to users within these scenarios. Thus, this chapter showcases the potential of nonverbal gesture-based feedback for autonomous social robots by making two major contributions. First, we provide a paradigm for roboticists to consider when designing nonverbal feedback for minimally supervised HRI; our approach is based on the education literature, as described below. Second, we show how to quantitatively analyze the effectiveness of several different kinds of nonverbal robot feedback. Existing research in gesture-based interaction has heavily focused on qualitative research methods such as surveys (Ham *et al.*, 2015; Sauppé and Mutlu, 2014; Michaelis and Di Canio, 2022) and thematic analysis (de Wit *et al.*, 2020; Michaelis and Di Canio, 2022), which are time intensive and disconnected from the user's physical performance of the task. We believe movement-based interactions such as exercising should primarily be evaluated via quantitative metrics that can also be used to strengthen and adapt robot feedback (Morasso *et al.*, 2009).

Drawing inspiration from the education literature, we propose that nonverbal robot feedback about human activities should be divided into two types based on the two most common assessment methods: formative and summative. Formative assessment is performed during a course of instruction; it is meant to help identify the student's strengths and weaknesses and also instill autonomy, responsible learning, and self-evaluation skills within the student (Andrade and Cizek, 2010; Morrow Jr. *et al.*, 2015). Additionally, formative feedback supports educators by helping them devise subsequent instruction strategies (Andrade and Cizek, 2010). On the other hand, an assessment that is administered at the end of a unit of instruction and categorizes student performance (Andrade and Cizek, 2010; Morrow Jr. *et al.*, 2015) is called summative assessment. Thus, we define the following two feedback strategies for a socially assistive robot:

•*Formative feedback* is real-time feedback that is provided during the interaction

and aims to improve the user's competence at the task.

- *Summative feedback* is feedback that gives the user an overall assessment of their performance at the end of a particular task.

To evaluate the two proposed feedback types, we draw the following hypotheses from past work in education and extend them to the domain of minimally supervised nonverbal human-robot interaction:

H1 Formative and summative feedback independently improve user performance (McMillan, 2013; Hattie and Timperley, 2007).

H2 Formative feedback promotes user understanding of the task (McMillan, 2013; Andrade and Cizek, 2010).

Experimental support for these hypotheses, as provided by our rigorous user study, indicates that social robots can better support humans in performing physical exercise by providing nonverbal performance feedback both during and after each activity. Though more complex to implement, formative feedback is particularly powerful at facilitating successful interactions because it improves user understanding. Given their impact in this minimally supervised exercise scenario, these types of robot behaviors combining real-time human observation with gestural communication would likely also benefit autonomous robots working with people in other challenging settings.

The design of both formative and summative feedback is described in Section 3.2. The results from the user study are detailed in Section 3.3, followed by the discussion and conclusion in Sections 3.4 and 3.5 respectively.

3.2 Materials and Methods

3.2.1 Objectives

This study was designed to study nonverbal feedback from a social robot to a human user in a minimally supervised setting. Inspired by educational assessments, the feedback provided could be either formative or summative. We studied the effects of formative and summative feedback using the Robot Interaction Studio described in Section 2.1 and the three task cues designed to evaluate that system. Thus, the study described in Section 2.2 acts as a control condition. We extend the three task cues to form three additional conditions based on the type of feedback provided by the robot. Since the cues now represent specific actions that the participant must perform, we now rename the three task cue types based on their function: the orange cues are location cues, the purple cues are pose cues and the gray cues are contact cues.

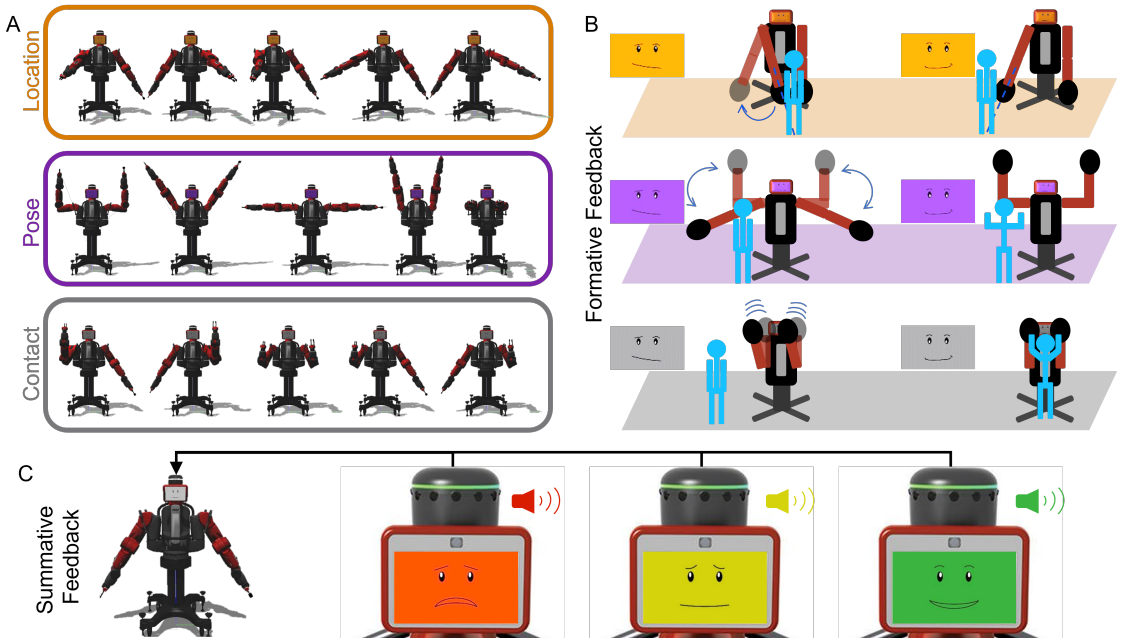


Figure 3.1: **(A)** The three cue categories (location, pose, and contact), which each had five variants. **(B)** The formative feedback adopted by the robot for each cue type. The location cue (top row) is modeled after pointing gestures. The robot points to a location on the ground within the Robot Interaction Studio for the user to move to. For formative feedback, the robot points to the user's current incorrect location and then to the desired location. The five variants of the pose cue (second row) show the symmetric arm pose the user should imitate. For formative feedback, the robot shows an intermediate arm pose between the desired pose and a neutral pose based on the magnitude of the user's current pose error. Finally, for the contact cue (third row), the robot signals for contact by holding up one or two of its hands similar to a high five. For formative feedback, the robot wiggles its forearm (elbow joint) for a duration proportional to the current distance from the user's hands to the robot's offered hand(s). **(C)** The way in which the robot conveyed summative feedback at the end of each trial, including presentation of a red, yellow, or green face with an animated facial expression and a matching auditory cue.

3.2.2 Study Design

This user study was designed as a mixed-design experiment with two between-subject factors with two levels each and one within-subject factor with three levels. The two between-subject factors were formative and summative feedback, with two levels each (with and without feedback). The within-subject factor was cue type, with the three levels being location cue, pose cue, and contact cue. We consider all five variants of each cue (different room locations, arm poses, and hand positions) to be largely equivalent to one another; thus, we do not consider the cue variant as a separate factor.

Each study participant tried to interpret the same three gesture-based robot cues under one of four feedback configurations: no feedback (fs), only formative feedback (Fs), only summative feedback (fS), or both types of feedback (FS). Fig. 3.1 shows the three cue types, their five variants, their associated types of formative feedback, and how summative feedback was conveyed across tasks. For all five variants of the three cue types (Fig. 3.1A), the robot gestures to try to make the user perform a specific task that is never verbally explained to the participant. For the location cue, the participant is supposed to go to the indicated location in the room and stand there. For the pose cue, the participant should mimic and hold the robot's arm posture. Finally, for the contact cue, the participant should constantly touch the robot's offered hand(s).

The study included 28 participants with 7 participants in each group. The study was approved by the Ethics Council of the Max Planck Society under the Haptic Intelligence Department's framework agreement as protocol F007A. Participants were recruited from the Max Planck Institutes in Stuttgart, the University of Stuttgart, and a local Facebook group for foreigners. Every participant provided written informed consent. Participants who were not employed by the Max Planck Society received a compensation of 8€ per hour for participating in this study. Since this study followed the same protocol as the evaluation study conducted for the Robot Interaction Studio, the finer details of the study are described in Section 2.2.2.

The participant first filled out a Robot Evaluation Survey (Table A.1), after which the experimenter read a set of pre-defined instructions out loud (Supplementary Materials Section A.1.1). Participants experienced 25 30-second-long trials of each type of cue, always presented in the repeating order location cue, pose cue, and contact cue for a total of 75 trials. Fig. 3.2 shows how the robot behaved over time during the study, highlighting three successive trials from one participant with plots of their error over time and sample video frames. Participants were told to do their best to interpret the robot's gesture-based cues to perform the tasks; the experimenter never explained any aspect of the robot's behavior and answered no questions.

At the beginning of a trial, the robot moved from the neutral pose to a cue-specific pose and then nodded its head to indicate that the trial was starting. Each trial lasted approximately 30 seconds after the robot moved into the desired pose. The system constantly checked whether the user was performing the desired task and calculated the score as the total duration of time that the task was performed correctly. If formative feedback was to be provided, it was triggered after two seconds in which the user error was continually greater than the defined threshold, and the provided feedback was based on the user's error at the end of that interval. Summative feedback, if present, was provided at the end of the 30-second interval.

The Robot Interaction Studio provided full-body tracking of the user as they interacted with the robot. These measurements were used in real time to drive the formative feedback provided by the robot during trials for participants in the Fs and FS groups, as shown in Fig. 3.1B. They were also used to calculate the score achieved on each trial by counting the amount of time the user performed the desired task with less than a specified

error threshold; the robot presented a binned version of this score to participants in the fS and FS groups at the end of each trial, as shown in Fig. 3.1C. The participant's heart rate (Fig. A.1) and facial expression were additionally collected throughout the study, and the user completed an extensive questionnaire (Section 2.2.2) after finishing all trials (Supplemental Material Section A.1.2).

3.2.3 Cue and Feedback Design

The initial task cue instruction that the robot provided was designed exactly as described in Section 2.2.1. Thus, each task cue was color coded: the location cue was represented by orange faces, the pose cue by purple faces, and the contact cue by gray faces. When presenting a task cue, the robot changed its face to an attentive expression (Fig. 3.7E, 3.8F, 3.9E). When formative feedback was present (Fs and FS conditions), the robot smiled when the user performed the task correctly and had an attentive face otherwise. User performance of each type of cue was determined by comparing the current value of an associated error metric defined for that task to a defined error threshold, which was selected to provide a moderate level of difficulty for healthy adults. If a formative feedback sequence was triggered by the error remaining above the threshold for two seconds, the face changed to a disappointed smirk while the formative gesture was being executed. After providing feedback, the robot's expression returned to switching between the smile and the attentive face based on the error. The robot used a white face and a neutral expression between task cues.

Location cue

For the location cue, which is based on deictic gestures (Bartneck *et al.*, 2020; McNeill, 1992), the robot points to a location on the ground; without being told, the user needs to figure out they are supposed to walk to that location and stand there for the duration of the trial. We quantified the user's error using the distance between the target and their current location on the floor. Thus, the error metric for the location cue (ε_ℓ) at time t was thus defined as the 2D Euclidean distance between the desired position ($\vec{p}$) where the robot was pointing to and the current projection of the user's hip position onto the ground ($\vec{u}_g(t)$), as follows:

$$\varepsilon_\ell(t) = \left|\vec{p} - \vec{u}_g(t)\right| \tag{3.1}$$

The threshold for the location cue error metric represents the radius of the circle within which the user was considered to be performing the task correctly. This constant threshold was defined as follows:

$$\bar{\varepsilon}_\ell = 0.3\,\mathrm{m} \tag{3.2}$$

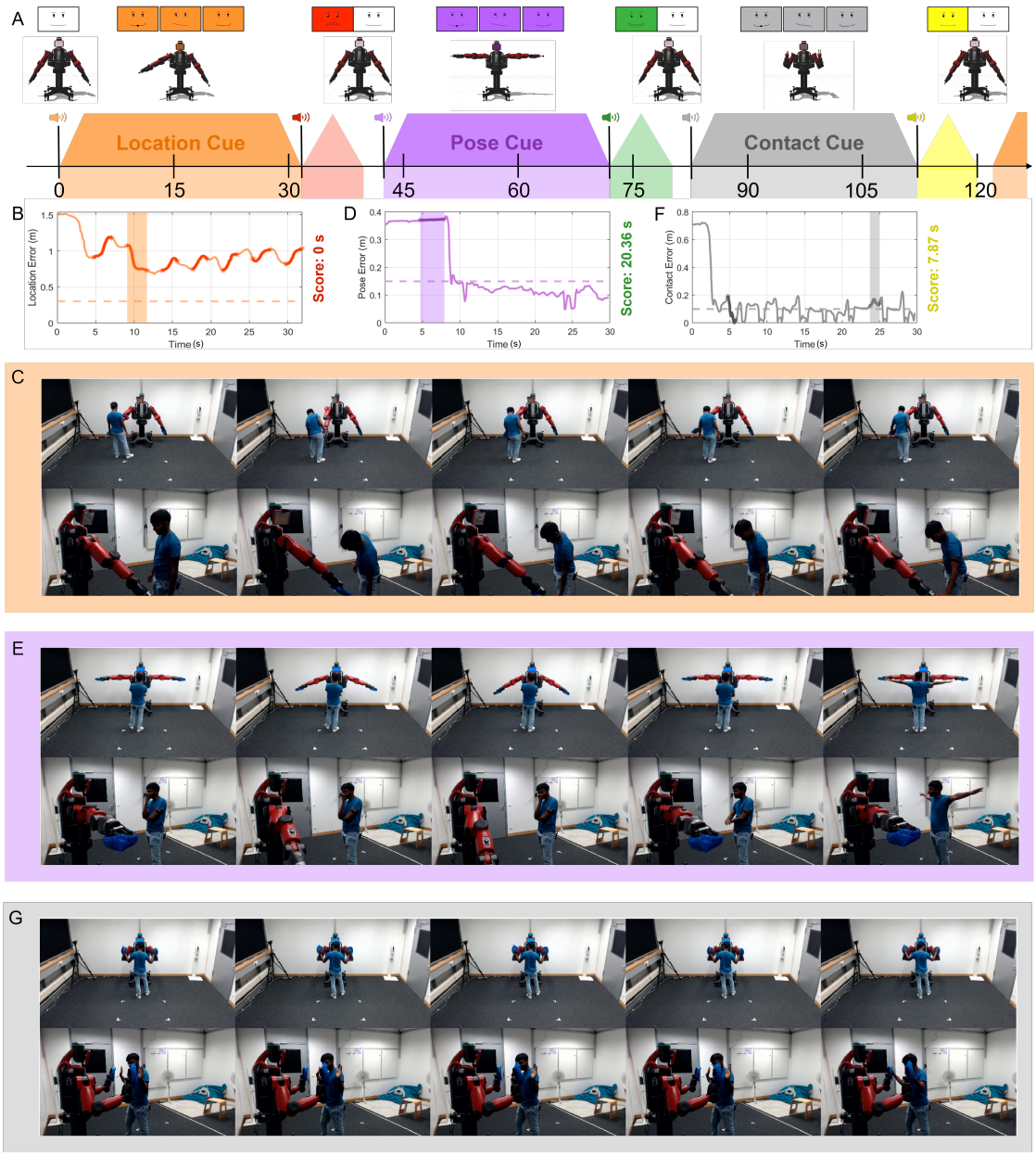


Figure 3.2: (Caption provided on the following page.)

Figure 3.2: **(A)** Sample timeline of three successive cues when both formative and summative feedback were provided (FS); the graphics above the timeline depict the robot facial expressions (Fitter and Kuchenbecker, 2016a) and arm gestures. At the end of each trial, the robot displays the summative feedback for several seconds before returning to its neutral pose. **(B)** User error over time for a sample trial of the location cue. The participant did not figure out this task and thus never achieved an error below the threshold (indicated with a dotted line), resulting in a final score of 0 seconds. The highlights in the line plot depict the seven instances of formative feedback. **(C)** Screenshots from one instance of formative feedback for the location cue. The user incorrectly attempts to hold the robot's hand; the robot points to the user's feet and then points to the correct location. **(D)** User error over time for a sample trial of the pose cue with only one instance of formative feedback; the user then copied the robot's arm pose. **(E)** Example formative feedback instance for the pose cue. The user starts out not performing the task. The robot provides formative feedback by moving its arms based on the error metric. The user then mimics the robot. **(F)** User error over time for a sample trial of the contact cue. The user repeatedly touches the robot's hand and then moves their hand back, causing oscillations in the error signal. Formative feedback is provided only when the user has not performed the task correctly for at least five seconds, which occurs twice. **(G)** Example formative feedback for the contact cue. The user can be seen mimicking the robot. The robot then performs a hand wiggle as formative feedback, and the user touches the robot's palm. All sample trials shown in this figure were taken from a participant who received both types of feedback and gave consent to share video.

The robot provides formative feedback for the location cue by moving its arm to point at the feet of the user standing outside the target circle and then again at the target position; therefore, its arm motion shows the magnitude and direction of the user's current task error.

Pose cue

The pose cue takes inspiration from mimicry or imitation-based nonverbal communication (Bartneck *et al.*, 2020; Wang *et al.*, 2015b); without being instructed, the user had to figure out that they should mimic the robot's arm pose for the duration of the trial. The error metric for the pose cue (ε_p) at time t is a curve-similarity measure based on Frechet's distance (Fréchet, 1957). For each robot arm, this metric can be defined as the sum of the shortest distance from every joint on the robot arm to a line-segment-based representation of the corresponding human arm. This metric is calculated after performing homogeneous transformations to scale the human arms to the length of the robot arms, mirror them, and transform them to start at the same locations as the robot arms.

Each robot arm has $N = 7$ revolute joints. We define $d_{Lj}(t)$ to be the shortest distance from the current location of the reference frame attached to joint j of the left robot arm

to the closest segment on the corresponding human arm; $d_{Rj}(t)$ has the same meaning for the robot's right arm. To ensure that both human arms are mimicking the pose of the respective robot arm, we define the overall pose error to be the minimum of the errors for the left and right arms. ε_p can then be calculated as follows:

$$\varepsilon_p(t) = \max\left(\sum_{j=1}^{n} d_{Lj}(t), \sum_{j=1}^{n} d_{Rj}(t)\right) \tag{3.3}$$

Thus, the error metric for the pose cue had to fall below the error threshold for both arms simultaneously for the user's action to be judged as correct.

The threshold for the pose cue error metric represents the total summed distance allowed between the pose of the user and that of the robot. This constant threshold was defined as follows:

$$\bar{\varepsilon}_p = 0.15\,\mathrm{m} \tag{3.4}$$

The formative feedback of the pose cue involves the robot moving its arms to an intermediate pose that is between the desired pose and the neutral pose, thereby demonstrating the amplitude of the user's current error.

Contact cue

Contact cues emulate variations on the high five, a simple interpersonal gesture (Gallace and Spence, 2010). Again, without being told the goal, the user needed to figure out that they ought to walk into the robot's workspace and touch its palm(s) with their hand(s) for the duration of the trial. For cues of this type, the user's instantaneous error was calculated as the distance between the robot's offered hand(s) and the closest human hand(s). Thus, the error metric for the contact cue (ε_c) at time t was defined as the Euclidean distance from the center of the offered robot hand ($\vec{b}_c$) to the center of the user's closest hand in 3D space. Assuming that $\vec{u}_R(t)$ and $\vec{u}_L(t)$ are the time-varying positions of the user's right and left hands, respectively, ε_c can be calculated as follows:

$$\varepsilon_c(t) = \min\left(\left|\vec{b}_c - \vec{u}_L(t)\right|, \left|\vec{b}_c - \vec{u}_R(t)\right|\right) \tag{3.5}$$

A special case of the contact cue occurs in the fifth variant of this cue when the robot offers both hands to the user. In this case, $\vec{b}_R$ and $\vec{b}_L$ are the right and left robot hand positions which we defined to be at the center of the respective parallel-jaw gripper. To enforce that both hands of the user should touch the robot hands, ε_c is defined as the maximum of the minimum distances from each of the user's hands to either of the robot's hands, as follows:

$$\varepsilon_c(t) = \max\left(\min\left(\left|\vec{b}_L - \vec{u}_R(t)\right|, \left|\vec{b}_R - \vec{u}_R(t)\right|\right), \min\left(\left|\vec{b}_L - \vec{u}_L(t)\right|, \left|\vec{b}_R - \vec{u}_L(t)\right|\right)\right) \tag{3.6}$$

In this case, each minimum distance must separately fall below the error threshold, i.e., each user hand must be near one of the robot's hands.

The threshold for the contact cue error metric represents the radius of the sphere within which the user was considered to be performing the task correctly. This constant threshold was defined as follows:

$$\bar{\varepsilon}_c = 0.1\,\text{m} \tag{3.7}$$

For formative feedback with the contact cue, when the error is above this threshold, the robot wiggles its hand(s) an integer number of times equal to the error rounded up to the nearest meter.

Summative Feedback

Summative feedback was presented consistently across all cue types, involving one of three differently colored facial expressions with sounds that showed the quality of the user's performance on the just-completed trial (Fig. 3.1C). If the participant error was less than the threshold for less than 10% of the 30-second trial time ($< 3\,\text{s}$), the robot displayed a sad red face and played an unhappy sound. If the participant performed the task within the desired limits for any duration between 10% and 60% of the time (3 – 18 s), the robot displayed a slightly happy yellow face and played a matching sound. Finally, if the participant performed the task correctly for more than 60% of the trial ($> 18\,\text{s}$), the robot displayed a happy green face and played a happy sound.

The user score (S) is a quantitative representation of summative feedback. The value of the score (S) at time t for each cue type was calculated as a function of the error over time. Specifically, the score for a particular trial is equal to the total duration of time the user spent performing the task with an error less than the error threshold defined for that task. Scores varied continuously from 0 s to 30 s (perfect performance) and were calculated during the trial as follows, where Δt is the time elapsed since the previous discrete time step:

$$S(t) = \begin{cases} 0 & \text{if } t = 0\,\text{s} \\ S(t-\Delta t) + \Delta t & \text{if } \varepsilon(t) < \bar{\varepsilon} \\ S(t-\Delta t) & \text{otherwise} \end{cases} \tag{3.8}$$

The final score for a trial was taken to be the score at the trial's end, i.e., $S(30\,\text{s})$.

Other Sources of Data

In addition to the metrics presented above, we were also interested in evaluating user activity and task cue comprehension. Thus, we looked at the following additional metrics.

User Grade: We created rubrics to evaluate the participants' written post-study explanations of what the robot wanted them to do for each cue type. Participants received one point each for correctly identifying how the robot indicated the cue, their expected response, and the duration component of the task. An external individual graded all explanations against these three-point rubrics with one-point resolution; they evaluated each set of responses in random order and were blinded to the participant's feedback condition. The grader was asked to focus on content, not specific words, and was allowed to award only integer scores, with no partial points. The possible grades were 0, 1, 2, and 3 (perfect comprehension).

For grading responses pertaining to the location cue, the total score for their response is the sum of the points earned for describing each of the following:

1 Point	Robot is pointing to location/position/place
1 Point	Participant must move to location/position/place/through the workspace
1 Point	Participant must stay in location/position/place

For grading responses pertaining to the pose cue, the total score for their response is the sum of the points earned for describing each of the following:

1 Point	Robot is making a pose
1 Point	Participant must mimic/copy pose
1 Point	Participant must hold pose

For grading responses pertaining to the contact cue, the total score for their response is the sum of the points earned for describing each of the following:

1 Point	Robot is lifting hand/holding out hand for contact
1 Point	Participant must make contact with robot's hand pads
1 Point	Participant must make repeated contact or hold contact with robot's hand pads

Participant Activity Measures: In addition to the metrics described in Section 2.2.3 (distance traveled and time spent in the robot's workspace), we also considered an attention metric which uses the user's head position and orientation to estimate their point of focus. The videos from a camera placed on Baxter's head were analyzed using the Affectiva AFFDEX (McDuff *et al.*, 2016) facial emotion recognition algorithm integrated into the iMotions software (iMotions A/S, 2020) to provide this information.

Similar to Section 2.2.3, heatmaps were created by discretizing the relevant plane into a 20×20 grid and then binning the position data. Corresponding distributions representing the desired user behaviors for each cue type were created using a 2D normal distribution with a mean equal to zero and a standard deviation equal to the relevant error threshold divided by three. The spread of this distribution was chosen such that approximately 99.6% of the points fall within the desired area, to represent perfect user behavior.

This distribution was then translated to be centered at the target location for the user's hips or hand(s), depending on the cue type and variant. We generated the same number of points as the actual data. The Wasserstein metric was calculated between these desired distributions and the actual distributions.

Though we collected participant heart rate throughout the study as beats per minute over time, we did not acquire a baseline for this data. Thus, we do not analyze the heart rate measurements for the purposes of this study. The raw measurements can be seen in Appendix Section A.1.3

3.2.4 Participant Information

The study included 28 participants with 7 participants in each group. The fs group (4 male, 3 female; aged 23–33) received no formative or summative feedback. The Fs group (3 male, 4 female; aged 19–39) received only formative feedback. The fS group (3 male, 4 female; aged 19–65) received only summative feedback, and the FS group (4 male, 3 female; aged 22–42) received both types of feedback. Fig. 3.3 shows how participants responded to four questions about their prior experience with robots, expe-

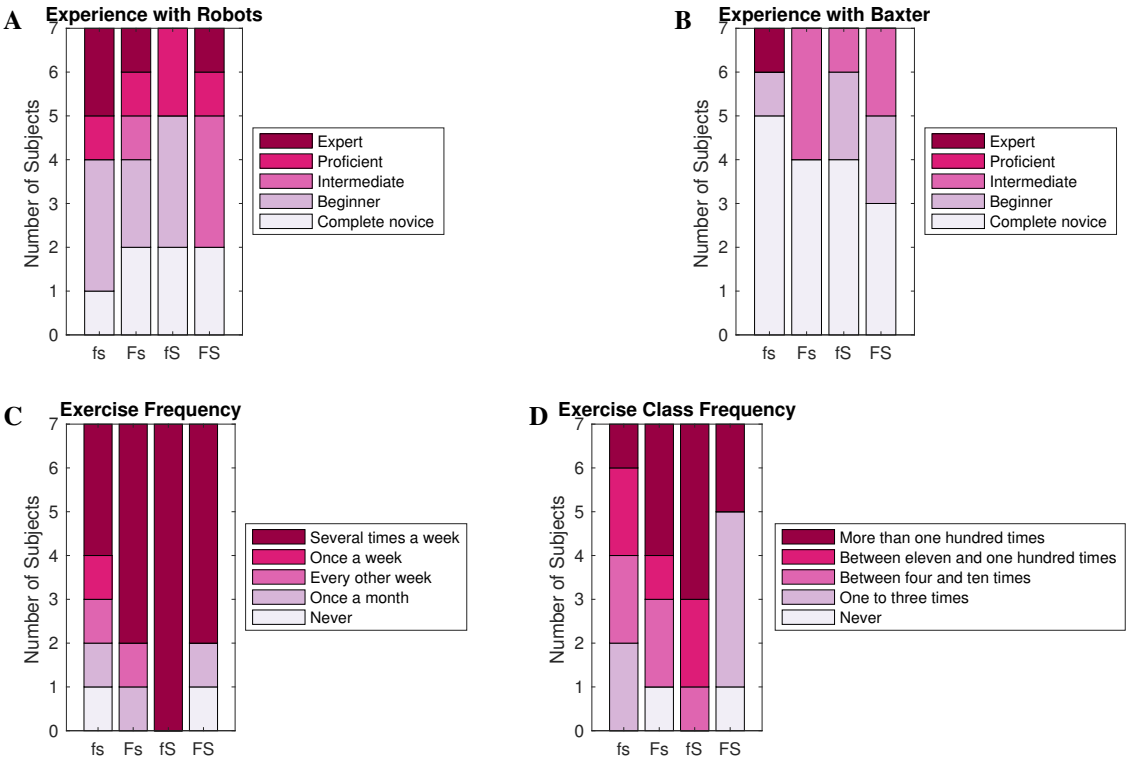


Figure 3.3: **(A)** Participant self-reported experience with robots. **(B)** Participant self-reported experience with a Baxter robot. **(C)** Participant self-reported exercise frequency. **(D)** Participant self-reported experience with exercise classes.

rience with Baxter, exercise frequency, and exercise class frequency. The full text of these questions is listed in Sec. A.1.2. The participants varied in their experience with robots, including 7 complete novices and 4 experts (Fig. 3.3A); only one participant had any substantial previous experience with Baxter (Fig. 3.3B). Most participants led relatively active lifestyles, reporting that they exercised several times a week (Fig. 3.3C), with some frequently attending exercise classes (Fig. 3.3D).

3.2.5 Statistical Analysis

All data were analyzed in R (version 4.2.3) using non-parametric Aligned Rank Transform (ART) analysis of variance (ANOVA) proposed by Wobbrock *et al.* due to the small sample size per group (Corder and Foreman, 2011) and because the data were non-normal. The unit of analysis for all analyses is the number of participants ($N = 28$ where $N = 7$ per group) except for the two-way ART ANOVA used for analyzing the Wasserstein metrics, which used the cue variants as the unit of analysis ($N = 5$). Pairwise post-hoc comparisons were conducted using the multiple comparisons for the ART procedure (Elkin *et al.*, 2021) with a Bonferroni correction and the Kenward-Roger method for approximating the degrees of freedom (Midway *et al.*, 2020). We report effect sizes using partial eta squared (η_p) for the ANOVA tests and Cohen's d for t-tests (Lakens, 2013). The significance-testing level, α, was considered to be 0.05 in all cases.

Missing Data

We do not have any data for two trials for one participant in the Fs condition due to an experimenter error. For the purposes of statistical analysis, these two trials were populated with the median of the dependent variable in question based on the other trials of the same cue type recorded for this participant.

3.3 Results

We examined our hypotheses quantitatively, evaluating the dynamic and unstructured behavior of naive users in response to formative and summative feedback delivered by a robot during minimally supervised activity sessions. The robot's behavior was carefully designed to convey the task cues and both types of feedback clearly and consistently via iterative pilot testing. To test our hypotheses, we evaluated the study results in terms of participant performance, cue comprehension, confidence in cue comprehension, activity, perception of robot behavior, self-reported cue exploration, and other survey responses.

3.3.1 Performance

For all three cues, the main outcome measure is the score the user achieved in each trial, calculated as the duration of time they performed the desired action with an error less than the defined threshold. A three-way non-parametric Aligned Rank Transform (ART) ANOVA (Wobbrock *et al.*, 2011) was performed to evaluate the effects of formative feedback, summative feedback, and cue type on the user's average score (Fig. 3.4), which could range from 0 s to 30 s. Importantly, the main effects of formative feedback ($F(1,24) = 10, p = 0.0036, \eta_p^2 = 0.3$), summative feedback ($F(1,24) = 7.4, p = 0.012, \eta_p^2 = 0.24$), and cue type ($F(2,48) = 24, p = 6 \times 10^{-8}, \eta_p^2 = 0.5$) were all significant. None of the interactions were significant, but there was a trending three-way interaction between formative feedback, summative feedback, and cue type ($F(2,48) = 3.1, p = 0.052, \eta_p^2 = 0.12$). Post-hoc analyses revealed that the presence of either formative feedback ($t(24) = 3.2, p = 0.0036, d = 1$) or summative feedback ($t(24) = 2.7, p = 0.012, d = 0.94$) significantly improved user performance. Furthermore, participants performed better in the pose cue ($t(48) = 5, p = 2.4 \times 10^{-5}, d = 1.3$) and the contact cue ($t(48) = 6.6, p = 7.7 \times 10^{-8}, d = 1.8$) compared to the location cue.

3.3.2 Cue Comprehension

Another important outcome measure is how well the participants understood the three task cues. As shown in Fig. 3.5A, we used an integer scale from 0 to 3 to grade their post-study written explanations of what the robot wanted them to do for each cue type (Supplemental Material Section 3.2.3). Interestingly, there were significant main effects of formative feedback ($F(1,24) = 4.81, p = 0.038, \eta_p^2 = 0.17$) and cue type ($F(2,48) = 6.8, p = 0.0025, \eta_p^2 = 0.22$) on the cue comprehension grade (Figs. 3.5B and 3.5D). In contrast, as shown in Fig. 3.5C, the main effect of summative feedback was not significant ($F(1,24) = 0.35, p = 0.56, \eta_p^2 = 0.01$). The grade was significantly higher for participants who received formative feedback ($t(24) = 2.2, p = 0.038, d = 0.56$), and it was higher for the pose cue ($t(48) = 3, p = 0.013, d = 0.8$) and the contact cue ($t(48) = 3.4, p = 0.0045, d = 0.9$) in comparison to the location cue. Additional self-reported metrics related to cue understanding can be seen in the supplementary materials (Figs. 3.10 and 3.11). Finally, the lack of explicit task instructions, the experimenter's refusal to answer questions, and the resulting uncertainty about what to do tended to frustrate participants, as evidenced by their slightly more negative perception of robots after the study (Supplemental Material Section 3.3.7). Feedback, however, positively affected the participants' feelings of self-efficacy (Fig. 3.14D).

3.3.3 Confidence in Cue Comprehension

The post-study survey also asked participants to rate their confidence in their explanation of the meaning of each of the robot's cues. The comprehension scores appear in

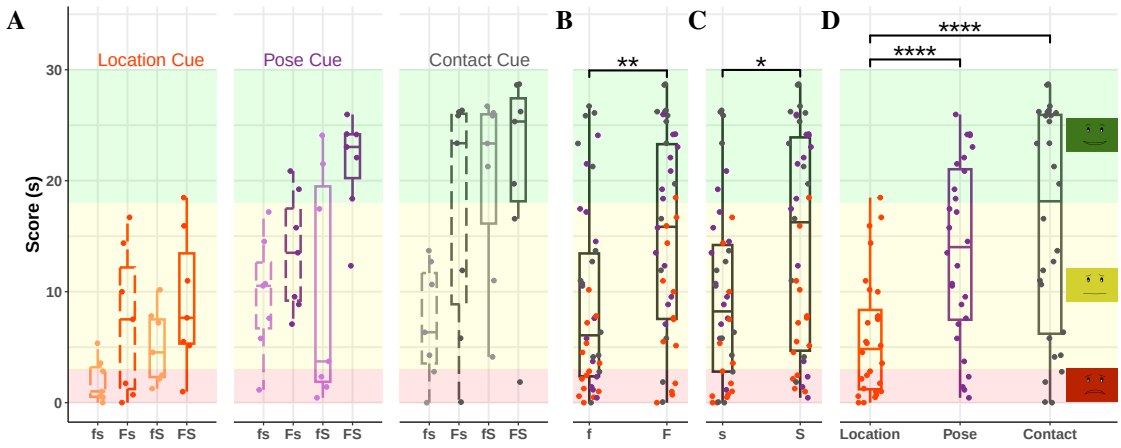


Figure 3.4: **(A)** Boxplots of mean participant score for each condition (no feedback (fs), only formative feedback (Fs), only summative feedback (fS), or both types of feedback (FS)) and cue types. Higher scores are better, and the maximum is 30 s; the background colors depict the three levels of summative feedback. Participants performed better when the robot provided either type of feedback. **(B)** Boxplots of mean participant scores separated by level of formative feedback. Formative feedback significantly increased participant scores. **(C)** Boxplots of mean participant scores separated by level of summative feedback. Summative feedback also significantly increased participant scores. **(D)** Boxplots of mean participant scores separated by cue type. Participants earned higher scores in the pose cue and the contact cue compared to the location cue. For all boxplots, the dots mark the data points (one per participant per cue), the central line shows the median, the box shows the lower and upper quartiles, and the whiskers show the range up to 1.5 times the interquartile range. Significance notation: $\star$ $p < 0.05$, $\star\star$ $p < 0.01$, $\star\star\star$ $p < 0.001$, and $\star\star\star\star$ $p < 0.0001$.

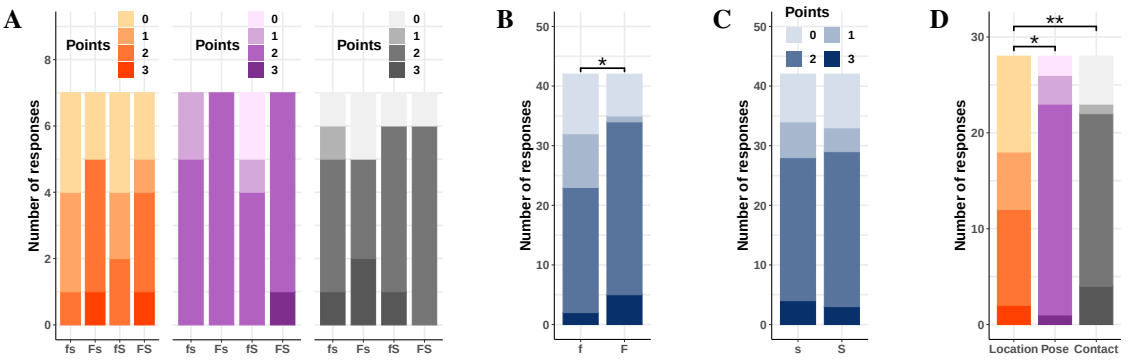


Figure 3.5: **(A)** Grades of participant comprehension of each task cue based on structured grading of written explanations (seven participants in each feedback condition). **(B)** Grades separated by level of formative feedback (fourteen participants in each formative feedback level with responses for all three cue types). Formative feedback significantly improved understanding, with more participants scoring 2 or 3 points out of 3. **(C)** Grades separated by level of summative feedback (fourteen participants in each summative feedback level with responses for all three cue types). Summative feedback did not affect user understanding. **(D)** Grades separated by cue type (all 28 participants for each cue type). Participants understood the pose and contact cues better than the location cue. Significance notation: $\star$ $p < 0.05$ and $\star\star$ $p < 0.01$.

Fig. 3.5, and the corresponding confidence results are shown in Fig. 3.11, separated by feedback condition. Participants used the entire range of confidence values from 0 to 10, as this self-reported measure is highly subjective. Confidence was generally highest for the contact cue, followed by the pose cue, and then the location cue. A three-way ART ANOVA showed there was a statistically significant two-way interaction between formative feedback and cue type on the user's confidence (Fig. 3.11) at interpreting the cue $(F(2,48) = 4.1, p = 0.023, \eta_p^2 = 0.15)$. However, we did not find any significant effects upon conducting post-hoc analyses. Visual inspection seems to show that the combination of formative and summative feedback (FS) might have made participants most confident of their understanding of the pose cue, and only summative feedback (fS) may have elicited the highest confidence for the location cue and contact cue. Interestingly, formative feedback points out mistakes as the user makes them and might have thus made participants more aware of their errors, reducing their confidence while also increasing their comprehension; the relationship between feedback, task difficulty, task comprehension, and confidence in comprehension deserves further study.

3.3.4 Activity

The effectiveness of formative and summative feedback can be understood by examining the actions participants performed as they tried to figure out the tasks cued by the robot in

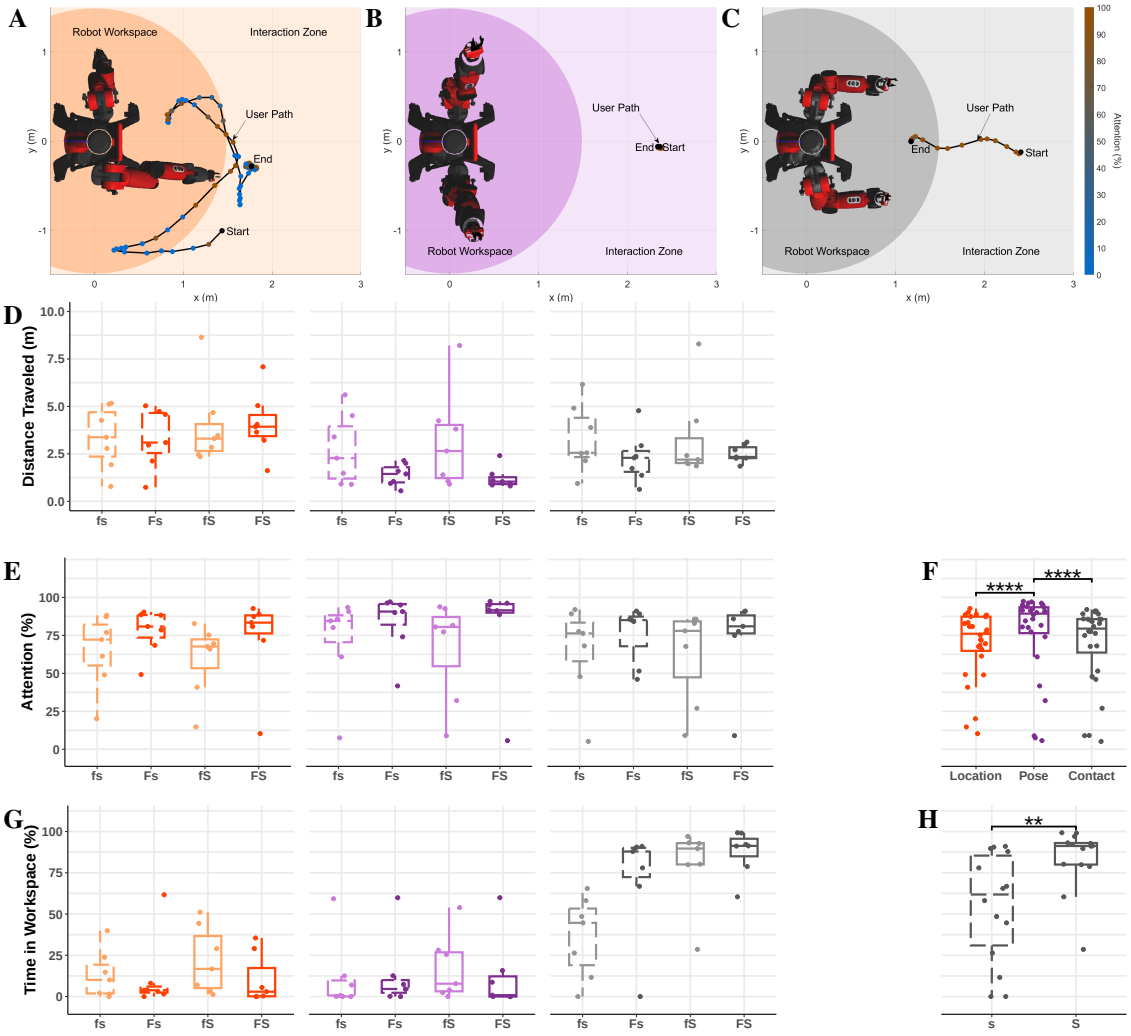


Figure 3.6: **(A)** Overhead view of a sample location cue trial from the condition with both formative and summative feedback (FS). The user walks around the room, leaving and entering the robot's workspace and occasionally looking at the robot's face. **(B)** Sample pose cue trial. The user stands in one position to perform the task. **(C)** Sample trial of contact cue. The user walks into the robot's workspace to touch its hands. **(D)** Mean distance travelled by each participant for each type of cue, organized by the four feedback conditions (fs, Fs, fS, and FS). **(E)** Mean attention level that the user paid to the robot. **(F)** Users paid significantly more attention to the robot during the pose cue compared to the location and contact cues. **(G)** Mean percentage of trial time that the participants spent within the workspace of the robot. **(H)** Summative feedback significantly increased the time participants spent in the robot's workspace for the contact cue. Though formative feedback also had a significant interaction with cue type for this metric, post-hoc results revealed no interesting observations. Significance notation: $\star\star$ $p < 0.01$ and $\star\star\star\star$ $p < 0.0001$.

each of the four feedback conditions. As shown in Fig. 3.6, we evaluated their behavior in terms of the distance they travelled within the room, their level of attention to the robot's face, and the amount of time they spent within the robot's reachable workspace.

Furthermore, heatmap visualizations were created to show the cumulative spatial distribution of the user's position in the room, their hand positions relative to their body, and their hand positions in the room across all trials of each of the five variants of the location cue (Fig. 3.7), the pose cue (Fig. 3.8), and the contact cue (Fig. 3.9). For these visualizations, the planar space under consideration was discretized into a grid, and the color shows the percentage of total trial time users spent in each grid cell. The resulting heatmaps provide a dense overview of what transpired during the study; for conciseness, we show all four feedback conditions for the heatmap type most relevant to each cue and only the condition with both formative and summative feedback (FS) for the other two heatmaps. The supplementary materials (Figs. A.2, A.3, and A.4) show all three heatmap types for all four feedback conditions across all three task cues.

To test whether formative and/or summative feedback changed how users moved for each task, we quantitatively analyzed the heatmaps by calculating the Wasserstein metric (Rüschendorf, 1985) (also known as the earth mover's distance) between each measured distribution and a normal distribution representing correct user behavior, which was centered on the target position and dispersed by the acceptable threshold. Heatmaps that are closer to the desired user behavior have a smaller Wasserstein metric.

Location cue

The desired behavior was standing where the robot pointed. In general, the presence of formative feedback tended to help guide participants toward the correct position (Fig. 3.7A). Although this task was difficult for most participants (low scores, high Wasserstein metrics), two resourceful individuals figured out most variants of this cue with only summative feedback. Participants would sometimes try mimicking the robot's arm poses, as seen in the blue and purple heatmaps (Fig. 3.7B). They frequently walked to and grasped the robot's pointing hand, as seen in the gray user-hand-position heatmap in Fig. 3.7C, which was similar to but not the same as the desired action. Despite these visual trends, a two-way Aligned Rank Transform (ART) (Wobbrock *et al.*, 2011) ANOVA evaluating the Wasserstein metric showed that neither type of feedback had a significant effect on how close users stood to the target location for this cue. A three-way ART ANOVA showed that the total distance travelled by the participant per trial ($F(2,48) = 5.4, p = 0.0076, \eta_p^2 = 0.18$) had a statistically significant two-way interaction between formative feedback and the type of cue (Fig. 3.6); however, we did not find any significant simple effects at $\alpha = 0.05$ upon further analysis of this interaction.

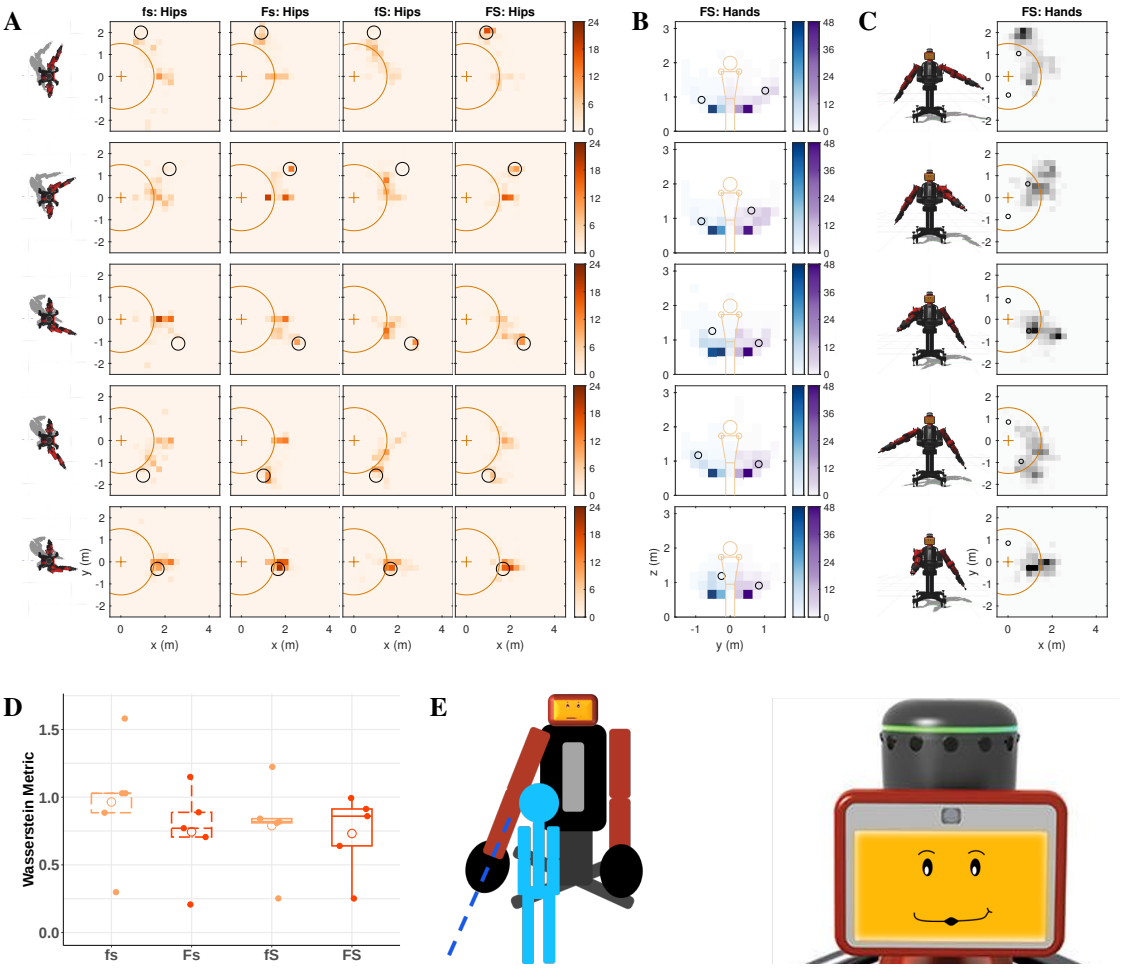


Figure 3.7: User activities across different feedback conditions of the location cue. **(A)** Heatmap visualizations of user position in the workspace of the Robot Interaction Studio. Across all conditions, users can be seen spending time at various locations around the room. When feedback is provided, the users tended to spend more time near the target position depicted by the black circle. **(B)** Heatmap visualization of the front view of the user's hands for the FS condition. Participants sometimes mimicked the robot's arm pose, depicted by the black circles, and they usually kept their arms at their sides as they walked to and stood at the correct location. **(C)** Heatmap visualization of the overhead view of the user's closer hand in the horizontal plane for the FS condition. This heatmap resembles the FS condition of the user position heatmap because users often walked to the desired location. **(D)** shows the Wasserstein metric representing how much participant position in the room deviated from the correct location. **(E)** A sample location cue and the face associated with the cue seen by participants across all conditions.

Pose cue

For cues of this type, participants generally performed the correct action of mimicking the robot across all feedback conditions. This behavior can be seen by the high concentration of blue and purple near the desired hand positions in all heatmaps (Fig. 3.8B). A two-way ART ANOVA on the Wasserstein metric based on the heatmap data revealed that there was a statistically significant main effect of formative feedback on the hand position ($F(1,16) = 22, p = 0.00023, \eta_p^2 = 0.58$). The presence of formative feedback significantly reduced the Wasserstein metric ($t(16) = 4.7, p = 0.00023, d = 2.1$), thus indicating an improvement in replicating the robot's movement. The overhead heatmaps show that participants tended to stand in the center of the room to perform the desired pose; accordingly, visual inspection of the pose cue data in Fig. 3.6 shows a statistically non-significant trend of formative feedback reducing the distance the user traveled ($F(1,26) = 5.4, p = 0.084, \eta_p^2 = 0.17$). Interestingly, users looked at the robot's face more often during the pose cue, as shown by a significant effect of cue type on attention ($F(2,48) = 20, p = 5.1 \times 10^{-7}, \eta_p^2 = 0.45$) with post-hoc analyses indicating that participants paid more attention to the robot for this cue in comparison to the contact cue ($t(48) = 4.6, p = 9.7 \times 10^{-5}, d = 1.2$) and the location cue ($t(48) = 6, p = 6.4 \times 10^{-7}, d = 1.6$). There is also a statistically trending effect of formative feedback on attention ($F(1,24) = 3.78, p = 0.06, \eta_p = 0.14$). Visual inspection of these data hint that formative feedback may increase attention to the robot for the pose cue, though this trend did not reach significance.

Contact cue

The desired behavior for the contact cue was to touch the robot's hands and maintain contact for the duration of the trial. Occasionally, participants incorrectly mimicked the robot's pose for cues of this type. Many participants were able to figure out this cue even without feedback. However, either type of feedback helped participants identify the correct behavior for this cue, as characterized by the top view of the hand positions (Fig. 3.9C), which shows darkly colored tiles overlapping with the correct hand positions. The two-way ART ANOVA on the Wasserstein metric showed that there was a two-way interaction between summative and formative feedback ($F(1,16) = 50, p = 2.6 \times 10^{-6}, \eta_p^2 = 0.76$). The presence of only formative ($t(16) = 5, p = 0.00079, d = 1.6$), only summative ($t(16) = 10, p = 1.6 \times 10^{-7}, d = 3.2$), or combined feedback ($t(16) = 15, p = 4.6 \times 10^{-10}, d = 4.7$) significantly improved the user's ability to perform the task. Additionally, participant performance was better when only formative feedback was present when compared to pure summative feedback ($t(16) = 5, p = 0.00079, d = 1.6$). The presence of both types of feedback led to a significantly better performance than providing only formative ($t(16) = 10, p = 1.6 \times 10^{-7}, d = 3.2$) or only summative feedback ($t(16) = 5, p = 0.00079, d = 1.6$).

Participants largely figured out that they needed to enter the robot's reachable

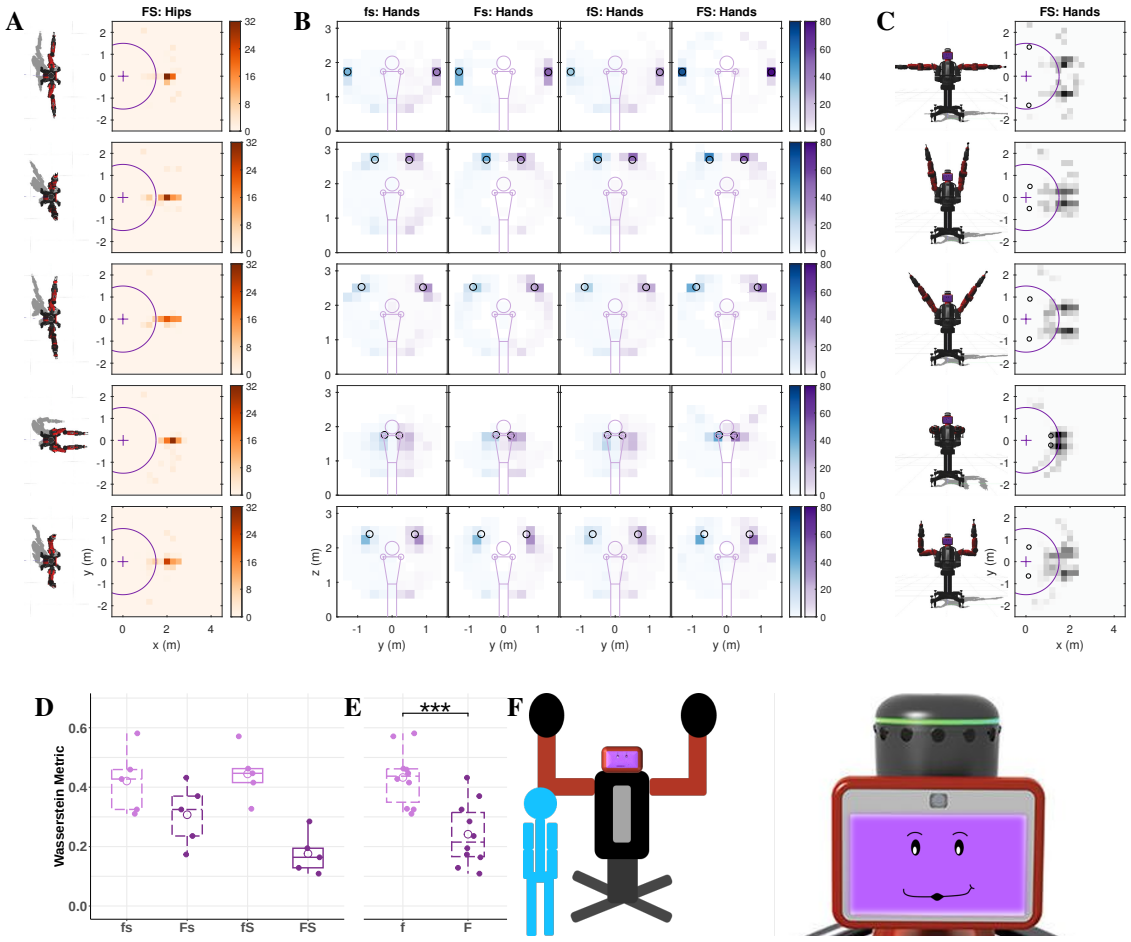


Figure 3.8: User activities across different feedback conditions of the pose cue. **(A)** Heatmap visualizations of user position in the workspace of the Robot Interaction Studio for the FS condition. Users tend to stand in front of the robot. **(B)** Heatmap visualizations of the front view of the user's hand positions. In general, participants across all conditions figured out the desired behavior of mimicking the robot's arm pose for this cue. However, their imitation of the pose was better when feedback was provided. **(C)** shows the overhead view of the user's hand(s) in the horizontal plane for the condition where both types of feedback were provided. Since participants generally stood in front of the robot, and since all poses are symmetric, one can see both the left and right hands. **(D)** Wasserstein metric representing how much participant hand positions deviated from the correct hand positions. The data presented here is the grand mean of the Wasserstein metric between actual and desired hand positions for both left and right hands. We look at the front view in this case, since the focus is on the arm pose. **(E)** Wasserstein metric separated by level of formative feedback. The presence of formative feedback had a positive impact on pose cue performance. Participants mimicked the robot more accurately when they had formative feedback to help correct their pose. **(F)** A sample pose cue and the face associated with the cue seen by participants across all conditions.

workspace to correctly perform this cue (Fig. 3.6). A three-way ART ANOVA on the time the participant spent in the robot's workspace showed that both formative feedback ($F(2,48) = 8.2, p = 0.00089, \eta_p^2 = 0.25$) and summative feedback ($F(2,48) = 5.3, p = 0.0086, \eta_p^2 = 0.18$) had significant two-way interactions with cue type. Post-hoc analyses revealed that the simple main effect of summative feedback ($F(1,26) = 13, p = 0.0033, \eta_p^2 = 0.34$) was significant only for the contact cue, with further analysis indicating that the time spent in the workspace was significantly higher when summative feedback was present ($t(26) = 3.7, p = 0.0011, d = 1.4$). Though there was a statistically significant interaction between formative feedback and cue type ($F(2,48) = 8.2, p = 0.00089, \eta_p^2 = 0.25$), post-hoc analyses did not reveal additional significant results.

3.3.5 Participant Perception of Robot Behavior

To understand the influence of combining robotic arm gestures, head motions, facial expressions, and sounds to convey the task cues and associated feedback, we asked participants to identify the ways in which the robot had communicated. Table 3.1 shows which robot actions participants remembered experiencing during the study. All but one of the participants noticed the head movements of the robot; the face screen rotated to face the user throughout the study in all feedback conditions. Interestingly, four of the 28 participants (14.3%) did not indicate that they saw the robot make arm gestures even though it moved its arms to convey every cue in the study. All participants who had formative feedback noticed that the robot had facial expressions, and all but one of them noticed the robot's nonverbal sounds. Most participants without formative feedback also noticed the facial expressions and nonverbal sounds. All participants correctly identified that the robot did not have lights on its hands or speak. In summary, participants across all feedback conditions correctly noticed the multimodal aspects of the robot's behavior.

Table 3.2 shows the robot actions participants wished for. Interestingly, most participants (85.7%) wished for the robot to speak, a modality commonly used by robots but not employed in this study to focus on nonverbal communication. Facial expressions (78.6%), head gestures (67.9%), and arm gestures (60.7%) were also requested by more than 60% of participants. Only about half (53.6%) of users wished for nonverbal sounds. A minority (28.6%) wished for lights on the robot's hands, though none of the participants who received both formative and summative feedback (FS) were interested in such a feature. These preferences validate the important role non-verbal communication can play in complementing verbal cues.

3.3.6 Participant Self-Reported Cue Exploration Measures

The post-study survey asked participants to report how often they performed movements only to find out how the robot would react, i.e., to explore how the system worked. We

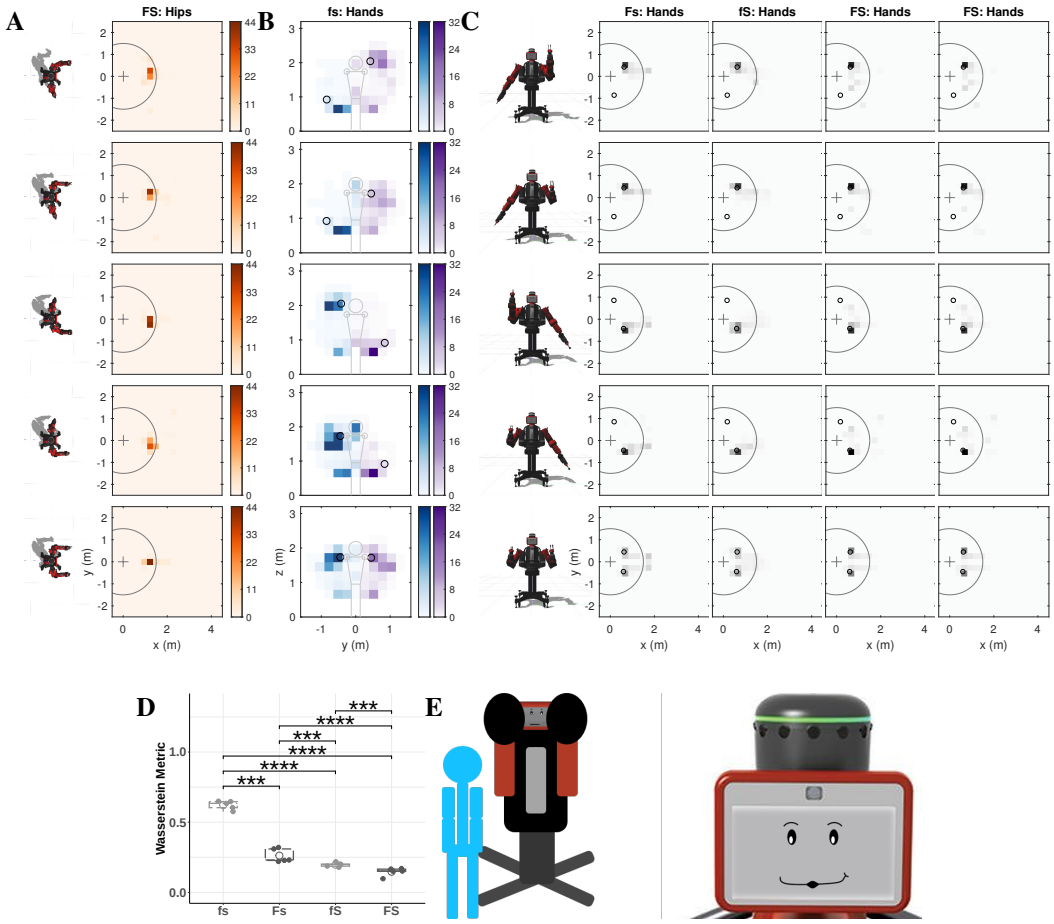


Figure 3.9: User activities across different feedback conditions of the contact cue. **(A)** Heatmap visualizations of user position in the room for the FS condition. Participants generally stood directly in front of the robot to perform this task. **(B)** Heatmap visualization of the front view of the user's hands relative to their body for the FS condition. From this angle, it appears as though the user is mimicking the robot's pose since they often hold their arms in a similar way to touch the robot's hand(s). **(C)** Heatmap visualizations of an overhead view of the user's hand(s) for all conditions. The first four rows show the position of the user hand that was closest to the robot's offered end-effector; both robot hands were presented in the fifth variant of this cue, so that row depicts both hands of the user. This task cue was the easiest of all cues. Almost all participants figured out the solution to this cue when feedback was provided: they entered the robot's workspace and touched the robot's offered hand(s). **(D)** Wasserstein metric representing how much participant hand positions in the room deviated from the correct hand positions. There was a significant interaction between formative and summative feedback. All feedback conditions were significantly different from the fs condition. [Continued on next page]

Figure 3.9: The FS condition was significantly better than all other conditions, and only formative feedback (Fs) was better than only summative feedback (fS). **(E)** A sample contact cue and the face associated with the cue seen by participants across all conditions.

	fs	Fs	fS	FS	Total %
head gestures	7	6	7	7	96.4%
arm gestures	6	5	6	7	85.7%
facial expressions	5	7	6	7	89.3%
lights on the hands	0	0	0	0	0%
nonverbal sounds	5	7	4	6	78.6%
speech	0	0	0	0	0%

Table 3.1: Number of participants who identified each listed type of action as being provided by the robot during the study, separated by feedback condition.

	fs	Fs	fS	FS	Total %
head gestures	3	6	6	4	67.9%
arm gestures	3	5	5	4	60.7%
facial expressions	5	7	6	4	78.6%
lights on the hands	3	3	2	0	28.6%
nonverbal sounds	3	6	3	3	53.6%
speech	7	6	5	6	85.7%

Table 3.2: Number of participants who wished for each listed type of action from the robot, separated by feedback condition.

found that there was a significant effect of cue type on this metric ($F(2,48) = 7.4, p = 0.0015, \eta_p^2 = 0.24$). Participants explored robot reactions much more for the location cue in comparison to the pose cue ($t(48) = 3.4, p = 0.0042, d = 0.91$) and the contact cue ($t(48) = 3.3, p = 0.0058, d = 0.88$). This finding matches well with the lower scores and lower comprehension that participants achieved in the location cue, as exploring how the robot reacts is a good way to improve one's comprehension and thus performance.

While analyzing these responses and the corresponding comments about trying different actions to explore robot behavior, we noticed that this question was interpreted in two distinct ways. Some participants assumed that this question was asking about actions they performed to elicit the right response from the robot. This was not our intended meaning, as we expected all participants to try to figure out the cues; this ambiguity probably reduced the power of this question. Others interpreted this question as we intended, asking about erroneous actions they intentionally performed to generate robot feedback after identifying the correct way to respond to a particular cue. Participants who answered positively with this interpretation reported to have enjoyed testing robot actions to see the formative feedback, several participants with formative feedback made

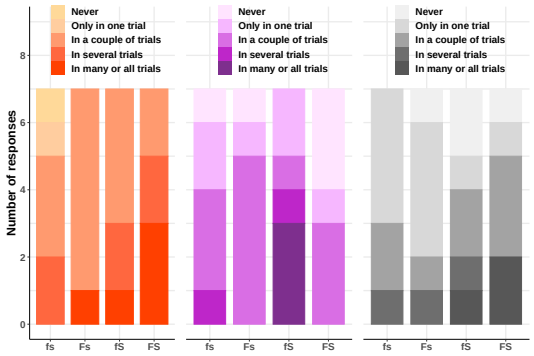


Figure 3.10: Participant self-reported frequency of exploring how the robot would react to different movements.

similar anecdotal comments to the experimenter.

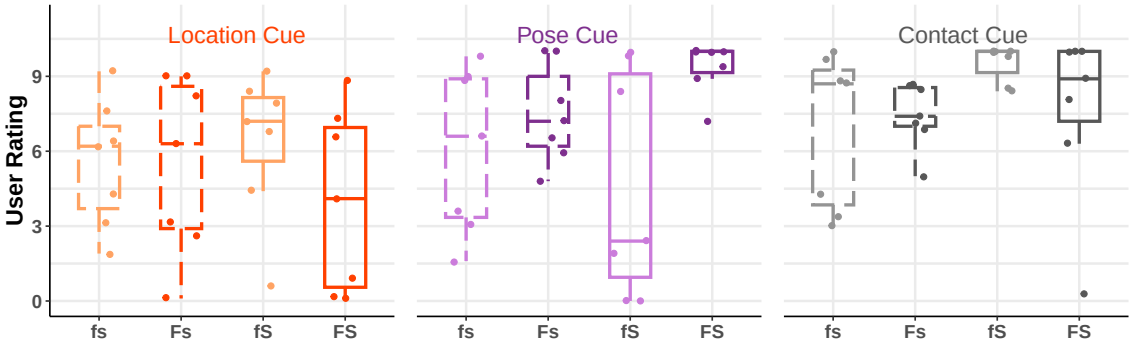


Figure 3.11: Participant confidence in their explanation of each of the robot's cue types.

3.3.7 Participant Survey Responses

We administered surveys to the participants as outlined in Section A.1.2. Here we present the responses to these surveys.

NASA-TLX Results

Fig. 3.12 shows the results of the NASA Task Load Index, with one data point for each participant's experience of each cue. As workload is a subjective indicator, high variations among participants are expected. Visual inspection hints that the location cue may have had higher workload than the other cues, and that summative feedback might have slightly increased workload for this cue alone. Visual inspection additionally shows that both kinds of feedback may have reduced participants' overall workload for the pose cue

and contact cue. However, a three-way mixed ART ANOVA showed that there was a significant three-way interaction between formative feedback, summative feedback and cue type on the total TLX results ($F(2,48) = 3.4, p = 0.04, \eta_p^2 = 0.13$). However, post-hoc analyses did not show any significant effects, likely due to the relatively low number of participants for this highly scattered between-subjects measure.

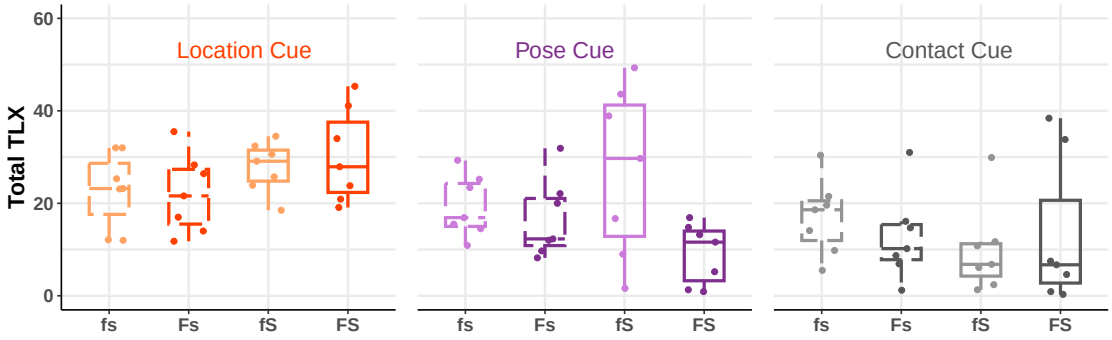


Figure 3.12: Self-reported total NASA-TLX workload results across the three cue types and the four feedback conditions (fs, Fs, fS, and FS). Table A.3 shows the six questions that constitute this survey; higher values indicate greater workload. Though there was a significant three-way interaction, post-hoc results were not significant.

SUS Results

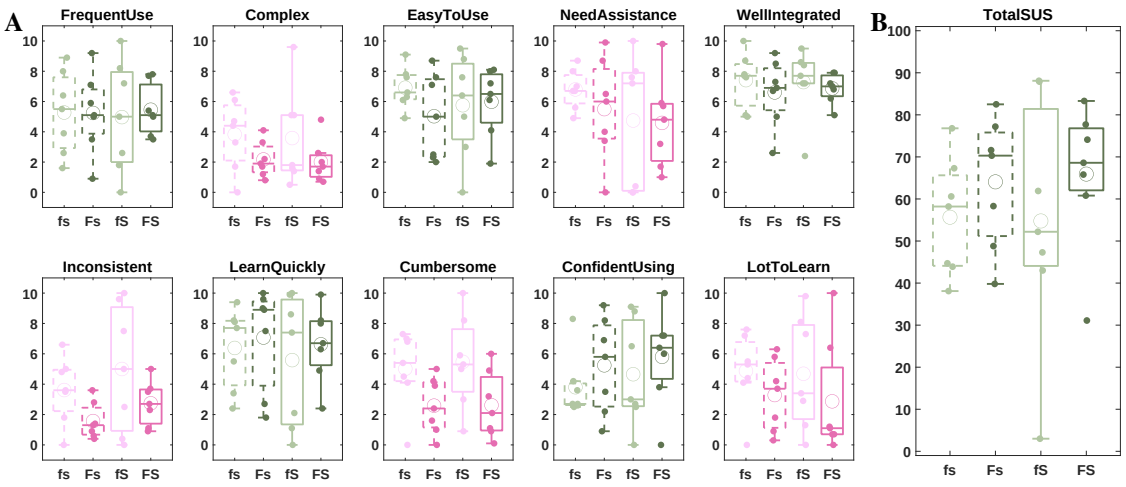


Figure 3.13: System Usability Scale results across the four feedback conditions (fs, Fs, fS, and FS). **(A)** Plots of the responses to the ten individual questions that make up this survey (Table A.2). Positive questions are colored green, and negative questions are colored pink. **(B)** Total score. Responses to negative questions were inverted before summing with the positive responses to yield the total score shown.

Fig. 3.13 presents the results of the System Usability Scale, which we analyzed by question and overall. Again, variability between participants is high for this subjective measure. Overall, a two-way ART ANOVA revealed that feedback condition had a significant impact on only one item. There was a main effect of formative feedback on how cumbersome the usability of the system was ($F(1,24) = 7.81, p = 0.01, \eta_p^2 = 0.25$). Participants found that formative feedback made the system significantly less cumbersome to use ($t(24) = 2.79, p = 0.01, d = 1.06$). There was also a statistically trending effect of formative feedback on how inconsistent users from the system to be ($F(1,24) = 3.89, p = 0.06, \eta_p^2 = 0.14$). As seen in Fig. 3.13A, formative feedback may have made the system seem more consistent. Statistical analysis indicated that feedback condition did not significant affect the total SUS ratings. However, visual inspection shows that formative feedback may tend to elicit higher system usability scores (mean $\pm$ standard deviation for fs: 55.7 ± 14.0, Fs: 64.1 ± 15.6, and FS: 65.9 ± 17.1), likely stemming from the two questions (cumbersome, inconsistent) with significant or trending benefits of formative feedback. Summative feedback does not appear to have affected this measure (fS: 54.8 ± 29.3). In retrospect, the System Usability Scale was not an ideal qualitative measure to use in this study, since it did not focus on the feedback itself.

RAS Results

Figs. 3.14 and 3.15 show the results of the Robot Acceptance Survey, which was administered both before and after the study. Overall, participants had a more negative attitude towards robots after the study. This view could be a result of the frustration they felt due to the lack of instructions. The RAS results were split into eight categories as recommended by Weiss et al. (Weiss *et al.*, 2008). Fig. 3.14 shows the results of four of these categories (performance expectancy, effort expectancy, attitude towards technology and self-efficacy) which were adapted from the UTAUT model (Venkatesh *et al.*, 2003). Fig. 3.15 depicts the other four categories (forms of grouping, attachment, reciprocity, cultural context), which were derived from the concept of object-centered sociality (Cetina, 1997). Each component of the RAS was analyzed using a three-way ART ANOVA that used the two feedback types (without and with formative feedback, without and with summative feedback) and study participation (before and after the study) as the three independent factors.

Fig. 3.14A shows the three RAS items that measure participants' expectations of robot performance. The study had an impact on how users felt about the usefulness ($F(1,24) = 12, p = 0.0018, \eta_p^2 = 0.34$), helpfulness ($F(1,24) = 8, p = 0.0094, \eta_p^2 = 0.25$) and supportiveness of the robot ($F(1,24) = 7.3, p = 0.012, \eta_p^2 = 0.23$). Fig. 3.14A shows that participants felt that robots are less useful ($t(24) = 3.5, p = 0.0018, d = 0.94$), less helpful ($t(24) = 2.8, p = 0.0094, d = 0.75$), and less supportive ($t(24) = 2.7, p = 0.012, d = 0.72$) after the study.

Fig. 3.14C shows participants' expectations of the effort required to use the robot. Although there was no effect of feedback or study participation on the perceived ease

of use of the robot (Fig. 3.14C), the study significantly affected how users felt about cooperating with the robot ($F(1,24) = 4.3, p = 0.05, \eta_p^2 = 0.15$). They felt that they could cooperate less with the robot after the study ($t(24) = 2.1, p = 0.05, d = 0.55$). This response could be attributed to the lack of instructions and participants' reduced feelings of self-efficacy.

Participants' attitudes toward technology were measured via three questions (Fig. 3.14B). The study had a significant impact on whether they thought that using the robot is a good idea ($F(1,24) = 14, p = 0.00099, \eta_p^2 = 0.37$). After the study, they had a lower opinion about using the robot ($t(24) = 3.7, p = 0.00099, d = 1$). There was a significant interaction between formative feedback and participation in the study on how afraid they were to make mistakes ($F(1,24) = 4.3, p = 0.049, \eta_p^2 = 0.15$); visual inspection of Fig. 3.14B indicates that participants who received formative feedback (Fs, FS) may have been less afraid of making mistakes after the study, but we did not find any significant simple effects upon further analyzing this two-way interaction. The study also had a significant impact on the user's feeling that they may break the robot ($F(1,24) = 9, p = 0.0061, \eta_p^2 = 0.27$); they felt that they were substantially less likely to break it after the study ($t(24) = 3, p = 0.0061, d = 0.8$).

The final category of the UTAUT-based questions measured self-efficacy (Fig. 3.14D). Perceived self-inefficacy increases arousal and anxiety and makes people feel they cannot perform (Bandura, 1986). On the other hand, a positive sense of self-efficacy reduces fear and increases the person's ability to perform 'threatening' tasks (Bandura, 1982) and is associated with acceptance for all robots (Latikka *et al.*, 2019). There was no impact of either feedback or the study on feelings of comfort in the presence of the robot. The presence of summative feedback had a significant impact on how users felt about whether they could use the robot with external help ($F(1,24) = 5.1, p = 0.033, \eta_p^2 = 0.18$). The presence of summative feedback made users feel that they could work with the robot if help was provided ($t(24) = 2.3, p = 0.033, d = 0.66$). There was a statistically significant two-way interaction between formative feedback and the study on being able to work with the robot without external help ($F(1,24) = 4.3, p = 0.049, \eta_p^2 = 0.15$). Though we did not find any significant simple effects upon further analyzing the two-way interaction, visual inspection appears to indicate that the presence of formative feedback made participants feel like they could use the robot without help. There was a significant three-way interaction between formative feedback, summative feedback, and participating in the study on being able to use this system with initial training ($F(1,24) = 7.5, p = 0.012, \eta_p^2 = 0.24$). Post-hoc analyses revealed that there was a two-way interaction between both types of feedback after the study ($F(1,24) = 10, p = 0.0079, \eta_p^2 = 0.3$). Additionally, the presence of summative feedback had a significant impact when formative feedback was absent ($F(1,12) = 7.6, p = 0.035, \eta_p^2 = 0.39$) but not when formative feedback was present. In this case, participants thought it would harder to work with the robot when training was provided when summative feedback was absent ($t(12) = 2.8, p = 0.017, d = 1.5$). Overall, participants felt they would re-

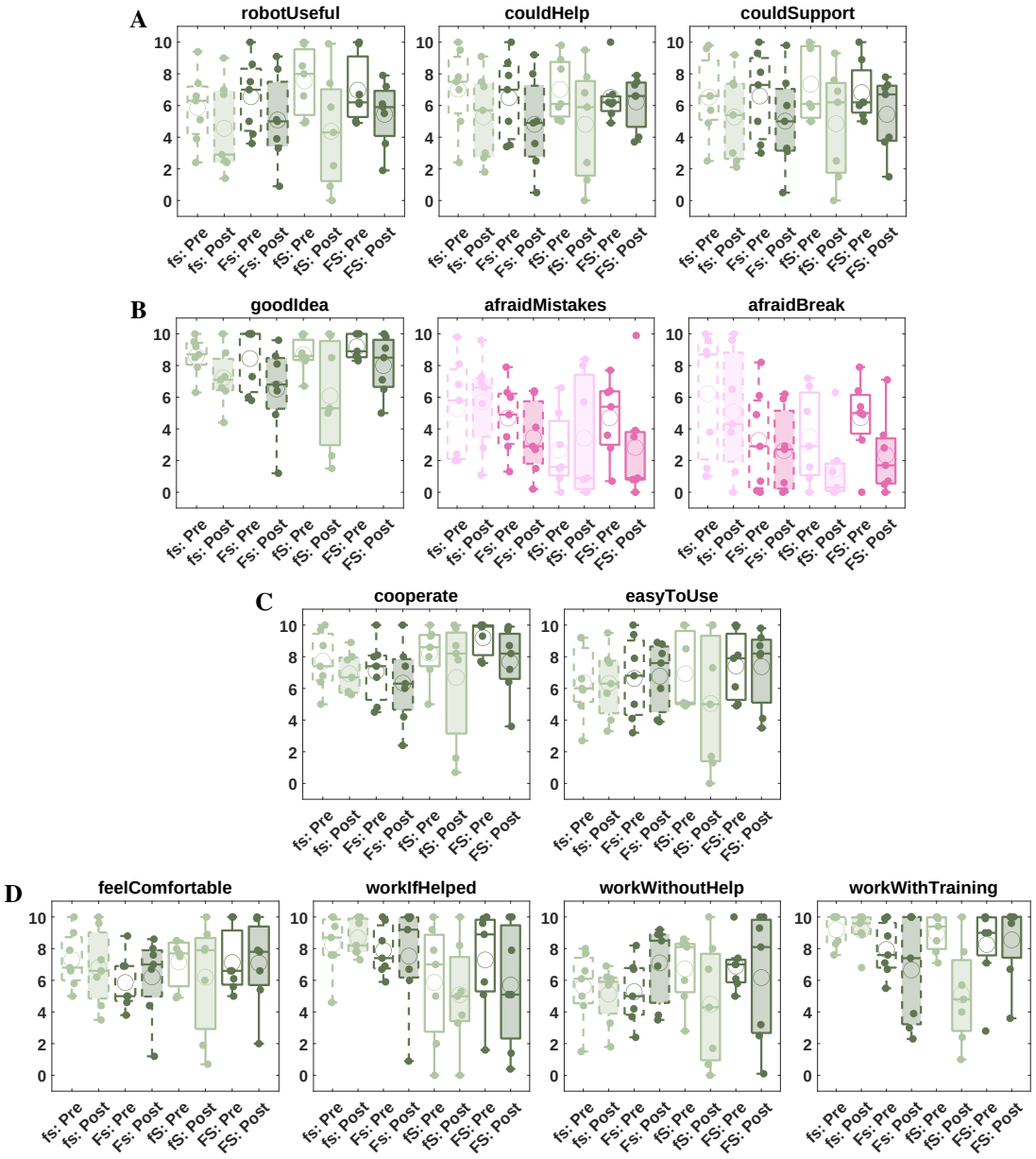


Figure 3.14: Robot Acceptance Survey categories based on the UTAUT model across the four feedback conditions (fs, Fs, fS, and FS). Positive questions are colored green, and negative questions are colored pink. **(A)** Participant responses regarding performance expectancy of robot. They felt that the robot was less useful, less helpful, and less supportive after the study. **(B)** Participant responses to questions assessing their attitudes toward technology. Overall, after the study they were less confident that using that robot was a good idea and were less afraid to break the robot after the study. **(C)** Participant responses about effort expectancy. They felt less able to cooperate with the robot after the study. However, the study had no impact on ease of use. [Continued on next page.]

Figure 3.14: **(D)** Participant responses to self-efficacy measures. Summative feedback made participants feel like they could work with the robot if external help was provided. There is also some indication that the presence of formative feedback made them think they could work with the robot without help.

quire some form of assistance or initial training to be able to use the system. This observation can be attributed to the fact that instructions were not provided to the participants to allow the effects of the autonomous robot's feedback to be measured.

Fig. 3.15 looks at user attributions of social and cultural constructs toward robots. There were no effects of the study or feedback conditions on participants' desire to build groups with and around the robot (Fig. 3.15A), their feelings of attachment toward robots (Fig. 3.15B), or their feelings of reciprocity (Fig. 3.15C). Finally, Fig. 3.15D looks at participant robot perceptions from a cultural context. The study had a negative effect on participant perceptions of impressing others with robot ownership ($F(1,24) = 6.1, p = 0.021, \eta_p^2 = 0.2$); users felt it would be less impressive to own a robot after the study ($t(24) = 2.5, p = 0.021, d = 0.66$). The study also had an impact on fear of losing one's job because of a robot ($F(1,24) = 8.5, p = 0.0076, \eta_p^2 = 0.26$); participating in the study made them less afraid of losing their jobs due to a robot ($t(24) = 2.9, p = 0.0076, d = 0.78$). However, their feelings about how nice robots are did not change after the study.

3.4 Discussion

For robots to interact effectively outside laboratory environments, they must be able to communicate with and engage users with minimal supervision from human experts. The presented study shows how an autonomous exercise robot can use multimodal nonverbal cues to convey task instructions, correct user behavior in real time, and communicate overall assessments of performance. Since education researchers have repeatedly proven the merits of both formative and summative assessments (McMillan, 2013; Schildkamp *et al.*, 2020), we tested these two kinds of feedback in a minimally supervised setting where users had to try to figure out three types of task cues based only on the actions of the robot. The results resoundingly support our hypotheses, showing that roboticists can use both of these time-tested feedback mechanisms to develop responsive interaction strategies that benefit users, while also opening several interesting discussion points for future work.

Role of Formative Feedback

Formative assessments are ongoing evaluations that support learners in improving their performance and increasing their understanding of the material (McMillan, 2013; Hattie and Timperley, 2007; Connors, 2021). As predicted by H1 and H2, the nonverbal formative actions our robot used to correct user errors in real time helped in both of these

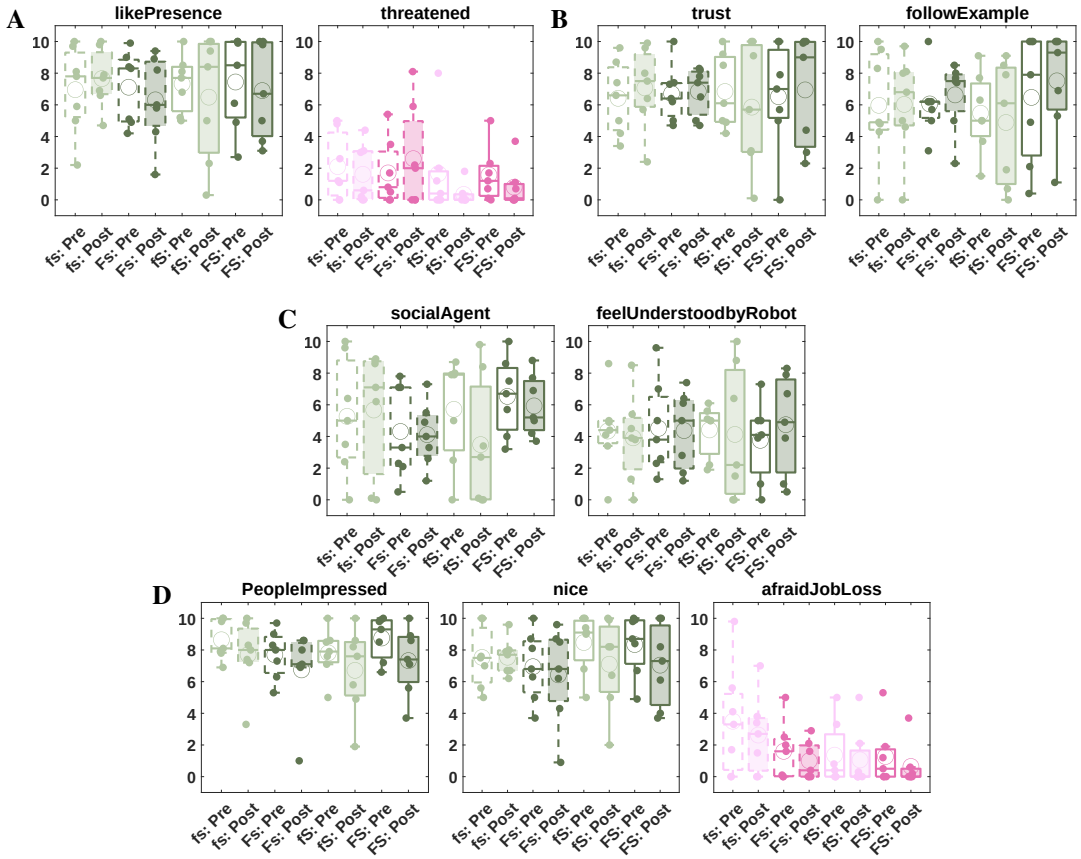


Figure 3.15: Robot Acceptance Survey categories based on the concept of object-centered sociality across the four feedback conditions (fs, Fs, fS, and FS). Positive questions are colored green, and negative questions are colored pink. **(A)** Participant responses to forms of grouping category items. There are no effects of the feedback types or the study on either liking the presence of the robot or feeling threatened. **(B)** Participant responses regarding attachment toward robots. There are no effects of the study or feedback on feelings of trust or the user's desire to follow the example of the robot. **(C)** Feelings of reciprocity reported by participants. There are no effects of the study or feedback on feelings of social agency of the robot or being understood by the robot. **(D)** Participant responses to cultural perception of robots. After finishing the study, participants felt that others would be less impressed if they owned a robot. They also were less afraid that they might lose their job because of a robot.

ways: formative feedback led to significantly higher average trial scores (Fig. 3.4B, H1) and significantly better post-study explanations of the tasks (Fig. 3.5B, H2). Formative assessments have been used within the classroom to improve both teaching strategies and student learning (Dixson and Worrell, 2016). Our results suggest that robotic coaches can similarly leverage the powerful paradigm of formative assessment; providing formative feedback that responds to what the user is currently doing provides highly useful granular information about correct and incorrect actions. Achieving the performance and understanding benefits of formative feedback requires a robotic system that measures user activity in real time and rapidly outputs corrective feedback when errors are detected. Effective formative feedback lies at the core of providing personalized support, facilitating skill development, and promoting user engagement.

Role of Summative Feedback

Summative feedback, on the other hand, conveys the final evaluation of the task. Summative assessments have repeatedly been shown to aid learner performance (Ismail *et al.*, 2022; Connors, 2021), but they are not thought to aid comprehension. Indeed, as predicted by H1, the summative feedback provided by our robot significantly increased average trial scores (Fig. 3.4C); it had no impact on the participants' ability to describe the tasks (Fig. 3.5C). These findings match well with existing education literature that states that summative feedback communicates the current state of the user's knowledge, whereas formative feedback supports the process of acquiring knowledge and skills over time (Dolin and Evans, 2017). Seeing the robot's red, yellow, or green face at the end of each trial helped participants with summative feedback achieve higher task scores, but it did not improve their ability to articulate what they were supposed to do. Formative feedback, on the other hand, supported participants' learning process by showing them whether what they were doing was correct at each instant in time during the trials. Nevertheless, robots can still use summative feedback to convey concrete performance metrics to help users improve. Advantageously, summative feedback has lower technical requirements than formative feedback, as it presents one summarized score at the end of an activity, with no real-time reactions. Robots can also combine formative and summative feedback, as our robot did in the FS condition, since the benefits of these two types of feedback are largely independent.

Cue Design

Since we wanted to study the impact of formative and summative feedback, we purposefully designed the tasks and robot cues to present a challenge for users so that they would undergo a learning process that depends on their assigned feedback condition. In this vein, we intentionally chose simple, static tasks to focus on the feedback. Our three resulting task cues can be categorized in terms of ease of understanding (Fig. 3.5D) and ease of performance (Fig. 3.4D). The location cue was more difficult to understand than

the other cues; since it was rarely understood, it was also rarely performed correctly, even though standing in a particular location is an easy activity. Most users (82.14%) correctly understood that they should copy the robot's pose in the pose cue, though the selected error tolerance was difficult for some individuals to maintain for 30 s. The contact cues were easy to understand and easy to perform. Most participants (78.57%) figured out that they needed to touch the robot's hand and could easily execute this action.

Summative feedback was consistently conveyed across all tasks, and participants seemed to understand its meaning through the redundant face colors, facial expressions, and sounds. The influence of summative feedback might have been greater if it had been designed to present the score with more granularity. The facial expressions that the robot used to provide formative feedback were also clearly understood and valued by participants; since they were more complex and unusual, the formative gestures often took several trials for participants to decipher. Simultaneously, it is useful to note that the formative gestures the robot executed for the three task cues provided different amounts of information: error magnitude and direction for location, error magnitude for pose, and quantized error magnitude for contact, roughly matching task difficulty. Though formative feedback provided less information for the pose and especially the contact cues, participants were still successful at identifying the appropriate solutions, as these cues were easier than the location cue.

The heatmaps created from the motion-capture data offer a dense overview of user activity during each task. The values of the corresponding Wasserstein metrics reiterate the relative challenge of the three cue types, with the location cue (Fig. 3.7D) being more difficult than the pose (Fig. 3.8D) and contact (Fig. 3.9D) cues. Though visible by eye (Fig. 3.7A), the distribution shifts caused by feedback in the location cue did not reach statistical significance. This finding may mean that the benefits of feedback are limited when a task is difficult to understand, or that our implementation of formative feedback for this cue was not sufficiently clear to participants. Formative feedback was particularly useful for inducing users to move their hands closer to the target locations in the pose cue (Figs. 3.8B and 3.8D), whereas both types of feedback independently helped users solve the contact cue (Figs. 3.9C and 3.9D). These observations could indicate that formative feedback may help with a wider range of tasks than summative feedback, likely due to its support of task understanding.

Formative and summative feedback should be synergistic in nature and should feed into one another (Mogboh and Okoye, 2019; Lau, 2016). Given our results, we believe that minimally supervised robots, especially comprehensive exercise coach robots, will be more effective when they provide both types of feedback. Formative feedback plays an essential role in improving the user's task comprehension, whereas summative feedback shows the performance level they have reached. Thus, the two types of feedback play different roles in facilitating user performance and should both be provided. Furthermore, drawing from the theory of dual-coding within education (Clark and Paivio, 1991), which states that information is more likely to be remembered when it is presented with both verbal and nonverbal symbolism, we recommend combining gesture-based feed-

back with other interaction modalities, such as speech. This idea is corroborated by our users, who recognized the different aspects of the feedback provided by our robot (head and arm gestures, facial expressions, and sounds) and wished it would also speak (Tables 3.1 and 3.2).

From Assessments to Feedback

Formative and summative assessments are two sides of the same coin (Dixson and Worrell, 2016); formative assessment fosters learning while summative assessment validates it (Stiggins and DuFour, 2009; Connors, 2021). They complement each other, as seen from the observation that each type of feedback independently had an impact on the score of the user (Figs. 3.4B and 3.4C). Thus, our findings strongly support H1. H2 was also strongly supported: only formative feedback enhanced user understanding of the task (Figs. 3.5B and 3.5C). These observations are fully consistent with the education literature, which states that feedback in general has been shown to improve performance (McMillan, 2013) and formative feedback has been shown to promote understanding of the task (McMillan, 2013; Andrade and Cizek, 2010).

Summative feedback is currently the most common type of feedback provided by social exercise robots (Guneysu and Arnrich, 2017; Shao *et al.*, 2019; Fitter *et al.*, 2020; Lewis *et al.*, 2016; Görer *et al.*, 2013). These robots utilize exergaming as their primary motivational strategy; such games provide summative performance feedback to try to cause behavior changes (Lyons and Hatkevich, 2013). Though it is more difficult to implement, a key role of formative assessments is to involve the individual in the learning process, enabling them to obtain personalized support (Connors, 2021; McMillan, 2013). The same sensors used to determine summative feedback can be leveraged by robots to provide formative feedback. When designing formative feedback, a one-size-fits-all approach is often not suitable, as seen from the custom formative feedback we designed for each cue type. Formative feedback is about using acquired data to influence the individual's learning and can lead to improved learning of the task at hand if implemented correctly (Schildkamp *et al.*, 2020). The feedback must be customized to the task at hand and must consider the user's performance over the course of the task. Similarly, a robot could use adaptive formative feedback by changing its feedback strategies based on user performance (Frost, 2016). This ideology could inform future formative feedback designs through adaptive feedback that changes over multiple interactions with the same participant. Ideally, a robot should adjust its strategy if a particular type of formative feedback is not effective at eliciting a desired change in the user's behavior.

Furthermore, formative feedback has been shown to have a positive impact on academic motivation and engagement (Ismail *et al.*, 2022; Chen *et al.*, 2021), test anxiety (Ismail *et al.*, 2022), and attitude towards learning (Ismail *et al.*, 2022; Babinčáková *et al.*, 2020). From this observation, one can extrapolate that incorporating formative feedback into robot behavior has the potential to help robots gain long-term acceptance. An interesting thread of future research would be to test this perspective in the context

of robotics.

Strengths and Weaknesses of Our Approach

One of the major strengths of this study is the quantitative information that we gathered about user interactions with our robot. Capturing high-frame-rate markerless motion-capture data enabled us to create heatmaps of user activity to depict what happened during each experiment session. As motion capture data is often extensive and hard to comprehend, we believe that heatmaps represent a practical alternative for exercise therapy experts seeking to observe and evaluate user performance. An interesting line of future research would be to validate the utility of these heatmaps for use by exercise therapy experts. Though this study focuses on group-wise heatmaps, these techniques could be extended to generate one heatmap per participant to observe individual task performance. Furthermore, the Wasserstein distance metric could be used to evaluate user performance over multiple sessions or days of exercise. The technical design of the Robot Interaction Studio (Mohan *et al.*, 2021b) enables us to provide consistent robot responses and interactions across all participants. Thus, the timing, type, and consistency of feedback were not affected by experimenter error or bias, as can happen with teleoperation. Finally, the use of educational assessments in such a context has not previously been explored in the realm of HRI. Though we examined them from the perspective of gesture-based interaction, these methods could generalize to any type of feedback mechanism within HRI.

When analyzing the results, we found that our participants could roughly be classified into two categories: the explorers and the imitators (Fig. 3.10). The explorers performed a variety of actions, walking around the workspace, touching the robot, and actively trying to figure out the solution to each cue type. Conversely, the imitators chose to stay in a single position and mimic the robot's arm pose across all 75 trials, irrespective of any feedback provided. This imitation behavior could be linked to motor resonance (Landmann *et al.*, 2011), which has been observed during human-robot interactions in the past (Bisio *et al.*, 2014), but it was the correct behavior for only one of the three cues tested in this study. To encourage them to perform diverse tasks, the experimenter used standardized language to remind all participants that they were allowed to move and make contact with the robot during each study break and when they asked for clarifications; however, some participants still persisted in imitating the robot throughout. Thus, when conducting studies without structured instructions, we suggest administering a questionnaire such as the Curiosity and Exploration Inventory (Kashdan *et al.*, 2009) to determine how well this personality attribute correlates with participant strategy.

Considering this diversity in participant responses to the robot actions, a larger participant pool would have been beneficial to assess the nuances of how each type of feedback affects participant engagement and self-learning. It would also be meaningful to tune the difficulty of each task by adjusting the error threshold to match the user's mental and

physical capabilities, as done by Fitter et al. (Fitter *et al.*, 2020) for older adult users. The error threshold for each task could also change over time, starting large to make the task easy and reducing with subsequent success to elicit more precise performance of the desired action.

Participants often glanced at the experimenter or asked for further clarification; experimenters in typical studies help participants in such situations, but our experimenter provided no additional support other than repeating a shortened version of the instructions. Though frustrating, this policy was necessary to create a consistent experience of minimal supervision for all participants and allow us to investigate the effects of formative and summative feedback provided by a robot. Interestingly, our survey responses show that feedback increased the participants' feelings of self-efficacy (Fig. 3.14D). Future research could investigate the impact of feedback on self-efficacy in more detail.

3.5 Conclusion

Exercise coach robots have the potential to improve compliance with exercise regimens. In an ideal scenario, such robots would guide users through exercise routines and provide effective feedback entirely autonomously without any supervision. We believe such robots would greatly benefit from providing compelling nonverbal feedback to their users, especially gestures, facial expressions, and sounds. However, effectively monitoring the user motion and adapting the robot's behavior accordingly continues to be a technical challenge (Matarić and Scassellati, 2016) due to the speed and accuracy requirements for both perception and action. Furthermore, designing effective feedback response systems lies at the core of personalizing robotic systems to their users. As substantiated by the results of our study, feedback design based on formative and summative feedback holds great potential for future research in this area.

Chapter 4

Expert Perception of Exercise Robots

4.1 Introduction

Now that we have established an environment for minimally supervised interactions with an exercise robot and have implemented associated feedback strategies, the subsequent phase involves determining an efficient approach to create specific behaviors for the robot. As mentioned in the introductory section, we consider teleoperation as a promising method to achieve this objective. Simple movements and exercises have been successfully imparted to robots using teleoperation (Tanguy *et al.*, 2016; Görer *et al.*, 2013). Nonetheless, to the best of our knowledge, no research has been conducted so far to gather professional perspectives on teleoperation within the realm of exercise-related fields. Furthermore, research shows that stakeholder involvement can enable the creation of technology that meets their needs and expectations (Fraune *et al.*, 2022).

Šabanović (2010) posits a bidirectional relationship between society and technology, emphasizing the imperative of involving stakeholders in the robot design process to foster the development of socially robust technologies. Thus, before delving deeply into the development of a teleoperation system, we sought to engage with expert stakeholders to ascertain their interest in such a system and, if so, gather insights on the features they would find valuable. We also took this opportunity to better understand the challenges that physical therapy experts face and the role that social exercise robots could play as tools that could support them. Thus, we sought to answer the following research questions:

[RQ 1] How is a typical therapy session facilitated?

[RQ 2] What are the challenges that experts face?

[RQ 3] What are experts' perceptions of a social exercise robot?

[RQ 4] What features do they desire for such a robot?

[RQ 5] Is teleoperation a capability that interests experts?

[RQ 6] What teleoperation features do they desire with social exercise robots?

Since, this study is currently ongoing, the scope of this thesis includes only a preliminary analysis of **RQ 4**, **RQ 5** and **RQ 6**. Thus, this chapter briefly delves into expert opinions on social exercise robots and teleoperation. The rest of the chapter is organized as follows: Section 4.2 describes our study design in detail. The results are presented in Section 4.3, followed by the discussion and conclusion in Sections 4.4 and 4.5, respectively.

4.2 Methods

To gain insight into expert perceptions of exercise robots and teleoperation, this user study was designed as a three-phase semi-structured interview process. The study was approved by the Ethics Council of the Max Planck Society under the Haptic Intelligence Department's framework agreement as protocols F020A, F020B and F020C. These dialogues were conducted virtually, employing teleconferencing platforms such as Zoom or WebEx. All experts were compensated for their time using gift cards to an online store of their choice, such as Amazon. The methodological framework of the study adhered to the guidelines put forth by Adams (2015), drawing inspiration from the user study orchestrated by Winkle *et al.* (2018).

In this investigation, the social exercise robot employed was the Rethink Robotics Baxter. A comprehensive timeline of the study is depicted in Fig 4.1, and a detailed breakdown can be referred to in Table 4.1. The study began with succinct introductions that segued into the three distinctive phases delineated in the forthcoming subsections. Each session involved one participant and two researchers performing the roles of interviewers. Supplementing these interviews, videos and surveys were utilized to garner more quantitative metrics. LimeSurvey was the chosen medium for all survey administration. The interviewers meticulously noted down detailed observations. All conversations were duly recorded, contingent on explicit participant consent, with transcripts set to undergo an inductive thematic analysis. However, the scope of this thesis is exclusively oriented towards the quantitative data aggregated via the surveys.

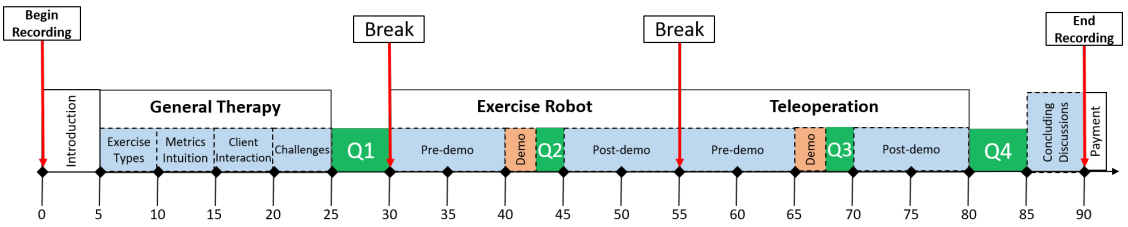


Figure 4.1: A visual representation of the experimental progression, highlighting the three primary phases, the approximate time allocation for each stage, and the administered questionnaires.

Section	4.2.1 General Therapy	A.2.1	4.2.2 Exercise Robots	4.2.3 Teleoperation	A.2.1, A.2.1	4.2.4
	I: Types of exercise	Q: Baseline Questionnaire	I: Pre-demo interview	I: Pre-demo interview	Q: Post-study, Demographic Questionnaire	I: Concluding Discussions
	I: Metrics / Intuition		D: Exergame with Baxter	D: Teleoperation Demo		
	I: Interaction with clients		Q: Robot Questionnaire	Q: Teleoperation Questionnaire		
	I: Challenges		I: Post-demo interview	I: Post-demo interview		

Table 4.1: Details of the different phases of the study presented in Fig. 4.1. In this table, **I** refers to an interview, **Q** refers to questionnaires or surveys, and **D** refers to video demos shown to the participant. The semi-structured interview commenced with an overview of the expert's job responsibilities, followed by an introductory survey. In the next phase, the discussion delved into the subject of exercise robots, and then transitioned to teleoperation. Finally, the participant completed both a closing survey and a demographics survey before debriefing with the experimenters. The detailed interview structure is provided in Appendix Section A.2.1

4.2.1 Phase 1: General Therapy

The primary objective of this phase was to gain a comprehensive understanding of participants' job responsibilities and build rapport with the participants. We also inquired about the nature of the key performance metrics they utilized, their social interactions with clients, and the challenges they encountered in their roles. The questions posed to the participants can be seen in Appendix Table A.6. At the end of this section, participants completed a slightly modified version of the Robot Acceptance Survey (Appendix Section A.2.1) based on the survey designed by Weiss *et al.* (2008). The purpose of this survey was to obtain an understanding of participants' initial perspectives on robots. Participants were given the option to take a break after the completion of this survey.

4.2.2 Phase 2: Exercise Robots

The primary objective of this phase was to familiarize participants with the concept of social exercise robots and serve as a stepping stone to introduce participants to the possibilities and benefits of utilizing social robots as exercise partners or guides. To begin, we initiated the process by posing initial questions to the participants to gauge their perceptions and understanding of a such a robot. Following the initial inquiry, we proceeded to showcase four sample exergame videos featuring social exercise robots. The intention was to provide practical demonstrations of how these robots can be integrated into exercise routines and promote social interaction. These included three videos of exergames from (Fitter *et al.*, 2020) and a video of the location cue with both formative and summative feedback from the study reported in Chapter 3.

After viewing the videos, participants were asked to complete a survey encompassing a set of self-reported metrics related to their perceptions of exercise robots, inspired by (Weiss *et al.*, 2008). This survey is detailed in Appendix Section A.2.1. Additionally, they were also prompted to express their preferences for specific features they would like to see in such a robot. Furthermore, the survey included several open-ended questions to gather more comprehensive and general feedback from the participants. Following the survey, this phase culminated with a short session when participants were asked specific questions to delve deeper into their opinions on exercise robots. The questions posed to the participants during this interview phase can be seen in Appendix Table A.7. Participants were given the option of taking a break at this stage if they so desired.

4.2.3 Phase 3: Teleoperation

The third phase of the investigation was designed to capture the participant's viewpoint on teleoperation within the sphere of socially interactive exercise robots. We provided a succinct explanation of the concept of teleoperation in this context and then sought their initial impressions of the same. Subsequent to this, they were shown three video demonstrations of robot arm teleoperation, accompanied by one showcasing head and facial teleoperation. Participants thereafter undertook the completion of a survey, in which they self-reported measures regarding their perspectives on teleoperation (refer to Appendix, Section A.2.1). The survey comprised inquiries into the features they desire in a teleoperation system, as well as open-ended questions to glean additional insights on the subject.

Finally, participants continued the dialogue with the interviewers, affording them the opportunity to provide more elaborate feedback. In these discussions, they were encouraged to deliberate on the advantages and disadvantages of teleoperation, propose potential applications for the teleoperation system, suggest novel behaviors they would like to facilitate, and, finally, to share their receptiveness towards employing it as a recording device to create exergames to clients in a more generalized manner. The questions posed to the participants during this interview phase can be seen in Appendix Table A.8.

4.2.4 Final Debrief

Upon the conclusion of the three phases, the participants completed the same Robot Acceptance Survey that they completed at the beginning of the task (Appendix Section A.2.1). The experiment then transitioned into a debriefing stage. At this juncture, participants were presented with the option to offer any additional feedback, express concerns, or pose any lingering questions they might have. Concurrently, the researchers utilized this opportunity to elucidate further our research objectives pertaining to this project, and they thanked the experts for their generous contribution of time and insights.

4.2.5 Statistical Analysis

Due to the small number of participants ($N = 8$) in this within-subjects study, we conducted a non-parametric Wilcoxon signed rank test on the paired samples of the items on the Robot Acceptance Survey administered before and after the study. We also report the effect sizes using Cohen's d Lakens (2013). The significance-testing level, α, was considered to be 0.05 in all cases.

4.2.6 Participants

We defined an expert as any individual engaged in exercise-centric vocations entailing client interactions. These professions spanned orthopedic practitioners, physiatrists, physiotherapists, occupational therapists, researchers specializing in physical therapy, exercise coaches, and trainers affiliated with gyms and fitness centers. In our endeavor to obtain a more comprehensive and global perspective, we recruited participants on a worldwide scale. Presently, we have conducted a total of eight interviews ($N = 8$) with participants originating from three distinct countries. Detailed demographics of the participants can be accessed in Table 4.2.

The range of expertise among our participants, in relation to their interaction with robots, varied considerably. Among them, three individuals ($N = 3$) were classified as absolute novices, devoid of any prior robotic interaction. Conversely, one participant ($N = 1$) was classified as a beginner, having engaged with commercial robot toys. A further four participants ($N = 4$) had experienced interactions with more sophisticated robots, such as Pepper and NAO. Notably, one participant had had the opportunity to engage with a Baxter robot ($N = 1$), whilst another reporting having read about it ($N = 1$). For the remaining six participants ($N = 6$), the study served as their initial introduction to the Baxter robot.

4.3 Results

In order to examine expert sentiment, we evaluated the self-reported metrics provided within their survey responses. This analysis initially investigates their perspectives on exercise robots, followed by their views on teleoperation, and finally delves into any potential alterations in their perceptions of robots as a consequence of the study.

4.3.1 Expert Views on Social Exercise Robots

Our exploration begins with an analysis of expert views on social exercise robots, based on surveys completed during the second phase of the study. Fig. 4.2 exhibits a visual summary of expert feedback regarding exercise robots. The expert attitudes towards social exercise robots demonstrate considerable diversity. The breadth of opinions on the

ID	Profession	Practicing Country	Client Age (yrs)	Client Concerns
$P1$	Physiatrist	India	0–15	Spinal cord injury (SCI)
$P2$	Physiatrist	India	0–100	Stroke, SCI
$P3$	Professor of Physiotherapy	India	35–75	Stroke, traumatic brain injury (TBI), post-fracture complications, nerve injuries, SCI
$P4$	Physiatrist	India	0–100	Movement dysfunction
$E1$	Exercise Coach	Great Britain	9–95	Scoliosis or postural imbalances / discomfort
$OT1$	Professor of Occupational Therapy	USA	30–70	Stroke, SCI, TBI, orthopedic conditions, concerns related to aging
$P5$	Physiatrist	India	20–50	Stroke, TBI, SCI, cerebral palsy, amputations
$RS1$	Rehabilitation Research Scientist	USA	most 0–7, up to 21	cerebral palsy, stroke, neonatal followup

Table 4.2: Participant demographics.

merit of the concept, as well as the willingness to recommend these robots is significant ($M = 5.2$; $SD = 2.82$). There is a similar diversity in readiness to use these robots with their clients ($M = 5.7$; $SD = 2.36$). With regard to the idea of using the exercise robot, opinions display an expansive range ($M = 6$; $SD = 3.62$), as does their level of worry about it ($M = 3.7$; $SD = 3.45$). The variety in views also extends to sentiments about the robot's presence ($M = 6.8$; $SD = 3.34$), its social agency ($M = 6.2$; $SD = 3.66$), and its potential utility in facilitating exercise ($M = 7.2$; $SD = 3.31$). However, the variation in opinions seems less when considering the possibility of cooperating with the robot ($M = 7.1$; $SD = 2.91$), its agreeableness to work with ($M = 7.2$; $SD = 2.56$), and the interest in learning more about it ($M = 7.4$; $SD = 2.04$).

The study yielded a rich array of open-ended responses, offering deeper insights into individual expert perspectives. For instance, $P2$ appreciated the physical presence of the robot, while $P3$ believed it could help reduce fatigue and boredom. Experts $E1$ and $OT1$ were charmed by the robot's facial expressions. $OT1$ commended the robot's capability to facilitate lower limb exercises, even in the absence of legs. Additionally, $OT1$ and $RS1$ found value in the robot's ability to merge physical and cognitive training. $P5$ saw the potential for the robot's use in home-based therapy, particularly in enhancing hand coordination.

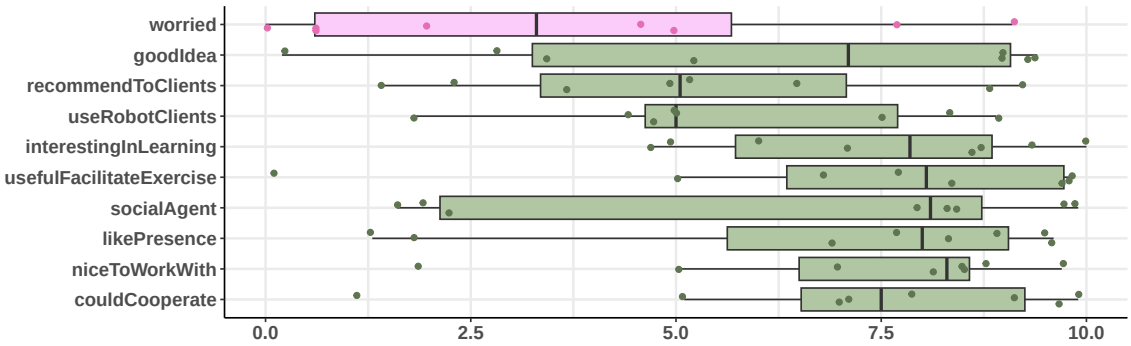


Figure 4.2: Expert feedback about exercise robots. Positive questions are colored green, and negative questions are colored pink. Expert responses are diverse but generally trending positive.

Exercise Robot Features

Further, we asked the experts about the desired features they envision in the robots (refer to Fig 4.3). The responses indicate a strong preference for robots to exhibit capabilities such as arm movement ($M = 8.6$; $SD = 1.24$), user recognition ($M = 8.5$; $SD = 1.22$), and verbal communication ($M = 8.1$; $SD = 1.61$). Additionally, the feature to play videos was generally appreciated ($M = 7.1$; $SD = 1.6$). The robot's ability to play music was moderately received ($M = 6.6$; $SD = 2.19$). Opinions varied more substantially regarding the robot's capacity to show emotions ($M = 5.7$; $SD = 3.61$). The capability for the robot to walk received a lower mean score, but also had a wider spread of responses ($M = 6.7$; $SD = 3$).

The open-ended question soliciting feedback about features unearthed intriguing suggestions and concerns. For example, $P3$ advocated for user feedback systems reflecting their performance level, alongside the robot's capacity to detect user disinterest. However, they also expressed a concern for the affordability of robotic technology, emphasizing the necessity for easy access in developing and under-developed countries, where it is currently beyond the reach of the majority in need.

On the aesthetic front, both $E1$ and $P5$ wished for improvements in the robot's cosmetic appeal. $E1$, in particular, envisioned the robot resembling Baymax from the movie 'Big Hero Six,' complete with articulated spinal and pelvic motions. $OT1$ expressed a desire for robotic lower extremities, while $RS1$ suggested a shorter robot height to better interact with toddlers.

4.3.2 Expert Views on Teleoperation

Next, we delve into expert opinions on teleoperation gathered from the surveys conducted during phase three of the interviews. As seen in Fig. 4.4, the experts expressed a strong appreciation for the concept of teleoperation ($M = 7.6$; $SD = 1.79$). This fondness

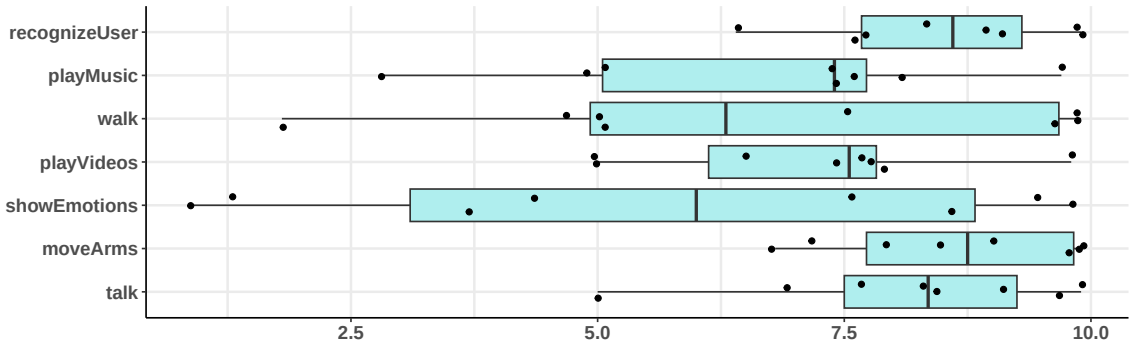


Figure 4.3: Expert responses about the features that they would like on a social exercise robot. Experts particularly like movement of arms, the ability to talk and recognize the user.

is reflected by notably high median ratings and minimal variability across all questionnaire items. Concerns about the system were minimal ($M = 1.6$; $SD = 1.6$), while enthusiasm for its customization potential was substantial ($M = 9.2$; $SD = 0.691$). Experts also recognized the system's potential in devising new exercises ($M = 8.7$; $SD = 1.5$) and enhancing social interaction with their clients ($M = 8.9$; $SD = 0.858$). Clear interest was expressed in learning how to operate the teleoperation system ($M = 8.2$; $SD = 1.79$), as well as in creating new robot behaviors ($M = 8.4$; $SD = 1.38$) and utilizing the system to teach these new behaviors ($M = 8.4$; $SD = 1.34$). Overall, the concept was deemed highly favorable ($M = 9.1$; $SD = 0.587$), and the experts agreed that it would be an enjoyable system to work with ($M = 8.5$; $SD = 1.25$). The open-ended feedback regarding teleoperation was overwhelmingly positive, with experts $OT1$, $P5$, and $RS1$ expressing a particular liking for the robot's facial expressions.

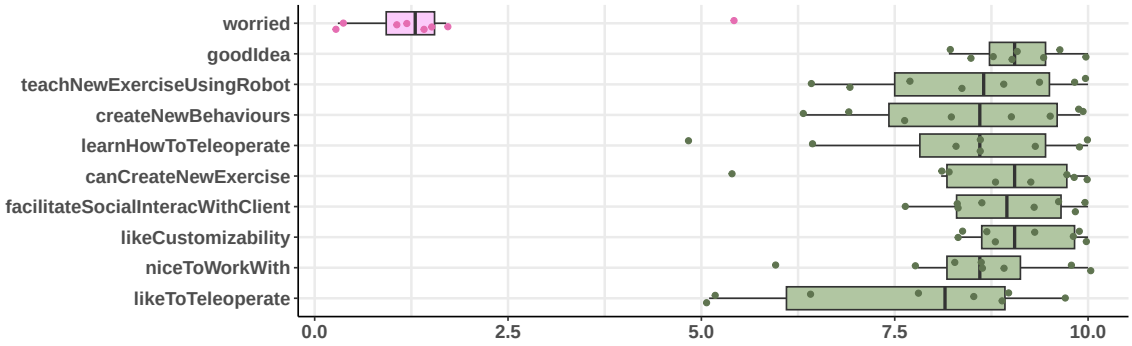


Figure 4.4: Expert feedback about teleoperation. Positive questions are colored green, and negative questions are colored pink. The experts were very positive about teleoperation systems.

Teleoperation Features

We further queried experts about their desired features for a teleoperation system (Fig. 4.5). The ability to control the robot's arms received the highest mean score ($M = 9.1$; $SD = 0.804$), indicating that this feature was particularly favored by participants. The capacity to move the robot's head also garnered strong approval ($M = 9$; $SD = 1.04$). Participants demonstrated a strong inclination for the feature allowing the robot to talk to the client ($M = 8.5$; $SD = 1.45$), suggesting that verbal interaction is highly valued. Controlling the robot's facial expressions was deemed highly beneficial, with a mean score of 8.8 and a standard deviation of 1.62, which implies a wide range of positive opinions.The function to display videos on an extra screen was generally well-received, albeit with a wider variation in responses ($M = 7.4$; $SD = 2.73$). The ability to control the robot's legs was also endorsed, though with a fair degree of variability ($M = 8.4$; $SD = 1.85$). Finally, the feature enabling the robot to play music was moderately appreciated ($M = 7.9$; $SD = 2.6$), indicating a more moderate consensus among participants on its utility.

The open-ended questions also garnered some interesting responses. $P2$ expressed a desire for the system to assist experts in transitioning between lying, sitting, and standing. $P3$ was interested in incorporating auditory and olfactory responses. $E1$ wished for more articulation along the robot's torso. $OT1$ expressed a desire for lower limb teleoperation to facilitate dance-based therapy. $P5$ advocated for the inclusion of more features to support activities of daily living, envisioning potential applications in speech therapy. Lastly, $RS1$ wished for hand teleoperation, including fingers.

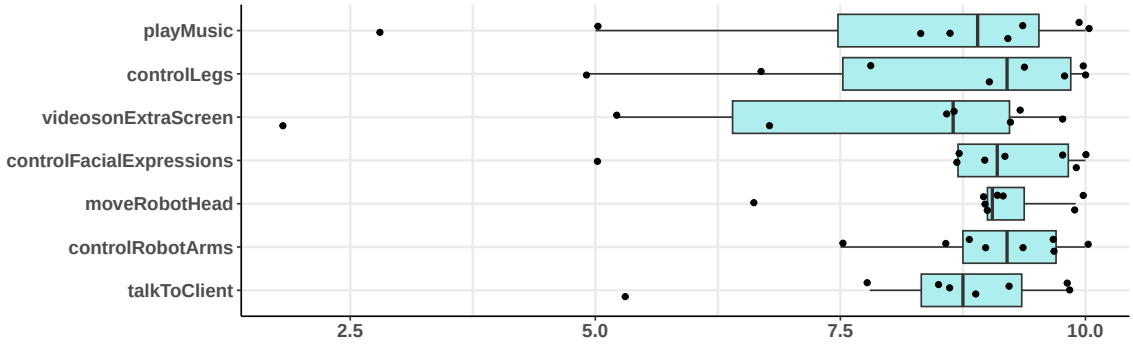


Figure 4.5: Expert responses to features they would like on a teleoperation system. Experts particularly like the ability to move the arms and head of the robots. They also like the ability to talk to clients during teleoperation.

4.3.3 Expert Opinions on Robots

Finally, we examined the evolution of expert opinion over the course of the study through pre- and post-study Robot Acceptance Surveys (Fig.4.6 and Fig.4.7). Analogous to Sec-

tion 3.3.7, these results were analyzed from the viewpoint of technology acceptance and object-centered sociality. Overall, expert perception towards robots had grown more positive post-study. The detailed summary statistics for these responses can be seen in Appendix Table A.9.

Fig. 4.6 showcases the survey results in light of the UTAUT model (Venkatesh *et al.*, 2003). There were no significant differences in expert expectations of robot performance (Fig. 4.6A). Although there were no significant differences in their attitudes towards technology, the trends illustrated a more favorable view of exercise robots post-study (Fig. 4.6B). In particular, statistical trends suggest they were less apprehensive about the robot causing harm to their clients ($T(6) = 20, p = 0.06, d = 0.82$); note that only $P1$ and $P2$ did not respond to this question as it was added later based on a conversation with $P2$). The study significantly impacted expert effort expectancy (Fig. 4.6C), with experts indicating a better ability to cooperate with the robot ($T(8) = 0, p = 0.02, d = 0.87$). The study did not affect their feelings of self-efficacy (Fig. 4.6D).

Fig. 4.7 addresses the categories related to user attributions of social and cultural constructs toward robots(Cetina, 1997). Generally, there were no significant differences in experts' desire to form groups with or around the robot (Fig. 4.7A). However, there was a statistically suggestive positive effect on how much they enjoyed the robot's presence ($T(8) = 5, p = 0.08, d = 0.6$). Their feelings of attachment towards robots also shifted favorably (Fig. 4.7B). No differences were observed in feelings of reciprocity (Fig. 4.7C). However, their perception of robots from a cultural context did change (Fig. 4.7D): after the study, they felt that people would be significantly more impressed if they possessed such a robot ($T(8) = 3, p = 0.04, d = 0.74$).

4.4 Discussion

This preliminary study sought to examine expert perspectives on social exercise robots and their teleoperation. The gathered data yields crucial insights into the perceived usefulness and appeal of various features associated with an exercise robot and its teleoperation system.

There was notable diversity in expert opinions regarding social exercise robots, as evidenced by the high standard deviations (Fig. 4.2). Nevertheless, a consistent theme emerges in their collective appreciation for the concept of teleoperation. This sentiment was supported by high mean scores and low variability throughout the teleoperation questionnaire items (Fig. 4.4), confirming a broad consensus among experts.

The post-study RAS results demonstrate a more favorable perception of the robot among the experts. Specifically, there was an improvement in attitudes towards robots (Fig. 4.6B), fondness for the robot's presence (Fig. 4.7A), and attachment to the robot (Fig. 4.7B). The desire among experts to cooperate with the robot was also positively influenced by the study (Fig. 4.6C). Furthermore, the notion of robot ownership was found to be notably impressive by the participants (Fig. 4.7D). While initial perceptions of the

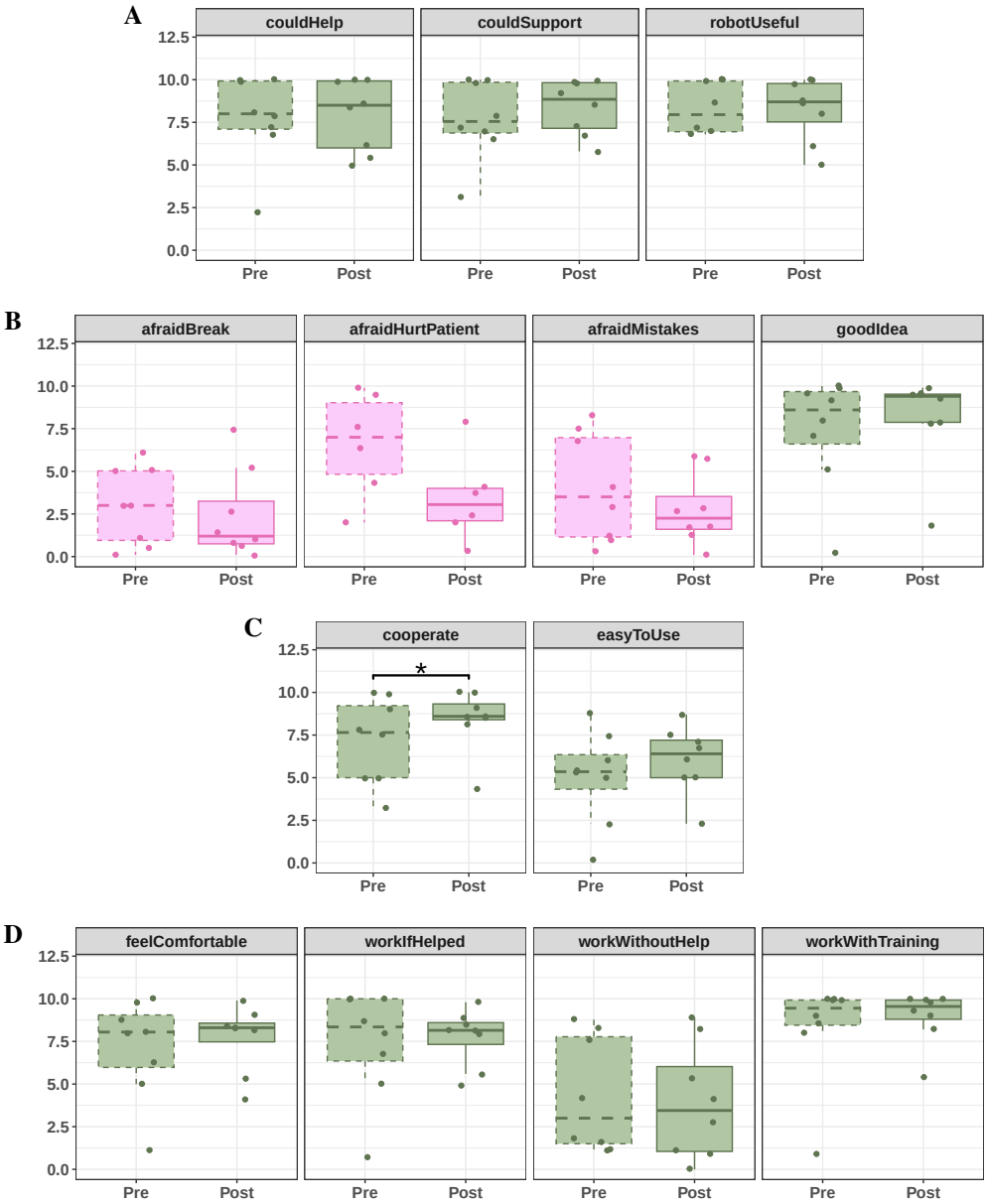


Figure 4.6: Robot Acceptance Survey categories based on the UTAUT model before (pre) and after (post) the study. Positive questions are colored green, and negative questions are colored pink. **(A)** Expert responses regarding performance expectancy of robot. There were no significant differences. **(B)** Expert responses to questions assessing their attitudes toward technology. Though there were no significant differences here, their attitudes towards robots improved after the study. **(C)** Expert responses about effort expectancy. Experts felt significantly more positive about cooperating with the robot after the study. **(D)** Expert responses to self-efficacy measures. There were no significant differences in this category.

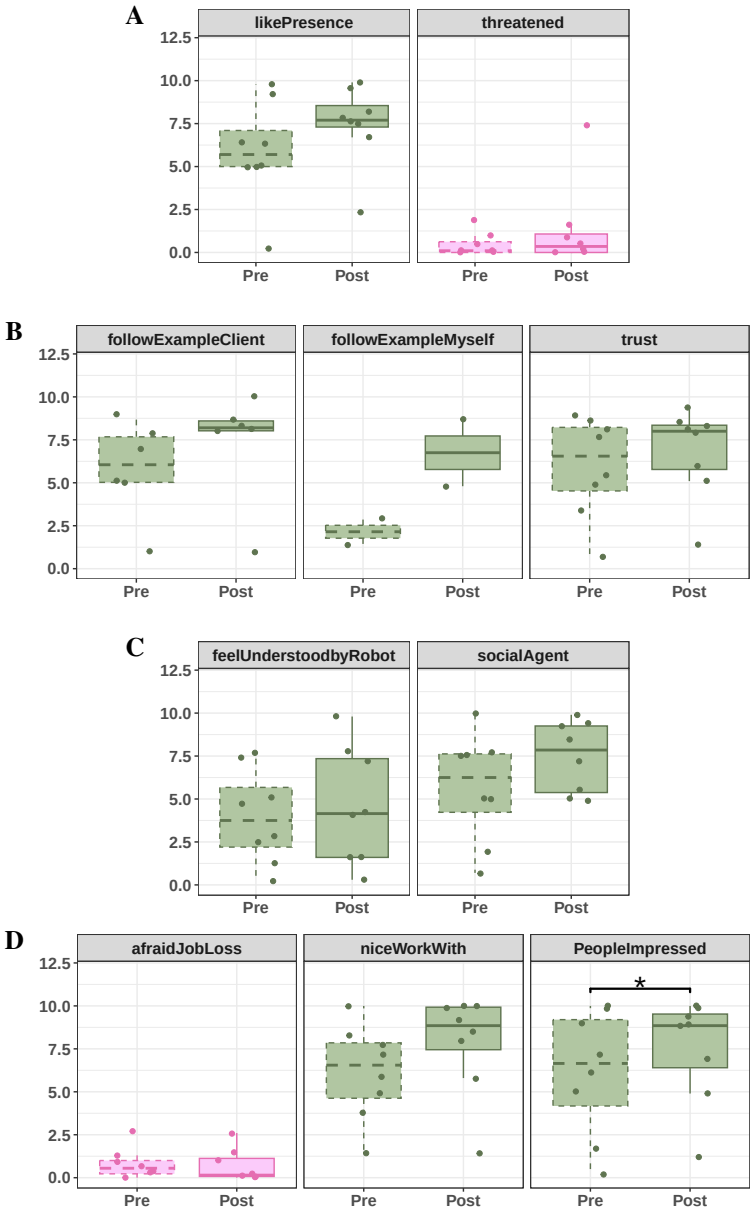


Figure 4.7: Robot Acceptance Survey categories based on the concept of object-centered sociality. Positive questions are colored green, and negative questions are colored pink. **(A)** Expert responses to forms of grouping category items. Though not significant, they like the presence of the robot a bit more after the study. **(B)** Expert responses regarding attachment toward robots. Experts were generally more positive about robots after the study **(C)** Feelings of reciprocity reported by experts. There were no significant differences sin this category. **(D)** Expert responses to cultural perception of robots. After finishing the study, experts significantly felt that people would be impressed if they had a robot.

exercise robot were mixed, the positive results regarding teleoperation indicate that this feature significantly alleviated initial concerns about the social robot. Hence, despite the initial variability in views, the study appears to have positively shaped the overall perception of social exercise robots and their potential for teleoperation.

With regards to the exercise robot's features, there was a marked preference for functionalities related to communication and arm movement (Fig. 4.3). In addition, the most favored teleoperation feature was arm control, followed by the control of the head, facial expressions, and client communication. These observations suggest that interactive and dynamic functionalities are deemed crucial for the exercise robot's effectiveness. Moreover, the experts' interest in cooperating with the robot, coupled with these preferred features, indicates potential acceptance of arm teleoperation technology for customizing robot behavior.

Intriguingly, the feature of user recognition was also highly appreciated, implying the value participants place on personalized experiences and the importance of the robot's ability to adapt to individual users.

Nonetheless, this study is in its preliminary stages. Given the diverse participant demographics, there is a need to gather additional data for a more comprehensive overview of expert perceptions. Moreover, the interview data needs to undergo detailed thematic analysis to grasp the nuances of participant responses. Finally, evaluating the phase one data may reveal key challenges that could be addressed by robots.

4.5 Conclusion

In conclusion, this study sheds light on the potential and challenges of social exercise robots and teleoperation, based on expert feedback. The enthusiasm shown for teleoperation confirms that using teleoperation to generate robot gestures for exercise is a promising avenue of research.

Chapter 5

OCRA: An Optimization-based Customizable Retargeting Algorithm for Teleoperation

5.1 Introduction

While socially assistive robots (SAR) have shown potential as exercise coaches (Matarić and Scassellati, 2016; Fitter *et al.*, 2020), they are currently far from being usable in real-world scenarios, where they will need to be able to adapt to changing user needs (Okamura *et al.*, 2010). Thus, this chapter diverges from the user within the minimally supervised environment and instead focuses on identifying an efficient approach to engineer novel robot behaviors. Specifically, we are interested in sophisticated socially assistive robots that can perform dynamic social-physical interactions with a user similar to those showcased by Fitter *et al.* (2020). These robots typically leverage pre-programmed gesture-based movements in the form of demonstrations (Pulido *et al.*, 2019; Matsusaka *et al.*, 2009; Lewis *et al.*, 2016) to teach exercises to their users. Experts in exercise therapy have the nuanced understanding and experience to know what is best for their patients. However, the transition from ideation to creation of these robotic movements with expert and roboticist involvement is a complex and time-consuming process Fitter *et al.* (2020). By directly involving these experts in the design phase, the development can be more streamlined, and the resulting motions and responses of the robot can be more precisely tailored. Therefore, we believe the motions and responses of the robot can best be designed by exercise therapy experts who are familiar with the needs and therapy objectives of their patients.

A system designed for clinicians must be easy-to-use with a low learning curve (Klaic *et al.*, 2022). Motion-capture-based teleoperation is known to be the most effective and intuitive methodology to enable a human operator to teach gestures to a robot (Hedlund *et al.*, 2021). Moreover, recent research in socially assistive robotics advocates for teleoperation as an ideal level of autonomy (Elbeleidy *et al.*, 2022). Bolstered by the strong therapist opinions reported in Chapter 4, we believe that *an intuitive teleoperation system would facilitate the long-term personalization and effectiveness of such exercise robots*.

Within the domain of human-robot interaction (HRI), teleoperation interfaces used by therapists are primarily graphical user interfaces used to control affective or verbal behaviors of the robot (Elbeleidy *et al.*, 2022, 2021). Teleoperation has been successfully used to teach robots simple movements and exercises (Tanguy *et al.*, 2016; Görer *et al.*, 2013). Of particular note is an example of teleoperated hugging achieved via an algorithm that combines force control and motion retargeting (Kaplish and Yamane, 2019). While such an algorithm is suitable for creating close physical contact between a human and a robot, it does not focus on creating free-space robot motion that is human-like; thus, such an approach is not suitable for our goals. We require a system that is easy-to-use, can create smooth human-like motions, and can replicate social interactions.

The so-called **retargeting problem** has been studied extensively by both the robotics research community (Darvish *et al.*, 2023) and the computer graphics research community (Akber *et al.*, 2023; Aristidou *et al.*, 2018). Roboticists have created algorithms that rely on inverse kinematics (Ju *et al.*, 2014; Aristidou *et al.*, 2018), optimization (Dariush *et al.*, 2009), admittance control (Ju *et al.*, 2014; Lenz and Behnke, 2023), morphological adaptation between kinematic chains (Sousa *et al.*, 2011; Lentini *et al.*, 2019), machine learning Wang *et al.* (2019), or a combination of these methods (Rakita *et al.*, 2017; Ayusawa and Yoshida, 2017). Other interesting teleoperation methods have also utilized dynamic motion primitives (Liang *et al.*, 2021) and generative models (Wang *et al.*, 2019) to map human motion onto robots. Various methods have distinct advantages and disadvantages. For example, though inverse-kinematics-based methods have been studied extensively (Ju *et al.*, 2014; Aristidou *et al.*, 2018), they can be prone to singularities, often have multiple solutions, and focus heavily on end-effector pose rather than that of the whole arm. They are thus used in combination with other techniques such as optimization Rakita *et al.* (2017); Wang *et al.* (2015a). Methods based on machine learning tend not to function in real time (Tosun *et al.*, 2014; Wang *et al.*, 2019) or are constrained to specific motions (Zhang *et al.*, 2022). However, the most significant challenge in selecting an appropriate algorithm lies in the absence of adequate validation or user studies, exacerbated by the lack of widely accepted standard evaluation methods (Darvish *et al.*, 2023).

Furthermore, researchers in HRI employ a wide variety of robots with differing kinematic configurations Matarić and Scassellati (2016). Most motion-mapping algorithms require substantial customization when adapting them to platforms other than their original design. The bespoke nature of these algorithms not only complicates implementation but also hinders performance comparison for specific applications. Consequently, we set out to create a customizable and generalizable teleoperation system that is not only user-friendly for non-expert operators but is also carefully evaluated. We posit that such an algorithm can only be completely evaluated by two sets of user studies: one that focuses on user perception of robot-generated motion and another that focuses on the usability of the teleoperation system. This chapter will focus on the development of the algorithm and evaluation of the user perception of the generated motion.

Our algorithm, OCRA, has been detailed in Section 5.2, and the associated study has

been outlined in Section 5.3. The performance of the algorithm and the results of the user study have been described in Sections 5.4 and 5.6, respectively. The associated discussion can be found in Section 5.7, followed by the conclusion in Section 5.8.

5.2 Algorithm

We created an optimization-based kinematic retargeting algorithm that functions in real time. OCRA takes inspiration from the method proposed by Tosun *et al.* (2014) due to its generalizability to arbitrary serial robots. However, that method ignores end-effector orientation, matches the user pose imperfectly, and does not function in real time. In this section, we detail the end-to-end process of taking a desired pose for the robot arm, measured from the pose of the human operator's arm, and transforming it into the commanded pose for the robot's arm. Our algorithm requires only the current configurations of both the input and output kinematic chains to perform retargeting. We consider robot arms with only revolute joints as they are far more common than prismatic joints. Fig. 5.1 provides a higher-level overview of the algorithm. We compose the retargeting problem as an optimization task that requires real-time knowledge of only the human joint angles, the human hand orientation, and the current joint angles of the robot, as well as the forward kinematics for both the human arm and the robot arm.

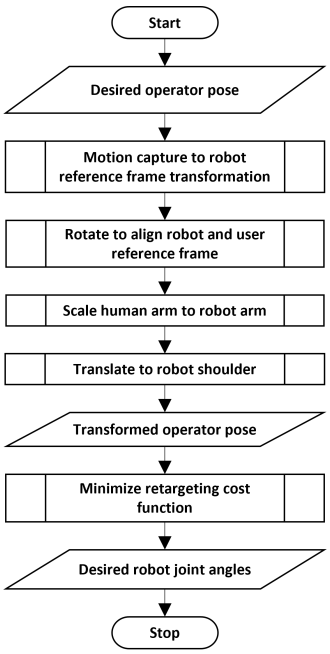


Figure 5.1: Flowchart of OCRA. The operator's arm pose is obtained from a motion-capture system. A set of homogeneous transformations are then applied to this pose before performing the minimization to obtain the desired robot joint angles.

5.2.1 Initial Transformations

Before performing the retargeting, we need to apply a set of homogeneous transformations to the kinematic chain of the human arm. Let us assume that the source chain (typically a human arm), h, has m joints. Each joint of the source has an orientation represented by a quaternion, q_h, and a position, $\vec{d}_h$. Thus, the source chain in its own reference frame, H, can be represented as follows:

$$h^H = \begin{bmatrix} q_{h_1}, q_{h_2} \ldots q_{h_m} \\ \\ \vec{d}_{h_1}, \vec{d}_{h_2} \ldots \vec{d}_{h_m} \end{bmatrix}^H \tag{5.1}$$

Similarly, the target chain (typically the robot arm being controlled) is represented by τ and has n joints. The reference frame of the target is T. Each joint of the target has an orientation represented by a quaternion, q_τ, and a position, $\vec{d}_\tau$. Thus, the target chain in its own reference frame, T, can be represented as follows:

$$\tau^T = \begin{bmatrix} q_{\tau_1}, q_{\tau_2} \ldots q_{\tau_n} \\ \\ \vec{d}_{\tau_1}, \vec{d}_{\tau_2} \ldots \vec{d}_{\tau_n} \end{bmatrix}^T \tag{5.2}$$

Rotate to world reference frame

We use the reference frame of the target robot's stationary base as our world reference frame. All rotations are performed using unit quaternions Hart *et al.* (1994). Since these frames are manufacturer defined, we manually calculate the rotation, represented by quaternion q_H^T, that transforms the source chain in reference frame H into the target reference frame, T, as follows:

$$\begin{aligned} h^T &= q_H^T \times h^H \\ &= \begin{bmatrix} q_{h_1}, q_{h_2} \ldots q_{h_m} \\ \\ \vec{d}_{h_1}, \vec{d}_{h_2} \ldots \vec{d}_{h_m} \end{bmatrix}^T \end{aligned} \tag{5.3}$$

Next we apply a combined rotation (q_w), scaling ($Scal_\tau$), and translation ($Trans_{sh}$) operation to the source chain using the below equation. These transformations are calculated in the following subsections.

$$\bar{h}^T = Trans_\tau Scal_\tau q_w h^T \tag{5.4}$$

Torso-invariant rotation

A rotation is performed to make the source kinematic chain invariant to the torso rotation of the human body. This step is optional and can be omitted if capturing the torso rotation is necessary. There are two different methods to accomplish this step based on the availability of the source chain's torso or waist orientation, represented by q_w. If this quaternion is available via the motion-capture system, then it can be directly used in equation (5.4). If not, q_w can be calculated from the bilateral shoulder positions of the human. Let the left and right shoulder positions be represented by vectors $\vec{d}_{\tau_{shL}}$ and $\vec{d}_{\tau_{shR}}$, respectively. We calculate the vector connecting the shoulders of the source chain through the clavicles (in case of the human) as follows:

$$\vec{d}_{\tau_{sh}} = \vec{d}_{\tau_{shL}} - \vec{d}_{\tau_{shR}} \tag{5.5}$$

Then, we find the angle by which it is rotated, θ_{sh}, about the vertical axis of the world frame or the target frame, τ. Let the horizontal axis, be represented by the vector $\vec{o}^{\tau}$ which goes through the shoulders of the target when it is in neutral position. Then the vector perpendicular to $\vec{d}_{\tau_{sh}}$ and $\vec{o}^{\tau}$, $\vec{v}^{\tau}$, can be calculated using a cross product as follows:

$$\vec{v}^{\tau} = \vec{o}^{\tau} \times \vec{d}_{\tau_{sh}} \tag{5.6}$$

Now, v_z^{τ} represents the element along the vertical axis. We can calculate θ_{sh} as follows:

$$\theta_{sh} = \begin{cases} -\tan^{-1} \frac{||\vec{v}^{\tau}||}{\vec{o}^{\tau} \cdot \vec{d}_{\tau_{sh}}} & \text{if } v_z^{\tau} < 0 \\ \tan^{-1} \frac{||\vec{v}^{\tau}||}{\vec{o}^{\tau} \cdot \vec{d}_{\tau_{sh}}} & \text{otherwise} \end{cases} \tag{5.7}$$

Thus, the quaternion q_w can then be calculated as a rotation by $\angle\theta_{sh}$ about the vertical axis of the world frame.

Scale to robot arm length

The next step is to scale the input chain such that it is of similar length to the target chain. Let the length of the source chain be ℓ_h and the length of the robot arm be ℓ_τ. The scaling parameter $Scal_\tau$ can be calculated as follows:

$$Scal_\tau = \frac{\ell_\tau}{\ell_h} \tag{5.8}$$

Translate to robot shoulder

Finally, we perform the translation. Here, we superimpose the base joint of the source chain ($\vec{d}_{h_{sh}}$) to the corresponding position of the target ($\vec{d}_{\tau_{sh}}$) chain. The choice of this position is very important and can affect the mapping. It determines how effectively the

end-effector of the robot can track the human hand. In the case of Baxter, we choose this position to be the projection of the S0 joint at the same height as the S1 joint to make it correspond to the human shoulder joint. The translation to be applied to the human arm to position it at the shoulder of the robot, $Trans_\tau$, can then be calculated as follows:

$$Trans_\tau = \vec{d}_{\tau_{sh}} - \vec{d}_{h_{sh}} \tag{5.9}$$

When applied to the full captured pose of the human arm, these transformations scale the human arm up or down in length to match the length of the robot arm, and they position and orient it so that the center of the human shoulder joint lies along the axis of the robot's first joint as close as possible to the axis of the second joint. This transformed version of the operator's arm is then ready to use in the retargeting process.

5.2.2 Retargeting

Assuming the human arm has been appropriately scaled and transformed so that the shoulder joint coincides with the robot's first joint, OCRA minimizes the retargeting error (ε_R), which is a weighted sum of the squared arm skeleton error (ε_s) and the squared hand orientation error (ε_o):

$$\varepsilon_R = \alpha\varepsilon_s^2 + \beta\varepsilon_o^2. \tag{5.10}$$

Each type of error (skeleton, orientation) is normalized by the maximum possible value and is thus unitless. The constant weights (α, β) are customized by the user before teleoperation and are subject to the following constraints:

$$0 \leq \alpha \leq 1, \quad 0 \leq \beta \leq 1, \quad \alpha + \beta = 1 \tag{5.11}$$

At each time step, the algorithm's goal is to find the robot joint values that minimize (5.10) for the operator's current pose. The optimization employs Newton's conjugate-gradient method (Nash, 1984) and uses the previous time step's solution as the next initial guess. The solution of the optimization is bounded by the joint limits of the target chain.

Arm Skeleton Error This error term is a curve similarity metric that measures how different the current configuration of the source chain (human arm) is from that of the target chain (robot arm); our formulation takes inspiration from Fréchet's distance (Fréchet, 1957; Alt and Godau, 1995). We define the arm skeleton error as the sum of the shortest distances from each joint of the source kinematic chain to the target chain and vice versa.

Let s_i be the shortest positive distance from the ith joint of the source chain to the closest segment of the target chain, and let t_j be the shortest distance from the target chain's joint j to the closest segment of the source chain. The source chain has m joints, and the target chain has n joints. To represent the maximum possible value, the normalization factor, $\bar{\varepsilon}_s$, is the sum of the cumulative sums of the segment lengths of the source

and target chains. Let the source segment lengths be $\ell_{h_1}, \ell_{h_2} ... \ell_{h_m}$ and the target segment lengths be $\ell_{\tau_1}, \ell_{\tau_2} ... \ell_{\tau_n}$. Then, we can calculate the normalization factor as follows:

$$\bar{\varepsilon}_s = \sum_{k=1}^{m} \sum_{h=1}^{k} \ell_{h_h} + \sum_{k=1}^{n} \sum_{h=1}^{k} \ell_{\tau_h} \tag{5.12}$$

Thus, the normalized skeleton error can then be defined as follows:

$$\varepsilon_s = \frac{\sum_{i=1}^{m} s_i + \sum_{j=1}^{n} t_j}{\bar{\varepsilon}_s}. \tag{5.13}$$

End-effector Orientation Error The orientation of the hand plays a vital role in both task-based teleoperation and social interaction. Thus, we add a second metric that quantifies errors in the orientation of the end-effector frame. Since robot and motion-capture manufacturers choose end-effector frames somewhat arbitrarily, we first define a custom frame that is positioned and oriented in a suitable way on each end-effector. For our application, it is in the middle of the human's palm and at the center of the robot's palm, which coincides with the base of Baxter's parallel jaw gripper. Let q_ς and q_τ be the quaternions representing the orientations of the source and target end-effector frames, respectively. We can then obtain the quaternion Q_d that represents the orientation difference between the two frames as follows:

$$Q_d = q_\varsigma / q_\tau \tag{5.14}$$

From Q_d we calculate the positive relative angle between these frames, θ_d, using the axis-angle representation. This angle can have a maximum value of π radians.

$$\theta_d = 2 \tan^{-1} \left(\frac{\pm \sqrt{1 - Q_d^2}}{Q_d} \right) \tag{5.15}$$

Thus, normalized end-effector orientation error is defined as follows:

$$\varepsilon_o = \theta_d / \pi \tag{5.16}$$

5.2.3 Initialization

The robot starts at an initial pose P before teleoperation begins. To facilitate a smooth start to controlling the robot's motion, we designed our system so that the human has to initially mimic the pose P in order to initiate the teleoperation. This procedure avoids sudden unexpected robot movements at the beginning; everything starts seamlessly because the robot's arms are already close to the pose commanded by the human's arms.

In our current implementation, the start of teleoperation depends only on the skeleton error and disregards hand orientation error as operators naturally tend to mimic the robot's hand orientation. When the shape error for an arm is less than a certain threshold, teleoperation automatically begins, so the robot's arm begins following the movements of the corresponding arm of the operator. This initialization routine also resolves the issue of optimization requiring a good initial guess for suitable robot joint angles at the start. For initialization poses, we generally recommend a non-straight robot arm pose with a bent "elbow" joint. The pose used in our current implementation incorporates shoulder and elbow flexion: the robot's upper and lower arms extend forward from its shoulders, with the wrists both bent upward at right angles, so the hands are vertical. Human operators tend to mimic this pose by bending their elbows and holding their forearms and hands up. This initialization pose works well despite the mismatch between robot wrist joint and human elbow joint because OCRA operates in task space (hand orientation and arm skeleton) rather than joint space.

5.3 Methods

To validate OCRA, we conducted three sets of experiments: a panel of benchmarking tests, a generalizability experiment and finally a user perception study. All validation experiments utilize motion-capture data that was collected using an Xsens Link (Movella Inc., 2018). The benchmarking experiments and the user study utilize a Rethink Robotics Baxter Research Robot (Robotics, 2013).

We chose a set of four motions to be used across all experiments. These were exercise-related motions of differing complexity representing a diversity of actions. Each motion was rendered with four values for the arm-skeleton weight: $\alpha = [0.001, 0.33, 0.67, 1.00]$. Fig. 5.2 shows sample images from the study. The rationale behind the selection of the motions, along with their particulars, is elaborated upon below.

Arm Lift: Adapted from the validated previously validated exergame Roboga (Fitter *et al.*, 2020). This movement commences with a full-range shoulder abduction, during which a forearm supination is performed mid-motion. The abduction continues symmetrically until both arms are held vertically up. Following this, the sequence transitions into a full-range shoulder adduction, accompanied by a forearm pronation midway through. The sequence concludes with the arms straight down.

Bent Arm Touch: Modified version of the initialization pose (Section 5.2.3) for OCRA. This modification was done to better showcase the orientation error. This movement amalgamates shoulder abduction and elbow flexion, then sequentially introduces a combination of forearm pronation and supination, coupled with internal and external rotation of the shoulder.

Dance Sequence: Adapted from the validated previously validated exergame Flamenco (Fitter *et al.*, 2020). The most complex of all four motions, this movement showcases two actions from flamenco dance. One move is a curvilinear motion with the right arm ascending and the left arm descending in synchrony. This creates an elegant arc, with one arm gracefully curving above the body and the other below. This motion is followed by a distinctive, sideways clap. The initial curvilinear motion is then repeated, succeeded by an analogous hand clap on the left side. The sequence concludes with a final replication of the curvilinear movement before returning to a neutral arm posture.

Front Clap: Selected to demonstrate contact between end-effectors and considerable motion in a distinct plane. In this movement, the person initially performs a slight shoulder abduction, followed by elbow flexion, positioning the palms to face each other. The hands then elevate in front of the face where they meet in a clap, before the arms return to rest at the sides in a neutral position.

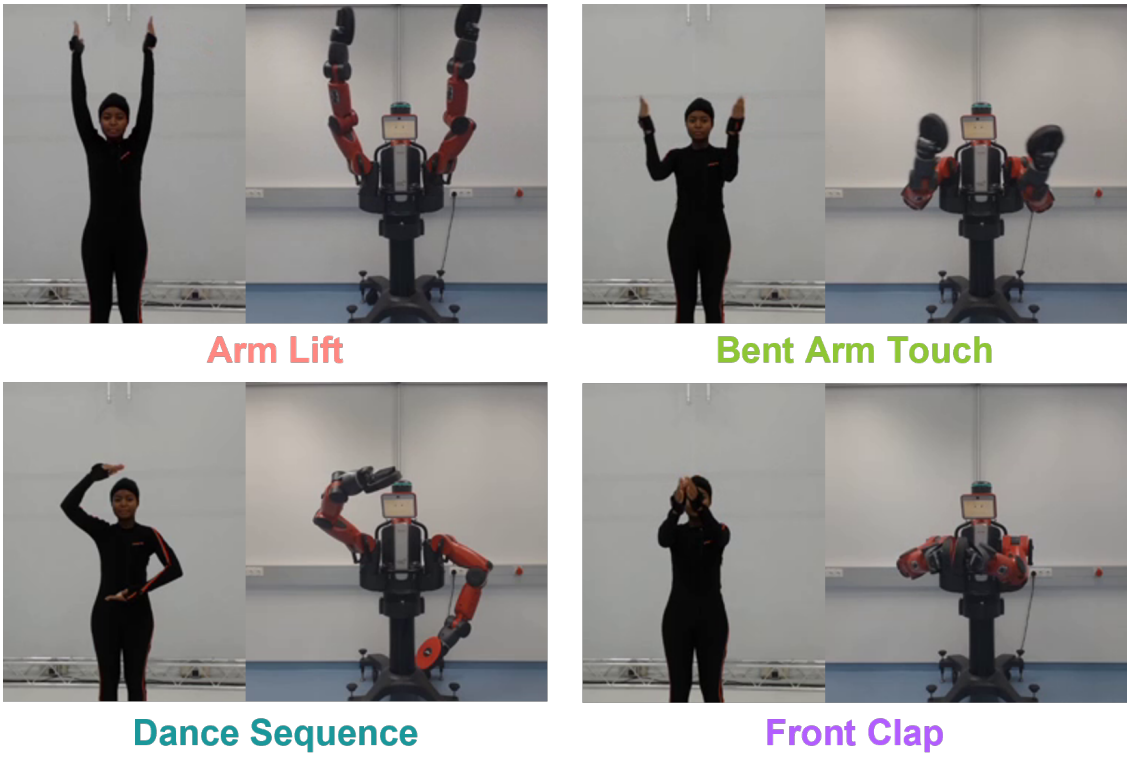


Figure 5.2: Sample screenshots of the four movement videos shown in the study for the $\alpha = 0.67$ condition. The two-armed robot tries to match the human operator's hand orientations while prioritizing matching their arm skeletons.

5.3.1 Benchmarking Tests

To evaluate OCRA's performance, we assess the impact that different weight sets exert on various movements through a one-way repeated measures analysis. Our examination encompasses both the commanded errors, referring to inaccuracies in the commands generated by the algorithm, and the actual errors, which denote deviations generated by the robot during the execution of the motion. Specifically, we scrutinize the skeleton error, orientation error, and retargeting error for all four motions and four weight sets previously described. All measures represent the grand mean, i.e., the mean of the mean errors of the frame-by-frame error for both the left and right arms. In addition to these considerations, we also gauge the latency of the algorithm, quantifying it as the grand mean of the time required for one complete iteration.

5.3.2 Generalizability Experiment

In order to evaluate the generalizability of the algorithm, we also implemented it on a Rethink Robotics Sawyer robot (Rethink Robotics, 2023). Although Sawyer possesses the same number of degrees of freedom as each of Baxter's arms, the configuration of its joints and the lengths of its links are markedly different. Since the experiment was conducted solely in simulation, our analysis focused on comparing the commanded joint angles generated by the algorithm for both robots. Specifically, we examined the impact of various errors including skeleton error, orientation error, and retargeting error, employing a non-parametric ANOVA for a comprehensive assessment.

5.3.3 User Perception Experiment

Human perception of the motion of others stems from our immense experience in performing activities ourselves (Blake and Shiffrar, 2007). Thus, an intuitive teleoperation system should create human-like motions. Additionally, we were curious if users could perceive the differences between the different weight sets. Thus, we hypothesize that

H1 Retargeting that combines both types of errors will outperform retargeting that considers only skeleton or only orientation error.

H2 The relative performance of different weight sets will be consistent across motions.

To examine our hypotheses, we conducted a two-factor within-subjects study to discover how users perceived robot motions created by OCRA (with different weights) compared to those of the human operator. The 4×4 repeated measures study evaluated four different operator movements as the first independent variable and four different weight sets as the second independent variable. The study was approved by the Ethics Council of the Max Planck Society under the Haptic Intelligence Department's framework agreement as protocol F004B.

The study was run completely virtually. The survey was conducted using LimeSurvey version 3.22. All survey questions, themes and templates were modified to ensure accessibility. The survey could be easily completed on any mobile device with a compatible browser or on a computer. Though participants were recommended to use landscape mode on a mobile device, the survey worked well on devices with a screen width greater than 320px in portrait mode. This resolution should work well for almost anyone completing the survey based on screen-size usage statistics (StatCounter, 2022).

The study survey link was accessible for a period of approximately 7 weeks. Participants could access the survey via the link provided in our recruiting materials (email announcement and flyer). This link directed them to the online consent form. Once they provided informed consent by typing their full name, they received a link to the user study survey. To ensure participant confidentiality, the survey and the consent form were stored separately. This measure was taken to de-identify the data from the participants, safeguarding their privacy and confidentiality.

The design of the video-based study format has been adapted from the guidelines provided for the simultaneous double stimulus for continuous evaluation (SDSCE) method created by the International Telecommunication Union (International Telecommunication Union, 2019). This methodology calls for four phases of the study: a training phase, a demonstration phase, a mock phase and then the study phase. To fulfill the requirements for the training phase, the study link directed participants to an introductory page with instructions on how to complete survey questions and a labelled image of the robot (Appendix Section A.3.1). The participants then completed a demographics questionnaire (Appendix Section A.3.2). Since the study was run completely virtually, a demonstration could not be run. As an alternative, clear instructions were provided with every video to support the participant (Appendix Section A.3.2). The mock phase must include sequences that are not part of the actual study to avoid inducing bias for those stimuli. Thus, this phase comprised four videos of two operator movements that are not included elsewhere in the analysis. These videos were presented in a random order to every participant and included one video from each weight combination.

Since the study was virtual, we also included attention-checking videos where the operator motion was very different from that of the robot. Every fifth video was an attention-checking video. The first attention-checking video was the same for all participants as this video utilized an operator motion presented in the stabilization videos. The remaining four attention-checking videos were presented in a random order to all participants, interspersed with their main trials. After the mock phase, the participants saw videos of the set of 16 motions (4 weights $\times$ 4 motions seen in Fig. refRetargeting:fig) plus the four attention-checking videos. The dependent variable was a similarity rating per video where the participants had to rate the similarity of the operator movement to that of the robot.

For each question in the main part of the study, participants were presented with videos of two-armed human motion positioned next to and synchronized with a corresponding Baxter robot motion (except attention-checking videos, which had a non-synchronized

and Baxter motion). Participants rated how similar each robot motion was to the human motion on a visual analog scale (VAS) from 0 (worst) to 100 (best). Based on the recommendations of Matejka *et al.* (2016), we opted for a slight variation of the banded design of the VAS with five labels (Strongly Different, Different, Neither Similar nor Different, Similar, Strongly Similar). Such a design reduces the artificial bias generated from the design of the VAS while maintaining a high precision of results. The slider had no anchor present unless the participant clicked on it. The second question presented for each video was an optional free-response question where the participant could explain their rating.

After participants saw all 20 videos in the main part of the study (16 study videos and four attention-checking videos), they were presented with two free-response questions inquiring about aspects that they considered when making comparisons between the operator and robot motions and what suggestions they had for enabling the robot to match the human movement better (Appendix Section A.3.2). The final survey page (Appendix Section A.3.2) presented a ranking question to enable them to rank the items they thought were important when comparing human and robot motion. Participants were then given an opportunity to add an item to the ranking list. The study ended with an optional free-response question where participants could provide any type of feedback about their experience. The entire study typically lasted around 20 minutes for completion, but participants were free to take as much time as they needed. There were no limitations on the number of breaks they could take during the process.

5.3.4 Participant Information

In order to determine the minimum sample size required for this user study, we conducted an a priori power analysis using MorePower version 6.0.4 (Campbell and Thompson, 2012). The results showed that a two-way repeated measures ANOVA test would require at least 52 participants to obtain 95%power for detecting a medium effect at a significance of $\alpha = 0.05$. We collected additional participants to account for variability in user ratings, differences in devices used to complete the survey and failed attention tests. Thus, the total of 72 participants who provided consent and completed the study can be considered sufficient. We recruited by posting on notice boards of local universities, internal mailing lists and robotics mailings lists. As mentioned above, physical recruitment notices contained a QR code linking to the consent form, and recruitment e-mails contained links to the online consent form.

Before conducting the relevant statistical analyses, we first identified those participants whose responses were reliable. We defined two exclusion criteria. Responders who erroneously rated more than one attention check question higher than 50 (Fig. 5.3) and had a very low standard deviation ($SD < 5$) in their similarity ratings of the main study videos were excluded from the study. One attention-check video had a segment where the robot and operator motions were quite similar, causing some participants to rate this video higher even though the full motions did not match. We took this into account when

assessing which participants to keep. We identified two responders who fulfilled both of these exclusion criteria. Thus, all the analysis and data presented in this chapter has $N = 70$.

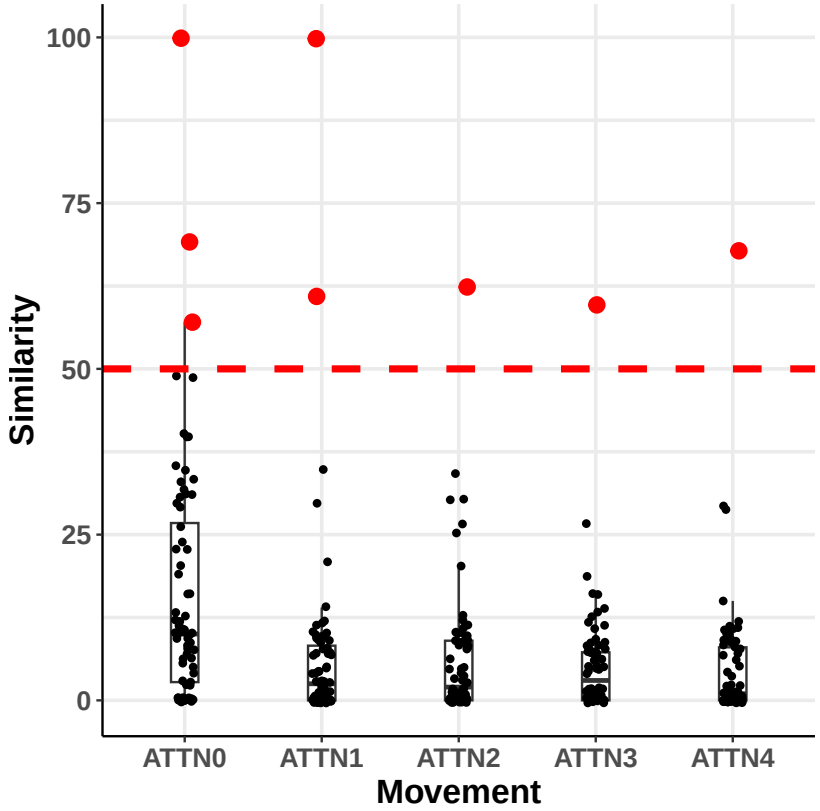


Figure 5.3: Participant similarity ratings of attention-check videos. The dashed red line shows the threshold above which participants were considered for exclusion. The larger red dots show the ratings that are above this threshold.

Our participants (age: $M = 35.44$, $SD = 12.297$) came from 18 different countries (Fig. 5.4A). 50% of our participants were women, and the remaining 50% were men, with 0% other. 70% of the participants had some form of technical education. We also asked participants about their expertise in assessing human arm motion, expertise with robots and expertise with the Baxter robot itself (Fig. 5.4B).

5.3.5 Analysis

All quantitative metrics derived from the algorithm were subjected to analysis using non-parametric statistical tests in R version 4.2.3. The selection of a non-parametric test was influenced by the limited number of data points available in each group. The analysis involved utilizing a repeated measures Aligned Rank Transform (ART) ANOVA, as

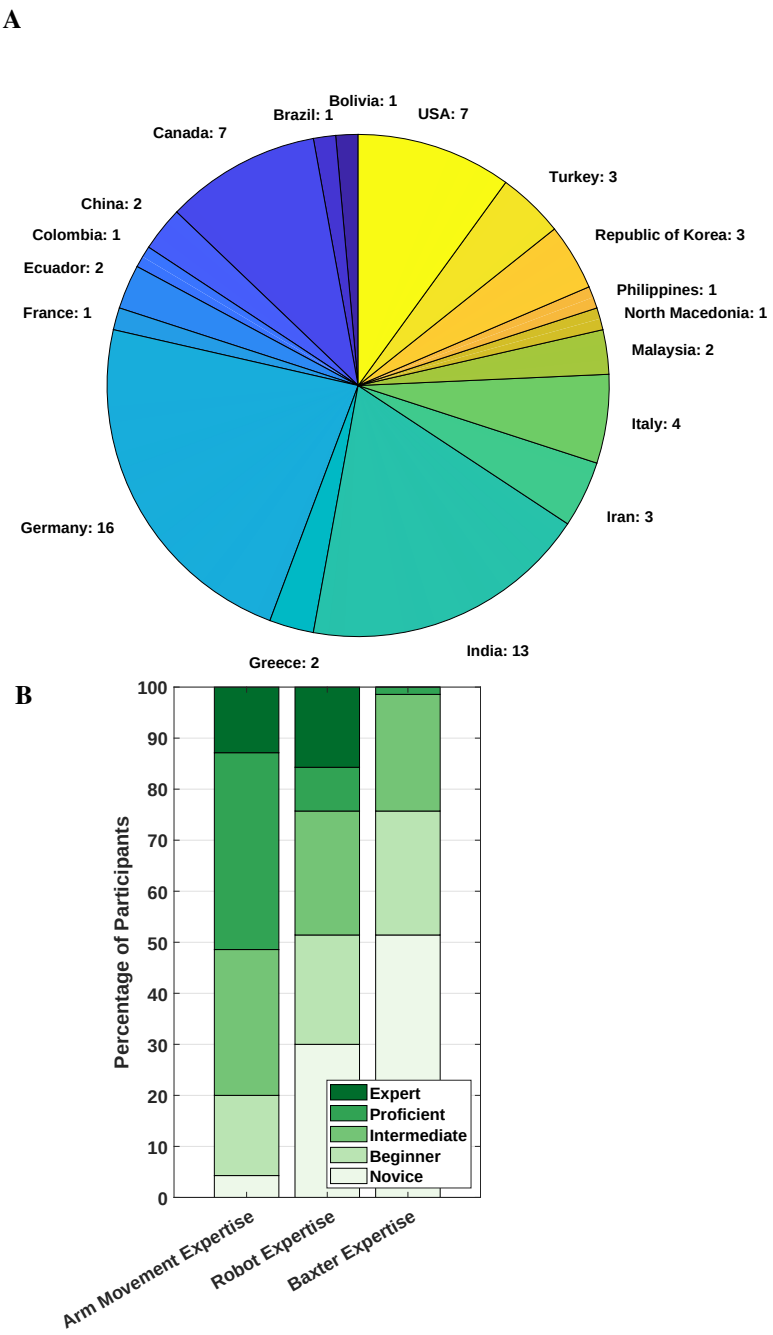


Figure 5.4: **(A)** Distribution of participant home countries. Participants hailed from 18 different countries. **(B)** Participant experience with arm movements, robots and Baxter. The participants exhibited a diverse range of experience in fields related to arm movements. They also had differing levels of expertise with robots. Finally, more than 50% of participants had never seen a Baxter robot before this study.

described by Wobbrock *et al.* (2011). To ensure a rigorous evaluation, all post-hoc comparisons were performed using the ART-C method with the Bonferroni correction for multiple comparisons, following the procedure outlined by Elkin *et al.* (2021).

All similarity ratings were also evaluated using a repeated measures two-way ART ANOVA. We decided to use this non-parametric test because our data included several extreme outliers, violated the Shapiro-Wilk normality test (Shapiro and Wilk, 1965) ($p < 0.05$ for several groups) and resulted in a significant Levene's test for homoscedascity ($p < 0.05$). Furthermore, the Q-Q plot shows mild kurtosis for several groups (Fig. 5.5). When the ART ANOVA indicated significant results, we performed post-hoc multiple comparison contrast tests with Bonferroni corrections using the method proposed by Elkin et al. (Elkin *et al.*, 2021). Additionally, we report the effect sizes in the form of partial eta-squared (η_p^2) for ANOVA tests and Cohen's d for t-tests. We use the benchmarks suggested by Cohen (Cohen, 1988) as small ($\eta_p^2 = 0.01; d = 0.2$), medium ($\eta_p^2 = 0.06; d = 0.5$), and large ($\eta_p^2 = 0.14; d = 0.8$). The ranking question presented as a part of the closing evaluation was analyzed using the pmr package in R (Lee and Yu, 2013).

5.4 Results: Benchmarking Tests

This section presents the results of the tests conducted to benchmark the algorithm.

5.4.1 Effects of Weighting Parameters

We use the dance sequence as a sample to show the effects of the weighting parameters. Fig. 5.6 shows how the weights impact the motion generated by the algorithm. This figure shows the errors that resulted from the optimization process, i.e., the errors that were commanded to the robot. Fig. 5.6A skeleton error is highest when α is almost 0, as this version of the algorithm almost exclusively prioritizes hand orientation error. Similarly, orientation error is highest when α is 1.00 (so β is 0) (Fig. 5.6B). The y-axes of these two figures also show the difference in ranges of both errors and hence emphasize the need for normalization. Fig. 5.6C shows the retargeting error and how it varies over time for a sequence of motions. This error tends to start on the higher side since the operator starts from rest or a neutral position. Hence, the robot needs to overcome gravity to catch up with the operator.

Effects on Skeleton Error

Fig. 5.7 depicts the commanded and actual skeleton errors. The changes in α (and hence, correspondingly, β) had a significant impact on the commanded ($F(3,13) = 19, p = 4.8 \times 10^{-5}, \eta_p^2 = 0.82$) and actual skeleton errors ($F(3,13) = 30, p = 4.5 \times 10^{-6}, \eta_p^2 = 0.87$).

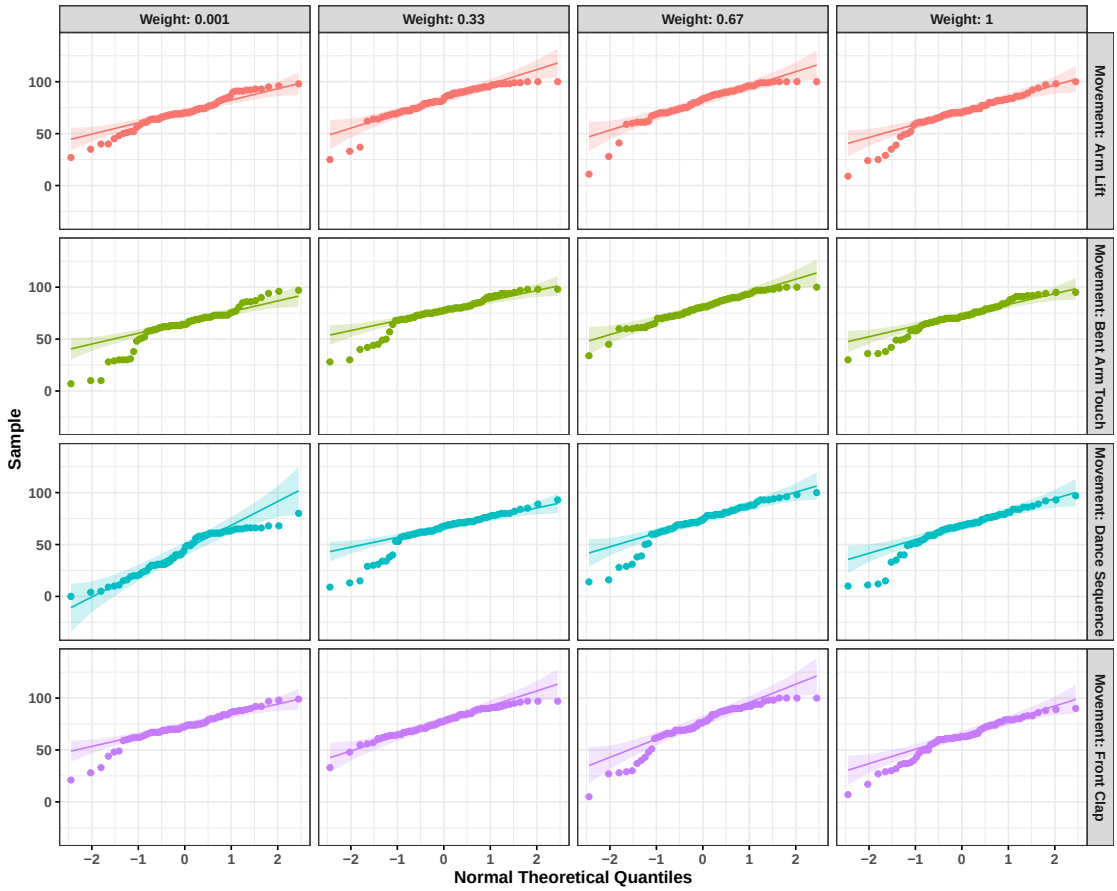


Figure 5.5: Q-Q plot for the similarity ratings shows kurtosis for almost all combinations of movements and weights. It is particularly severe for the dance sequence (row 3) and for bent arm touch (row 2).

For the commanded skeleton error, $\alpha = 0.001$ was significantly higher than $\alpha = 0.33$ ($t(13.008) = 4.3, p = 0.0057, d = 2.7$), $\alpha = 0.67$ ($t(13.008) = 7.4, p = 3.2 \times 10^{-5}, d = 4.7$) and $\alpha = 1$ ($t(13.008) = 5.1, p = 0.0011, d = 3.3$). Additionally, $\alpha = 0.33$ was significantly higher than $\alpha = 1$ ($t(13.008) = 3.1, p = 0.048, d = 2$). These significant differences were replicated in the same manner for the actual errors of the robot. Thus, $\alpha = 0.001$ was significantly higher than $\alpha = 0.33$ ($t(13.0004) = 5.2, p = 0.00094, d = 3.3$), $\alpha = 0.67$ ($t(13.0004) = 7.2, p = 4.2 \times 10^{-5}, d = 4.6$) and $\alpha = 1$ ($t(13.0004) = 8.8, p = 4.4 \times 10^{-6}, d = 5.6$). Similar to the commanded errors, $\alpha = 0.33$ was significantly more than $\alpha = 1$ ($t(13.0004) = 3.6, p = 0.019, d = 2.3$).

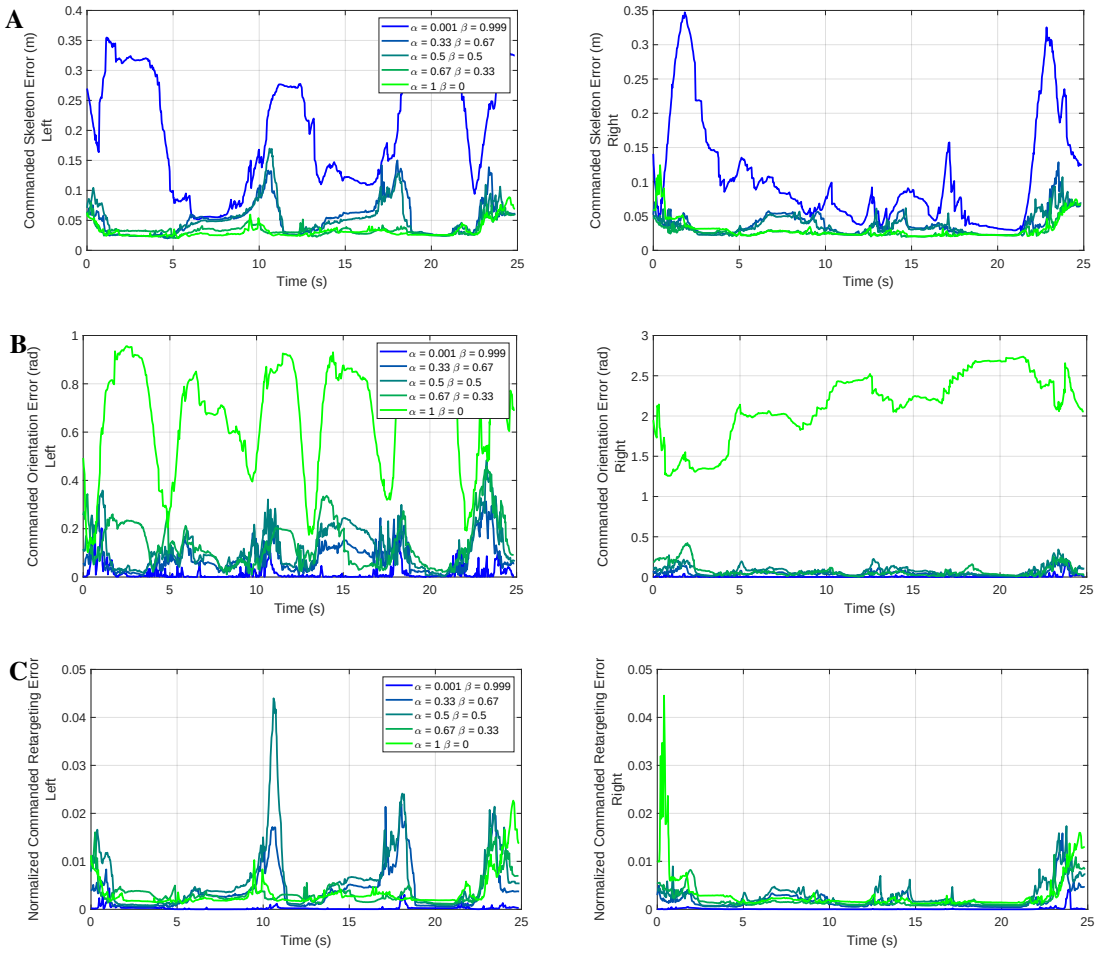


Figure 5.6: The various errors relevant to the algorithm. The blue to green gradient in these plots represents the increasing skeleton error weight and decreasing orientation error weight. The bright blue curve represents almost pure orientation error whereas the bright green represents pure skeleton error. **(A)** Commanded skeleton error for the four different weights of the dance sequence movement. The blue curve has the highest skeleton error whereas the green curve has a very low skeleton error. **(B)** Commanded orientation error for the four different weights of the dance sequence movement. The green curve has the highest orientation error. **(C)** Commanded retargeting error for the four different weights of the dance sequence movement. The retargeting error is lower in magnitude because of the normalization and the weights. The two bumps in the left arm's retargeting error are the result of collision with the robot's pedestal during the left arm's curvilinear motion.

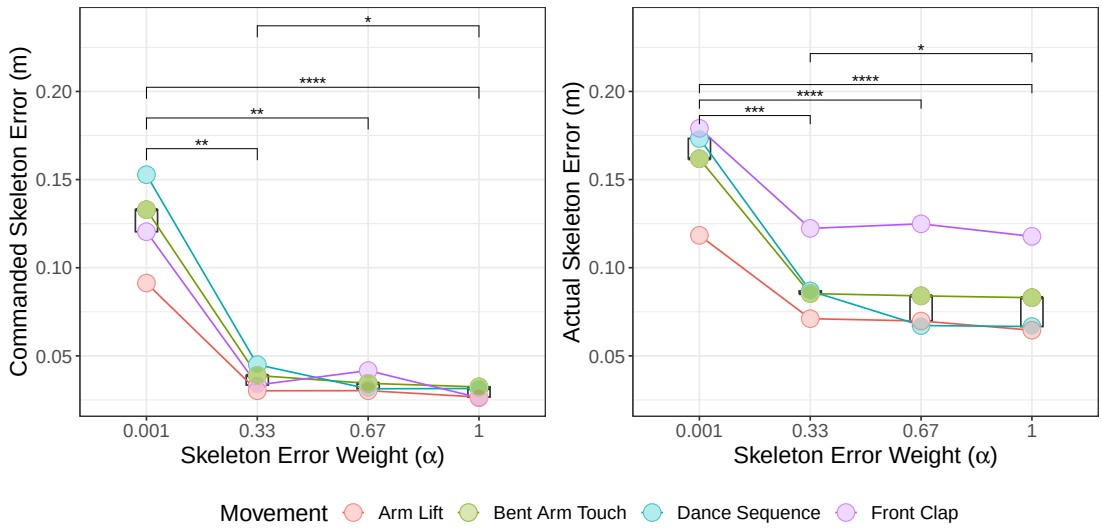


Figure 5.7: Skeleton error grand means across all motions and weight sets. The patterns of statistically significant differences are the same for both the commanded and actual errors.

Effects on Orientation Error

Fig. 5.8 shows the grand mean of the orientation error for the commanded joint angles and the actual joint angles. Once again, the ART ANOVA revealed a significant difference in the orientation error as a result of varying the weights for both the commanded errors $(F(3,13) = 100, p = 2.4 \times 10^{-9}, \eta_p^2 = 0.96)$ and the actual errors $(F(3,13) = 21, p = 3 \times 10^{-5}, \eta_p^2 = 0.83)$.

Interestingly, the post-hoc differences between the commanded and the actual were not the same for this error. All categories of the commanded errors were different from each other. $\alpha = 0.001$ was significantly lower than $\alpha = 0.33$ $(t(13.0.04) = -5.6, p = 0.00053, d = 3.5)$, $\alpha = 0.67$ $(t(13.0.04) = -11, p = 2.9 \times 10^{-07}, d = 7.1)$ and $\alpha = 1$ $(t(13.0.04) = -17, p = 2.1 \times 10^{-9}, d = 11)$. $\alpha = 0.33$ was significantly lower than $\alpha = 0.67$ $(t(13.0.04) = -5.6, p = 0.00053, d = 3.5)$, and $\alpha = 1$ $(t(13.0.04) = -11, p = 2.9 \times 10^{-7}, d = 7.1)$. Finally, $\alpha = 0.67$, was significantly lower than $\alpha = 1$ $(t(13.0.04) = -5.6, p = 0.00053, d = 3.5)$.

The only difference between the weight sets for the actual errors was that $\alpha = 0.001$ was significantly lower than $\alpha = 0.33$ $(t(13.002) = -7.6, p = 2.2 \times 10^{-5}, d = 4.8)$, $\alpha = 0.67$ $(t(13.002) = -5.5, p = 0.00066, d = 3.4)$ and $\alpha = 1$ $(t(13.002) = -5.1, p = 0.0012, d = 3.2)$.

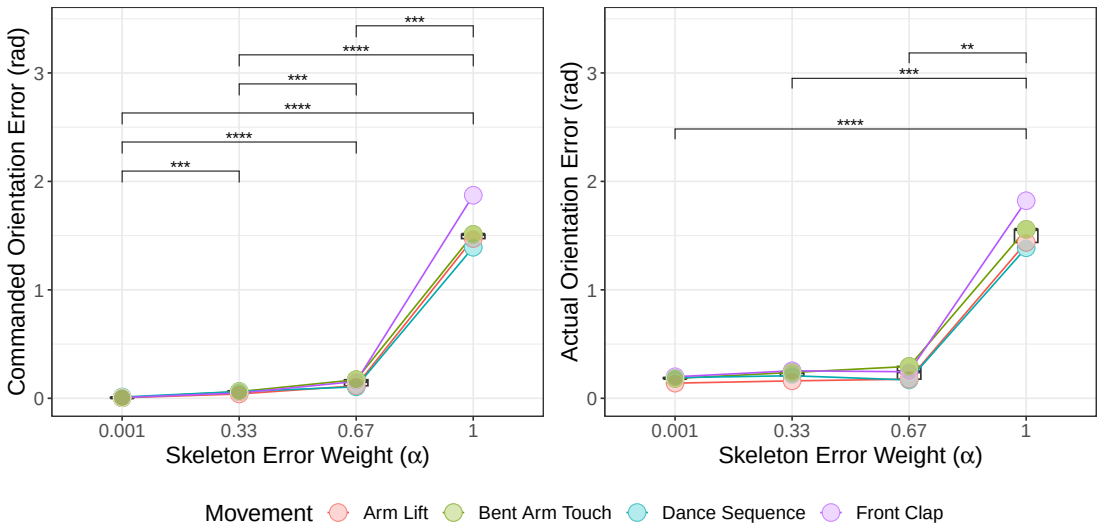


Figure 5.8: Orientation error grand means across all motions and weight sets. There were differences in the post-hoc comparison results between the commanded and actual errors.

Effects on Retargeting Error

Finally, similar to the shape and orientation errors, there was a significant main effect on the retargeting error for the commanded ($F(3,13) = 31, p = 3.2 \times 10^{-6}, \eta_p^2 = 0.88$) and actual ($F(3,13) = 23, p = 1.8 \times 10^{-5}, \eta_p^2 = 0.84$) errors (Fig. 5.9).

Post-hoc tests for both the commanded and actual errors revealed that $\alpha = 0.001$ was significantly lower than $\alpha = 0.33$ (commanded: $(t(13.02) = -3.6, p = 0.021, d = 2.2)$; actual: $(t(13.002) = -4.8, p = 0.0021, d = 3)$), $\alpha = 0.67$ (commanded: $(t(13.02) = -8.7, p = 5.3 \times 10^{-6}, d = 5.5)$; actual: $(t(13.002) = -6.3, p = 0.00016, d = 4)$) and finally $\alpha = 1$(commanded: $(t(13.02) = -7.5, p = 2.7 \times 10^{-5}, d = 4.7)$; actual: $(t(13.002) = -7.8, p = 1.7 \times 10^{-5}, d = 4.9)$). These differences are the same as those of the skeleton and orientation errors.

Post-hoc tests additionally reveal that $\alpha = 0.33$ is significantly lower than $\alpha = 0.67$ for the commanded errors $(t(13.02) = -5.1, p = 0.0011, d = 3.2)$. Finally, with $\alpha = 0.33$ was significantly lower than $\alpha = 1$ $(t(13.02) = -3.9, p = 0.01, d = 2.5)$ with a similar statistically trending effect for the actual error $(t(13.002) = -3.029, p = 0.06, d = 1.9)$.

5.4.2 Latency

Finally, we looked at the latency within the entire retargeting pipeline. Though there are significant differences between the weights (Fig. 5.10), the mean latencies are of more interest in this case. As, shown in the plot there are significant differences between

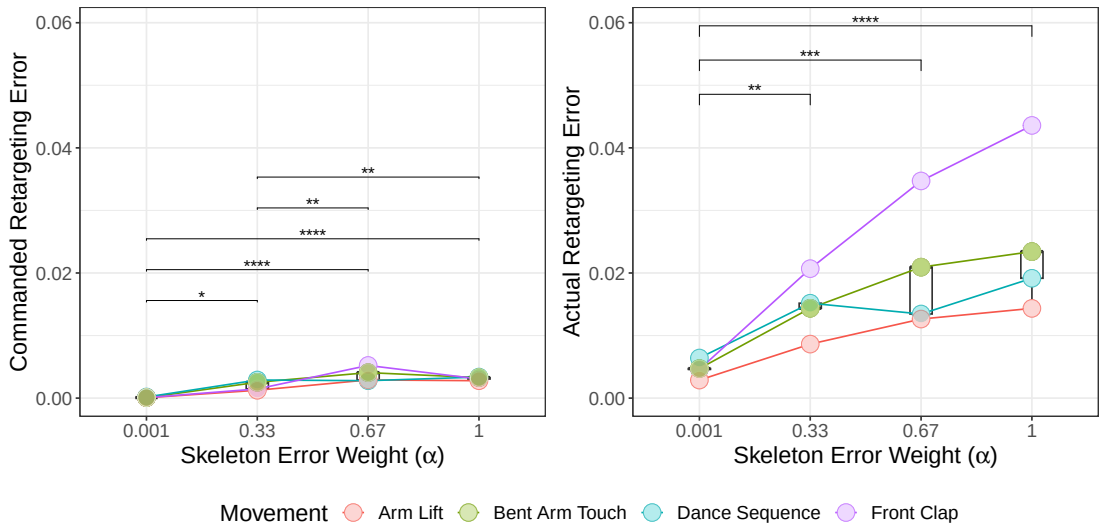


Figure 5.9: Retargeting error grand means across all motions and weight sets. This error is the weighted sum of the normalized skeleton and orientation errors. The significant errors reflect some of the differences in the skeleton and orientation errors.

the latencies of different weight combinations. However, these range between 0.034 s and 0.045 s. Across all movements and weight sets, there is a latency with mean, $M = 0.0378$ s and $SD = 0.00273$ s, which implies that the pipeline runs at a frequency of 26.46 Hz.

5.5 Results: Generalizability Experiment

The algorithm's functionality on Sawyer showed equivalence to its performance on Baxter, as depicted in Fig. 5.11. This section's analyses take a more comprehensive look at the error metrics, considering an additional robotic arm with an alternative configuration.

Fig. 5.12 reveals that the skeleton error trends, when combining both robots, mirror those of the benchmarking test conducted with Baxter (Fig. 5.12) with significant impacts of the weights ($F(3,29) = 15, p = 3.4 \times 10^{-6}, \eta_p^2 = 0.62$) and similar post-hoc trends. The commanded skeleton error for $\alpha = 0.001$ is significantly higher than $\alpha = 0.33$ ($t(29.06) = 4.5, p = 0.00059, d = 2.1$), $\alpha = 0.67$ ($t(29.06) = 5.4, p = 5.5 \times 10^{-5}, d = 2.5$) and $\alpha = 1$ ($t(29.06) = 6.6, p = 1.9 \times 10^{-6}, d = 3.1$).

The orientation error trends, as shown in Fig. 5.13, also follow a similar pattern to those in the same figure reference for both robots during benchmarking tests with Baxter (Fig. 5.13). Thus, the significant impact of the weights ($F(3,29) = 120, p = 3 \times 10^{-16}, \eta_p^2 = 0.92$) shows similar post-hoc trends. The commanded skeleton error for $\alpha = 0.001$ is significantly lower than all the others

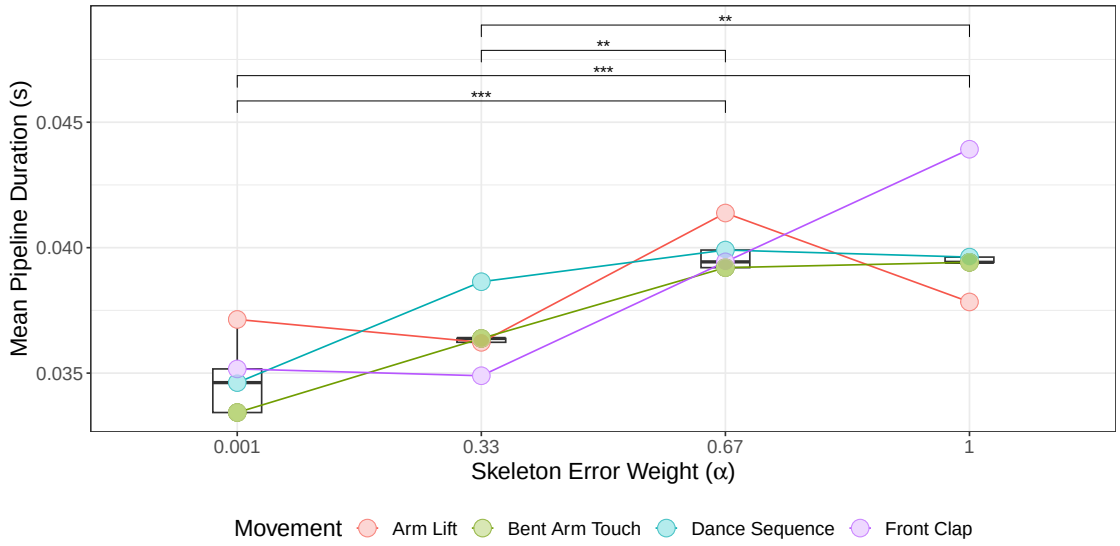


Figure 5.10: Pipeline duration grand means across all motions and weight sets. The higher the skeleton error component, the higher the latency, as skeleton error is more computationally intensive to compute than hand orientation error.

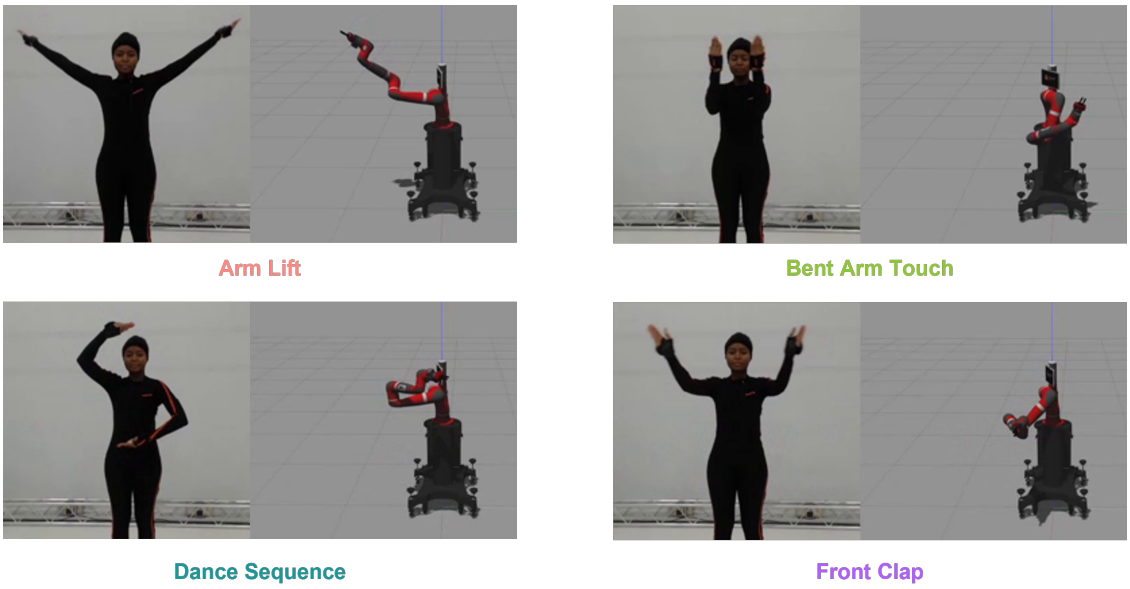


Figure 5.11: Screenshots of the simulated Sawyer robot performing the human's right-arm motion for the four motion sequences. All simulations have a skeleton error weight of $\alpha = 0.67$.

($\alpha = 0.33$: $t(29.06) = -6.5, p = 2.6 \times 10^{-6}, d = 3.1$; $\alpha = 0.67$ $t(29.06) = -12, p = 4.9 \times 10^{-12}, d = 5.7$; $\alpha = 1$ $t(29.06) = -19, p = 7.6 \times 10^{-17}, d = 8.7$). Additionally,

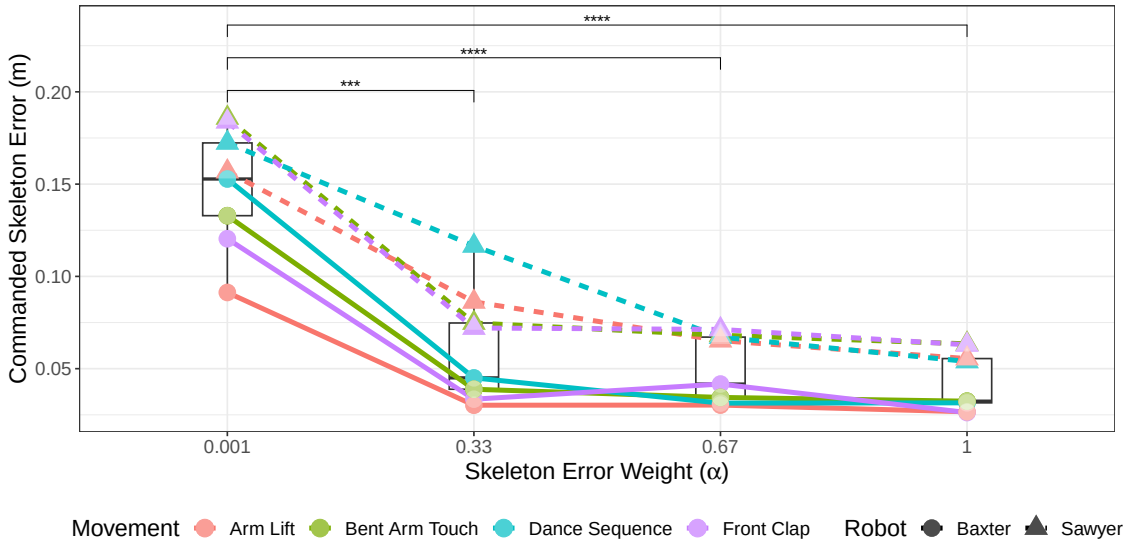


Figure 5.12: Skeleton errors for Baxter and Sawyer across all motions and weight sets. The solid line shows Baxter, and the dotted line shows Sawyer. Statistically significant differences are similar to those of the benchmarking tests. The skeleton error functions similarly even for a robot that is highly different from the human arm such as Sawyer.

$\alpha = 0.33$ is significantly lower than $\alpha = 0.67$ ($t(29.06) = -5.6, p = 3.2 \times 10^{-5}, d = 2.6$) and $\alpha = 1$ ($t(29.06) = -12, p = 4.9 \times 10^{-12}, d = 5.7$). Finally, $\alpha = 0.67$ is significantly lower than $\alpha = 1$ ($t(29.06) = -6.5, p = 2.6 \times 10^{-6}, d = 3.1$).

Finally, a substantial impact on the retargeting error by the weight sets is evidenced by the statistics provided: ($F(3,29) = 15, p = 3.7 \times 10^{-6}, \eta_p^2 = 0.61$). The retargeting error was significantly lower when there was little to no skeleton error component ($\alpha = 0.33$: $t(29.06) = -4.5, p = 0.00067, d = 2.1$; $\alpha = 0.67$: $t(29.06) = -6.2, p = 4.9 \times 10^{-6}, d = 2.9$; $\alpha = 0.1$: $t(29.06) = -5.9, p = 1.4 \times 10^{-5}, d = 2.8$).

5.6 Results: User Study

To examine our hypotheses and user perspectives on robot-generated motion, we break the results into three sections. The first section looks at user perception of movement generated by OCRA. The second section aims to elucidate the reasoning behind participant ratings. The third analyzes the relationship between the errors and participant ratings.

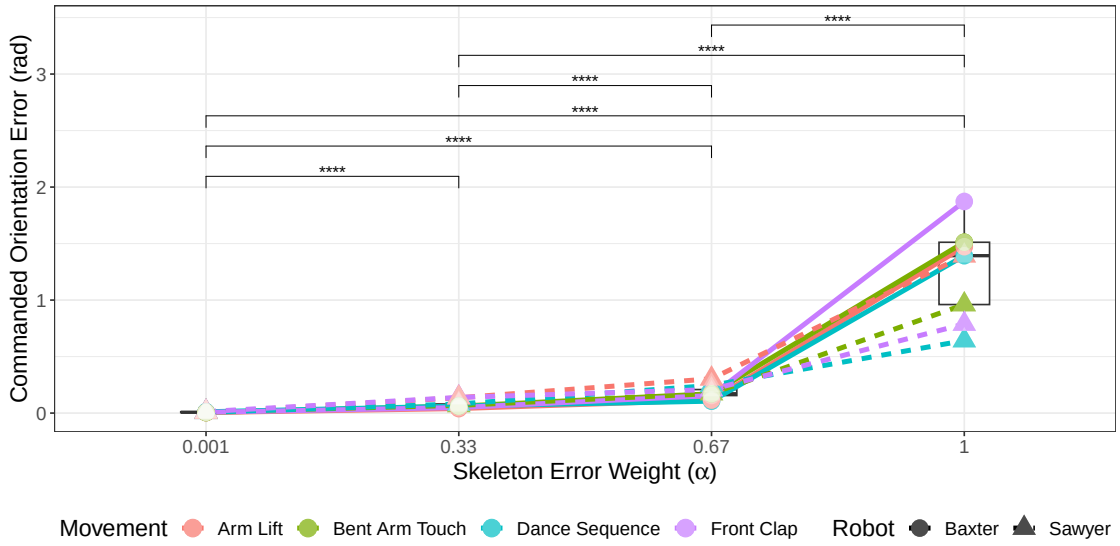


Figure 5.13: Hand orientation errors for Baxter and Sawyer across all motions and weight sets. The solid lines show Baxter, and the dotted lines show Sawyer

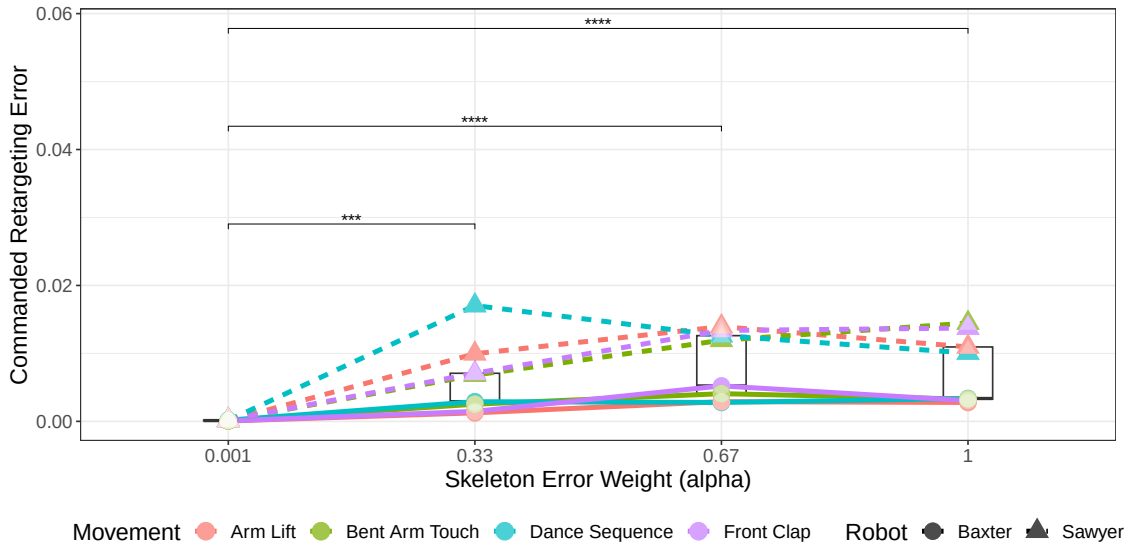


Figure 5.14: Retargeting errors for Baxter and Sawyer across all motions and weight sets. The solid lines show Baxter, and the dotted lines show Sawyer

5.6.1 User perception of motion

Participants provided ratings on the perceived similarity between the robot's motion and that of the operator as seen in Fig 5.15A. The ART ANOVA revealed that there was a significant interaction effect between the weighting of the errors and the movement

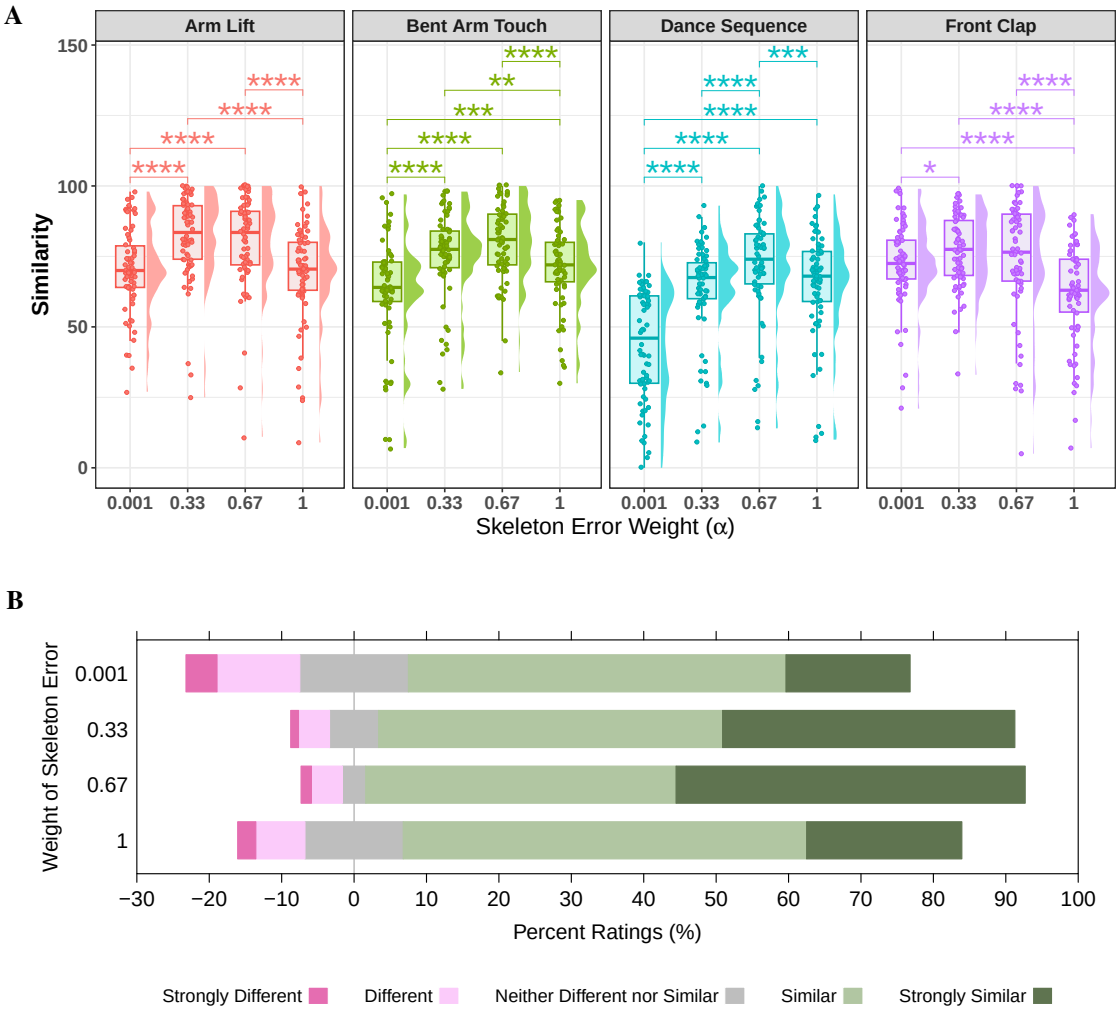


Figure 5.15: **(A)** Rain cloud plots of participant ratings for the movements shown in the user study, each with four weights for skeleton error. Significant pairwise differences between medians (the central bars in the boxes) are indicated by lines with stars: Significance notation: $\star$ $p < 0.05$, $\star\star$ $p < 0.01$,$\star\star\star$ $p < 0.001$, and $\star\star\star\star$ $p < 0.0001$.
(B) Participant similarity ratings presented as a stacked bar plot. This layout shows that a large percentage of participants found the robot motion to be strongly similar to that of the human operator.

$(F(9,1035) = 17, p = 5.1 \times 10^{-26}, \eta_p^2 = 0.13)$. Upon conducting post-hoc tests, we identified that the simple main effects of the weights were significant for each of the four movement types (arm lift: $(F(3,207) = 30, p = 2.8 \times 10^{-16}, \eta_p^2 = 0.31)$, bent arm touch: $(F(3,207) = 34, p = 4.5 \times 10^{-18}, \eta_p^2 = 0.33)$, dance sequence: $(F(3,207) = 72, p = 5.1 \times 10^{-32}, \eta_p^2 = 0.51)$, front clap: $(F(3,207) = 26, p = 1.5 \times 10^{e-14}, \eta_p^2 = 0.28)$).

Pair-wise comparisons revealed that there is a significant difference between a skeleton

error weight of 0.001 and 0.33 across all movement types (arm lift: $(t(207) = -7.2, p = 7.8 \times 10^{-11}, d = 1.2)$, bent arm touch: $(t(207) = -7.6, p = 7.7 \times 10^{-12}, d = 1.3)$, dance sequence: $(t(207) = -9.3, p = 1.6 \times 10^{-16}, d = 1.6)$, front clap:$(t(207) = -2.7, p = 0.042, d = 0.46)$). Similarly, a skeleton error weight of 0.67 and 1 are also significantly different across all movement types (arm lift: $(t(207) = 6.2, p = 1.4 \times 10^{-08}, d = 1.1)$, bent arm touch: $(t(207) = 5.2, p = 3.6 \times 10^{-6}, d = 0.87)$, dance sequence: $(t(207) = 3.9, p = 0.00088, d = 0.65)$, front clap: $(t(207) = 7.4, p = 2.2 \times 10^{-11}, d = 1.2)$).

Participants were able to differentiate between a skeleton error weight of 0.001 and 0.67 only for arm lift $(t(207) = -6.3, p = 1.3 \times 10^{-8}, d = 1.1)$, bent arm touch $(t(207) = -9.5, p = 3.8 \times 10^{-17}, d = 1.6)$ and dance sequence $(t(207) = -14, p = 1.4 \times 10^{-31}, d = 2.4)$. Additionally, participants could distinguish between skeleton error weights of 0.33 and 1 for arm lift $(t(207) = 7.2, p = 8.3 \times 10^{-11}, d = 1.2)$, bent arm touch $(t(207) = 3.2, p = 0.0084, d = 0.55)$, and front clap $(t(207) = 8, p = 4.8 \times 10^{-13}, d = 1.4)$. Participants also found 0.001 to be significantly different from 1 for bent arm touch $(t(207) = -4.3, p = 0.00014, d = 0.73)$, dance sequence $(t(207) = -10, p = 1.3 \times 10^{-19}, d = 1.7)$ and front clap $(t(207) = 5.3, p = 1.8 \times 10^{-6}, d = 0.89)$. Interestingly, the weights of 0.33 and 0.67 were significantly different only for the most complex motion, the dance sequence $((t(207) = -4.9, p = 1.1 \times 10^{-5}, d = 0.83))$.

Since we utilized a banded-type visual analog scale, we also examined these results from a different lens using a Likert style analysis (Fig 5.15B). When the skeleton error weight was 0.33, 87.86% of the ratings were either similar or strongly similar. This percentage increased to 91.07% when the weight was 0.67.

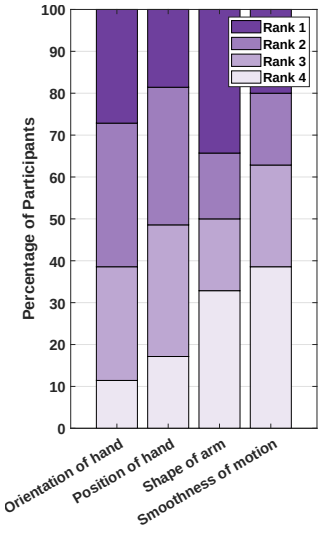


Figure 5.16: Participant rankings of rating criteria

	Orientation of hand	Position of hand	Shape of arm	Smooth motion
Orientation of hand	0	38	41	45
Position of hand	32	0	34	41
Shape of arm	29	36	0	41
Smooth motion	25	29	29	0

Table 5.1: Pairwise frequencies of ranking data

5.6.2 Insight into user ratings

We presented participants with four criteria that we deemed important and asked them to rank these criteria based on their rating system. A total of 19 different ranking combinations were chosen by the participants out of 24 possible combinations. Upon analyzing this data, we found that the most commonly chosen ranking order (selected by 8 participants) is shape of arm, position of hand, orientation of hand and finally smoothness. We then calculated the mean rank of all four options. The mean rank is the average rank of a particular option, weighted by how often an option is at that rank. Orientation of the hand has the highest rank (2.23), followed by position of the hand (2.47), shape of the arm (2.49), and smoothness of the motion (2.81). There is not much difference between the ranks of orientation of the hand and shape of the arm. However, users seem to have considered the smoothness to be the least important. The percentage of marginal frequencies showing how often a particular item was ranked at a particular position can be seen in Fig. 5.16. The pairwise frequencies showing how often an item was ranked above another can be seen in Table 5.1

We also gave participants the opportunity to suggest an additional item for the ranking list. These suggestions varied between participants. A majority of these comments ($N = 13$) were about the orientation or position of the joints of the robot's arm. Synchronicity of the robot's movement to that of the human and speed of the robot motion were both mentioned $N = 6$ times. Participants assigned varying levels of importance to items in different categories, leading to a lack of consistency in their rankings. Other items mentioned were relative symmetry between left and right arms ($N = 1$), operator intention ($N = 1$) and a slimmer physique for the robot ($N = 1$). The remaining participants did not have any additional feedback for this question.

5.6.3 Relation between errors and participant ratings

We were curious if there was a quantitative relationship between participant ratings and the errors themselves. So, we fit a polynomial curve to the similarity rating (S_p) data using MATLAB's fit function, which uses the least squares method to perform the fit. The equation below represents the polynomial model.

$$S_p = \zeta_0 + \zeta_1 \varepsilon_s + \zeta_2 \varepsilon_o \tag{5.17}$$

where ζ represents the coefficients of the polynomial.

This fit was performed separately for each movement type since there was a significant difference in participant ratings for each move, as seen in Fig. 5.17. The actual equations and the goodness of fit measures can be seen in Table 5.2. The sum of squares due to the error measures the total deviation of the similarity ratings from the curve fit to the similarity ratings. This value is relatively low for all movement types. The R^2 and the adjusted degrees of freedom (adj. R^2) are close to 1, indicating that a greater proportion

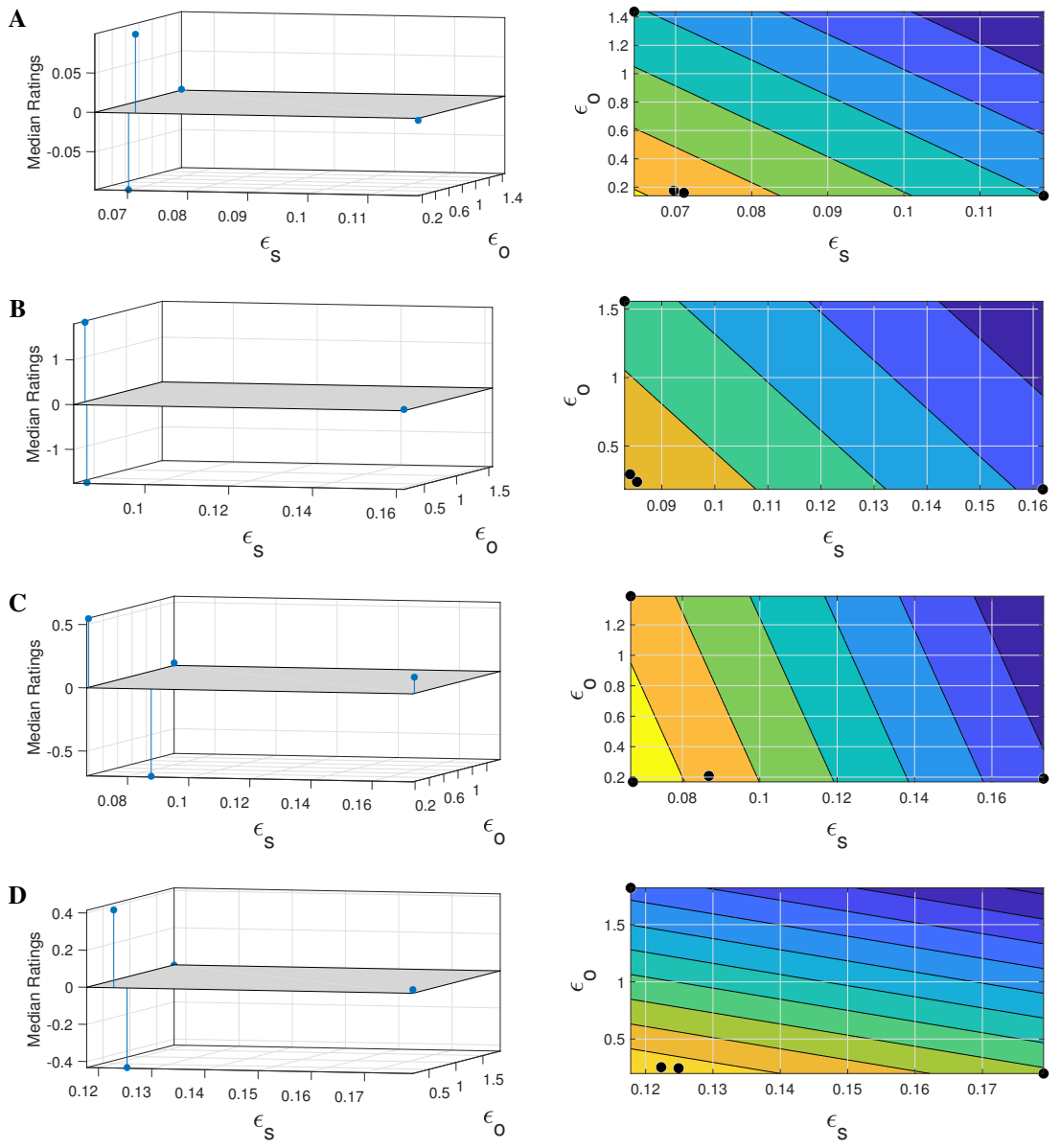


Figure 5.17: **(A)** Residual and contour plot for the arm lift motion. **(B)** Residual and contour plot for the bent arm touch motion. **(C)** Residual and contour plot for the dance sequence motion. **(D)** Residual and contour plot for the front clap motion.
Note: Axis limits of the plots differ between different movement types to enhance visibility.

of the variance is accounted for by the polynomial fit model. The root mean squared error is also relatively low, showing that the standard error of the regression is relatively low. Interestingly, bent arm touch has a higher SSE and RMSE compared to the other movements.

Movement	Equation	SSE	R^2	DoF	adj. R^2	RMSE
Arm Lift	$105.77 - 288.51 * \varepsilon_s - 11.58 * \varepsilon_o$	0.0196	0.9999	1	0.9997	0.140
Bent Arm Touch	$98.05 - 203.98 * \varepsilon_s - 5.81 * \varepsilon_o$	6.3251	0.9616	1	0.8848	2.515
Dance Sequence	$91.64 - 258.92 * \varepsilon_s - 4.61 * \varepsilon_o$	0.7996	0.9982	1	0.9947	0.8942
Front Clap	$90.44 - 90.07 * \varepsilon_s - 9.24 * \varepsilon_o$	0.3593	0.9973	1	0.9918	0.5994

Table 5.2: Curve-fitting equations associated with the fits in Fig 5.17, along with the associated goodness of fit statistics. SSE is the sum of squares errors due to the fit.

5.7 Discussion

Although many retargeting algorithms exist, only a few are generalizable and even fewer have undergone rigorous validation Darvish *et al.* (2023). We aimed to develop an algorithm that produces human-like motions, serving as a valuable tool for human-robot interaction researchers and enhancing teleoperation systems for novice users.

Assessing the Algorithm

Results from the user study can be seen in Fig. 5.15. Participants rated conditions having a substantial combination of both arm skeleton and hand orientation error to be significantly more similar to the human motion than the tested conditions with purely orientation or purely skeleton error. User comments also demonstrate that participants cared most about the orientation of the robot's hand and the shape of its arm skeleton, rather than the position of the hand or the smoothness of the motion. These observations support our first hypothesis, **H1**. Thus, skeleton error and orientation error can indeed be considered suitable ingredients for creating human-like robot motion from motion-capture data.

Contrarily, our second hypothesis (**H2**) was not conclusively validated by our data. The variance in participant ratings across the two presented motions, most notably seen in the complex dance sequence, highlights the intricate relationship between motion complexity and error thresholds. Our data points to the possibility that the complexity of motions affects how users perceive error, emphasizing the need to consider motion complexity in future studies on motion retargeting.

Analyzing errors, after a threshold, skeleton error ceased to enhance the generated motion (Figures 5.7 and 5.12). Participants could not distinguish between the intermediate skeleton error weights unless the motion sequence was long ad complex such as the dance sequence movement. Additionally, the skeleton error's versatility is evident from consistent performance on the Sawyer robot (Figures 5.11 and 5.12).

While altering weight combinations influenced orientation error significantly (Fig-

ures 5.8 and 5.13), the retargeting error mirrored shape error trends (Figures 5.9 and 5.14). However, participants were consistently able to discern differences between the extreme and intermediate weight sets. In general, we recommend operating the algorithm within the range of $0.3 < \alpha < 0.67$, prioritizing the orientation as required.

Rigorous Validation

A notable contribution of this work is the robust validation performed to benchmark and evaluate OCRA. Despite the extensive history of teleoperation research in robotics, spanning several decades (Niemeyer *et al.*, 2016; Darvish *et al.*, 2023), the field suffers from a significant deficiency in standardized benchmarking procedures. Many teleoperation algorithms, especially those crafted for humanoid robots, are assessed using a lone operator (Stanton *et al.*, 2012; Zhao *et al.*, 2016; Kaplish and Yamane, 2019; Schultz *et al.*, 2017). This approach often lacks the depth and breadth of extensive user studies, resulting in potential biases and limited insights.

To augment the benchmarking framework detailed in this research, we identify a few additional steps that can offer a more comprehensive analysis of OCRA. Specifically, we recommend employing three objective measures: Procrustes distance (Goodall, 1991), for assessing the congruence of the retargeted pose; root mean square error (RMSE) between the robot and human end-effector positions, to gauge the algorithm's capability to execute particular tasks (Tosun *et al.*, 2014; Darvish *et al.*, 2019); and a smoothness metric (Wang *et al.*, 2018), designed to illustrate the transitions between motions. Collectively, these metrics will provide a more nuanced and reliable evaluation of the algorithm, aiding future development and application.

Following this additional benchmarking, the next step for OCRA will be to conduct a detailed quantitative user study to determine how operators perceive the algorithm. We propose the use of embodiment metrics such as peri-personal space (Sierotowicz *et al.*, 2020) and the methodology proposed by Alzayat *et al.* (2019).

Implementation and Additional Features

The algorithm is implemented in Python 3.8 and has been rigorously tested on Ubuntu 20.04 with ROS Noetic (Open Robotics, 2022). This implementation includes a specific ROS package, designed with a custom class containing robot-specific functions and parameters. Making the code compatible with a different robot requires only modification of this class to link to the chosen robot's Unified Robotics Description Format (URDF) file and related parameters.

Furthermore, the teleoperation interface comes with a voice command feature, utilizing a set of pre-defined commands powered by the PocketSphinx library (Huggins-Daines *et al.*, 2006). The voice commands allow users to initiate or halt teleoperation, switch among four different pre-set weight configurations, toggle between a mirrored mode (where the robot mirrors the operator) and a direct mode (where each arm of the

operator is mapped to the corresponding arm on the robot, e.g., left arm to left arm and right arm to right arm), and change between two motion capture systems (Xsens and Captury). This voice-enabled functionality frees the operator's hands for an intuitive and efficient teleoperation experience.

Emotion recognition is another added capability, using Baxter's face screen and the Affectiva AFFDEX algorithm (McDuff *et al.*, 2016). This allows the user to display one of seven pre-defined emotions (anger, disgust, fear, happiness, sadness, contempt and surprise) with the robot. This feature has been implemented based on expert feedback, based on their interest in leveraging the robot's emotion teleoperation features for therapy.

Finally, this extended version of the teleoperation system will be utilized in our upcoming user study.

Future Outlook

Furthermore, there were practical concerns with the implementation of the algorithm that should be considered for future work. For example, there are several manual inputs required by the algorithm such as the initialization pose, choice of translation point, and the quaternions for source frame to world frame transformation. Automating these aspects could help increase the usability and robustness of the algorithm, which would be beneficial for roboticists wishing to use this approach as a starting point for teleoperating their own robots.

It is also worth noting that the lack of publicly available code for motion-retargeting algorithms is a substantial barrier to advancement in the field. Greater transparency in methodology and implementation would facilitate replication studies and enhance the overall knowledge base on kinematic retargeting. We plan to release our code after publication of this chapter in a conference paper or journal article.

Our algorithm has been tested and confirmed to work with serial link manipulators with revolute joints. It is uncertain, however, if the same effectiveness can be achieved with parallel manipulators or with prismatic joints. Future studies should evaluate the performance of the algorithm on different types of manipulators, as the current algorithm might not be universally applicable.

In terms of latency, OCRA's performance is in line with previous studies that have reported latencies ranging from 30 ms (Koenemann *et al.*, 2014) to 0.005 ms (Darvish *et al.*, 2019). It is crucial to remember that although high latencies can disrupt the smoothness of the robot's motion, a balance must be struck between achieving real-time function and maintaining the similarity to human motion. Future advancements should aim to lower these latencies to increase the real-time applicability of such algorithms.

Finally, the error metrics presented in our study can also serve as evaluation tools for other retargeting algorithms. By quantifying both hand orientation and arm skeleton error, we offer a more comprehensive measure of motion similarity. This dual approach to

error measurement could help other researchers refine and enhance their own kinematic retargeting methods.

In conclusion, this work represents a significant step forward in kinematic retargeting, demonstrating the crucial role of both skeleton and orientation error in achieving human-like motion. Despite the limitations and challenges highlighted, the algorithm presents a promising foundation for future research, offering a fresh perspective and improved tools for achieving high-quality motion retargeting.

5.8 Conclusion

OCRA offers a versatile solution to transpose human arm motion to robot arms. Perceptual tests suggest our algorithm convincingly replicates human-like motion for optimal weight values. Designed primarily for robots with revolute joints, modifications might be needed for prismatic joints. We intend to launch OCRA as a ROS package with the associated future publication, promoting broader accessibility. Our immediate plan is a real-time usability study for this teleoperation system.

Chapter 6

Conclusion

While socially assistive robots have demonstrated their potential as exercise coaches (Matarić and Scassellati, 2016), their applicability in real-world settings, where adaptability to evolving user needs is critical, is yet to be realized. These robots predominantly employ pre-programmed gesture-based interactions, mainly as demonstrations (Pulido *et al.*, 2019; Matsusaka *et al.*, 2009; Lewis *et al.*, 2016), to instruct exercises to users. Moreover, in these situations, the robots need to operate as self-sufficient standalone systems, requiring minimal supervision, even when interactions become complex or sophisticated. Nonetheless, the design of robot movements and responses can be most effectively carried out by exercise therapy experts, given their deep understanding of their patients' needs and therapy goals (Tanguy *et al.*, 2016; Görer *et al.*, 2013).

Therefore, this thesis investigates dynamic gesture-based interactions in minimally supervised scenarios, focusing on facilitating such interactions with a robot, and incorporating experts in the development of such behavior. This section briefly summarizes the contributions of this thesis, revisits the two overall research questions posed in Chapter 1, and discusses the implications of this work as well as future avenues of research.

6.1 Summary of Contributions

This section briefly summarizes the contributions of the four main chapters of this thesis:

6.1.1 Robot Interaction Studio: An Environment for Minimally Supervised HRI

Within Chapter 2, we showcase a minimally supervised environment that facilitates the measurement of complex gesture-based interactions with an exercise-oriented robot. This objective is achieved by merging Captury Live, a markerless motion-capture system operating in real-time, and a Rethink Robotics Baxter Research Robot, culminating in the formation of the Robot Interaction Studio. We evaluated this platform for minimally supervised human-robot interaction (HRI) through a 75-minute-long user study with seven adults who were given minimal instructions and no feedback about their actions. The robot used sounds, facial expressions, facial colors, head motions, and arm motions to

sequentially present three categories of cues in randomized order while constantly rotating its face screen to look at the user. Analysis of the captured user motions shows that the cue type significantly affected the distance subjects traveled and the amount of time they spent within the robot's reachable workspace, in alignment with the design of the cues.

6.1.2 Formative and Summative Feedback in Minimally Supervised HRI

When teaching or coaching, humans augment their words with carefully timed hand gestures, head and body movements, and facial expressions to provide feedback to their students, but robots rarely leverage these complementary cues. A minimally supervised social robot with such skills could help people exercise, do physical therapy, and learn new activities. Thus, in Chapter 3, we leverage the rich quantitative data obtained from the Robot Interaction Studio to propose and evaluate a methodology for designing gesture-based feedback in minimally supervised scenarios. We investigated how nonverbal feedback provided by an upper-body humanoid robot affects human behavior with only nominal instructions. Inspired by the education literature, we evaluated formative feedback (real-time corrections) and summative feedback (a final score after task completion) both alone and in combination for three distinct tasks: standing in a particular location in the room, mimicking the robot's arm pose, and contacting the robot's hands.

Including the seven participants who received no feedback during the validation study from Chapter 2, twenty-eight adults participated in this mixed-design study, attempting to follow the robot's cues for seventy-five trials with no experimenter help. Groups of seven participants received either no feedback (fs), only formative feedback (Fs), only summative feedback (fS), or both formative and summative feedback (FS) from the robot. Detailed statistical analysis of markerless motion-capture data indicates that both formative and summative feedback from the robot significantly aided user performance, as hypothesized. Additionally, grading of post-study task explanations shows that formative feedback (but not summative feedback) improved task understanding. Detailed analysis of participant activity via metrics and heatmaps further elucidated the complementary positive effects of these two kinds of feedback. These results should encourage roboticists to leverage nonverbal cues based on the user's movement and view feedback though formative and summative lenses when designing social robots.

6.1.3 Expert Perception of Exercise Robots

Chapter 4 presents a user study comprising semi-structured interviews with eight exercise therapy experts, probing their views on social exercise robots and teleoperation. Despite varied initial perspectives on social robots, a consensus favoring teleoperation emerged. This shared interest in teleoperation led to enhanced attitudes towards robots,

marked by increased appreciation for the robot's presence, stronger emotional attachment, and heightened desire to cooperate among the experts. There was a preference for functionalities related to communication and arm movement, suggesting that interactive and dynamic features are pivotal for a robot's effectiveness in exercise scenarios. The study underscores the potential of teleoperation in shaping exercise gestures for robots, pinpointing a promising pathway for future research.

6.1.4 OCRA: An Optimization-based Customizable Retargeting Algorithm for Teleoperation

Encouraged by the positive views that exercise therapy experts held towards teleoperation, we developed an optimization-based retargeting algorithm, OCRA, detailed in Chapter 5, with the goal of creating human-like robot motion. OCRA maps between two kinematically dissimilar serial linkages by minimizing a retargeting error, which is a combination of an end-effector orientation error and an arm pose-based error, termed the skeleton error. Experiments demonstrate OCRA's effectiveness in terms of performance, latency, and generalizability. A video-watching study with 70 participants demonstrated that when this algorithm considers both the hand orientation and the arm skeleton, it creates robot arm motions that users perceive to be highly similar to those of the human operator. These positive results indicate that OCRA would be highly suitable for telerobotics and telepresence through avatars.

6.2 Revisiting Overall Research Questions

ORQ 1 posed the question, *"What strategies can be employed by social exercise robots to effectively leverage gesture-based interactions within minimally supervised settings?"* The research reported in this dissertation demonstrates that within minimally supervised settings, markerless motion capture can be effectively used to initiate interactions between a user and a robot, eliminating the need for additional setup involving physical sensors. This approach yields extensive quantitative information that robots can utilize to generate dynamic gesture-based interactions. In these environments, it's crucial that the robot can operate effectively without the presence of a human expert, necessitating seamless interaction with the user through intuitive feedback. Our research also reveals that drawing inspiration from the educational literature can provide a robust foundation for a feedback strategy, which robots can use to support users smoothly and consistently. Robot designers should consider both formative and summative feedback when crafting their feedback systems to ensure intuitiveness for users.

ORQ 2 asked, *"What methods can be implemented to streamline the development process of gesture-based interactions?"* Given that experts understand their clients' needs best, we aimed to involve them in the process of generating interactions. We proposed that teleoperation could be a viable method to achieve this goal. To verify this approach,

we sought expert opinions to gauge their interest in such a system. The enthusiasm they demonstrated suggests that the combination of teleoperation of social robots with an intuitive algorithm like OCRA could be used to effectively generate these motions. We conducted comprehensive tests on OCRA, confirming its ability to intuitively generate human-like motions through quantitative evidence.

6.3 Implications and Future Outlook

We propose a system that integrates OCRA's functionality into the Robot Interaction Studio, as shown in Fig 6.1. The operator wears an inertial motion-capture suit that can measure their detailed movements for robot control. Beyond arm and hand motions, the proposed future platform also includes a face teleoperation system that uses an emotion-recognition algorithm such as Affectiva AFFDEX (McDuff *et al.*, 2016) to cause the robot to show the operator's facial expression in real time. Additionally, we propose incorporating naturalistic haptic feedback of the high-frequency vibrations of the robot's two wrists to the respective hands of the operator, as previously done for a surgical teleoperation system (McMahan *et al.*, 2011), to give the operator better awareness of the robot's motions and any physical contacts it makes. The operator receives audiovisual feedback via a monitor and speakers; multiple camera views are shown, including a video feed from a camera on the head of the robot, which enables the operator to directly look at their interaction partner. Additionally, the user can wear physiological sensors such as a heart rate sensor. Such a system can enable several avenues of research, as described in the subsequent sections.

6.3.1 Exercise Robots in Real-World Facilities

The Robot Interaction Studio can serve as an innovative addition to gymnasiums or physical therapy facilities, transforming them into hubs for community rehabilitation. Users can engage in interactive sessions with the robot through this innovative system. The data gathered during these interactions can be used to create detailed reports on user performance metrics, on both a daily and a long-term basis. Such information would be invaluable in helping users exercise regularly and achieve their therapy goals. Furthermore, this setup would allow experts such as physical therapists or fitness trainers to use OCRA to periodically adjust robot behaviors without the painful expert programming and iterative adjustments normally required, as noted by (Fitter *et al.*, 2020). Moreover, the platform's versatility facilitates easy replacement of Baxter with other robots. This flexibility implies strong potential for facilitating more dynamic lower-limb rehabilitation or interactions.

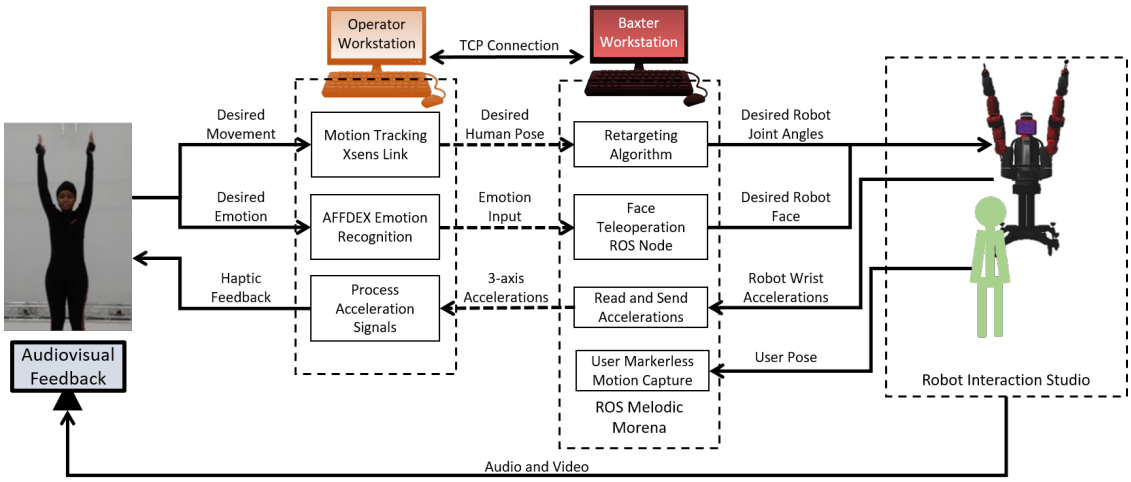


Figure 6.1: System architecture. The operator workstation handles Xsens motion tracking and runs the AFFDEX algorithm for facial expression tracking. The robot workstation is ROS compatible and is mainly responsible for the kinematic retargeting algorithm and markerless motion capture of the user. Visual, audio, and haptic data also flow from the robot-user interaction to the operator. The two systems exchange data via a TCP (Transmission Control Protocol) connection.

6.3.2 Designing Interactions

As noted in Chapter 5, OCRA could be utilized to record and generate generalizable interactions using learning from demonstration (LfD) techniques. However, the system proposed in Fig. 6.1 offers the added advantage of quantifying the interaction itself. This capability suggests that teleoperation with this system can be employed to create more than just simple movements; it can generate dynamic physical contact with a user. These interactions could range from straightforward actions like a high five, akin to the concept in contact cue, to more complex interactions such as hugging (Kaplish and Yamane, 2019). Moreover, the application of dual motion-capture systems further enables the detailed study of human-robot handovers (Ortenzi *et al.*, 2021) and their creation. Finally, the rich data of all these different types of interactions can then be used to create datasets that can be used by the community.

6.3.3 Implications for the Field of HRI

The field of human-robot interaction (HRI) currently grapples with a significant reproducibility crisis, compounded by an overreliance on subjective measures, a dearth of quantitative data, and insufficient reporting (Zimmerman *et al.*, 2022). Our system's capacity to offer detailed quantitative information paves the way for a paradigm shift toward quantitative, reproducible studies. Furthermore, despite the prevalent use of tele-

operation to analyze user behavior in HRI (Elbeleidy *et al.*, 2022), especially within Wizard-of-Oz studies (Riek, 2012; Martelaro, 2016), most of these require operators to switch between predefined sequences. Our system enables HRI researchers to transcend pre-programmed gestures, dynamically adapt to situations, and obtain quantitative user responses.

By leveraging systems like the ones explored in this dissertation, researchers can gather quantitative data from user studies that remain unaffected by the Hawthorne effect (the alteration of behavior due to the awareness of being observed) (Kachouie *et al.*, 2014). A promising research direction involves a deeper exploration of participant eye gaze, going beyond attention metrics based on head position and orientation. Fine-grained measures such as blink rate, re-glance frequency, and glance duration can enrich our understanding of user eye gaze and aid the robot in determining user attentiveness. Another avenue worth pursuing is the study of human-robot proxemics (Samarakoon *et al.*, 2022).

In conclusion, this thesis paves the way for a multitude of research opportunities in HRI. The potential of rich quantitative data and intuitive teleoperation remains to be fully harnessed and explored for social robots in exercise settings and beyond.

Appendix A

Supplementary Materials

A.1 Feedback Study

This section provides supplementary materials for the user study on formative and summative feedback reported in Chapter 3. The methodological details are also relevant for the validation study of Chapter 2, as those seven users constitute the fs group in this study.

A.1.1 Experimenter Script

The experimenter read the following script to the participant at the start of the study:

> This study is designed to understand how humans react to different robot behaviors during a movement-focused activity session with a humanoid robot. In each trial of the study, the robot will present you with a command communicated using movements of the robot's arms and head, facial expressions on the robot's screen, and sounds. You will have a set amount of time to complete the desired task communicated by the robot. Once the time has run out, you will move on to the next trial, and the robot will present a new command.
>
> We ask that you please stay in front of the marked line (*experimenter pointed at tape line on the floor*) and do not stick your hand or arm into the joints of the robot (*experimenter pointed out Baxter's joints*). You can ignore the four bits of tape on the floor. In order to complete the desired task, you may need to move throughout the room and make contact with the robot.
>
> We will be collecting both video and audio recordings, as well as heart rate data. Once you have put the heart rate sensor on your left upper arm, we ask that you move like this (*experimenter showed arm movements desired for Captury Live calibration*) so that we can confirm that the necessary data is being collected. The trials are broken into three blocks, and you will get a break between blocks. After each of the breaks, we will ask you to perform these similar movements to confirm that the necessary data is still being collected.
>
> After you have completed all three blocks of the study, you will complete a set of surveys about your experiences interacting with the robot.

> If at any point you would like to end the study, please let me know. What questions do you have?
>
> If there are no further questions, we can get started. Please do your best to interpret the robot's cue and complete the desired task.

Beyond this basic introduction, participants were not given any explicit instructions about what to do in the study. After each break, the experimenter read the following text:

> As a reminder, stay in front of the marked line. You can ignore the four bits of tape on the floor. Please do not touch the joints of the robot; otherwise, you may move around (*experimenter gestured arm broadly to indicate the wide workspace*) and make contact with the robot to complete the task. Please do your best to interpret the robot's cue and complete the desired task.

If the participant asked any further questions or requested assistance with the tasks, the experiment repeated the following statement:

> Please do your best to interpret the robot's cue and complete the desired task.

A.1.2 Administered Surveys

The following surveys were administered to each participant via the secure online platform LimeSurvey.

At the beginning of the study:

Please rate how much you agree or disagree (continuous slider bar, strongly disagree – strongly agree) with each of the following statements (shown in Table A.1).

This Robot Acceptance Survey (RAS) was developed by Weiss et al. (Weiss *et al.*, 2008).

After completing the study:

1. Please rate how much you agree or disagree (continuous slider bar, strongly disagree – strongly agree) with each of the following statements (shown in Table A.2).

   This System Usability Scale (SUS) was developed by Brooke (Brooke, 1996).

2. Please rate how much you agree or disagree (continuous slider bar, strongly disagree – strongly agree) with each of the following statements (shown in Table A.3).

   This NASA-TLX was developed by Hart and Lowell (Hart and Staveland, 1988).

3. Cue evaluation (Table A.4) for the location cue

4. Cue evaluation (Table A.4) for the pose cue

I think using the robot is a **good** idea.
I am afraid to make **mistakes** while using the robot.
I am afraid to **break** something while using the robot.
People would be **impressed** if I had such a robot.
Robots are **nice** to work with.
I am afraid that I could lose my **job** because of a robot.
I could **cooperate** with the robot.
I think the robot is **easy** to use.
I **like** the presence of the robot.
I feel **threatened** by the robot.
This robot would be **useful** to me.
This robot could **help** me.
This robot could **support** me.
I consider the robot to be a **social** agent.
I feel **understood** by the robot.
I feel **comfortable** while interacting with the robot.
I could work with the robot, **if** someone **helped** me.
I could work with the robot **without** any **help**.
I could work with the robot, if I had good initial **training**.
I **trust** the robot.
I would **follow** the example of the robot.

Table A.1: Robot Acceptance Survey (RAS) (Weiss *et al.*, 2008).

5. Cue evaluation (Table A.4) for the contact cue

6. Feedback evaluation

   a) What feedback did you receive from the robot? (check all that apply: head motions, arm gestures, facial expressions, lights on the hands, nonverbal sounds, speech, other (with text entry))

   b) In your ideal scenario, what feedback do you wish the robot would provide? (check all that apply: head motions, arm gestures, facial expressions, lights on the hands, nonverbal sounds, speech, other (with text entry))

7. Please rate how much you agree or disagree (continuous slider bar, strongly disagree – strongly agree) with each of the following statements (shown in Table A.1).

8. Do you have any additional comments about the system or the study? (optional open response)

9. Demographics questionnaire

   a) Age (open response)

I think that I would like to use this system frequently.
I found this system unnecessarily complex.
I thought this system was easy to use.
I think that I would need assistance to be able to use this system.
I found various functions in this system were well integrated.
I thought there was too much inconsistency in this system.
I would imagine that most people would learn to use this system very quickly.
I found this system to be very cumbersome/awkward to use.
I felt very confident using this system.
I needed to learn a lot of things before I could get going with this system.

Table A.2: System Usability Scale (SUS) (Brooke, 1996).

How mentally demanding was the task? (continuous slider bar, Very Low – Very High)
How physically demanding was this task? (continuous slider bar, Very Low – Very High)
How hurried or rushed was the pace of this task? (continuous slider bar, Very Low – Very High)
How successful were you in accomplishing what you were asked to do? (continuous slider bar, Failure – Perfect)
How hard did you have to work to accomplish your level of performance? (continuous slider bar, Very Low – Very High)
How insecure, discouraged, irritated, stressed and annoyed were you? (continuous slider bar, Very Low – Very High)

Table A.3: NASA Task Load Index (TLX) (Hart and Staveland, 1988).

b) Gender (multiple choice: male, female, other)

c) Profession (open response)

d) Is your education/background technical? (multiple choice: yes, no)

e) Please rate your level of experience interacting with robots (prior to this study): (multiple choice:
Complete novice – I have never used a robot before today
Beginner – I have played with commercial robot toys
Intermediate – I have interacted with more sophisticated robots (like Pepper, NAO, Baxter, Kuka arm, Kinova arm, etc.)
Proficient – I have some experience interacting with and designing, building, and/or programming different kinds of robots

In your own words, please describe what the robot is telling you to do in this video (Separate videos were shown for location cue, pose cue and contact cue. Versions of these videos can be seen in Supplementary Video SV1. The videos shown to the participants were somewhat zoomed in and silent).
How confident do you feel about your understanding of this robot cue? (continuous slider bar, Not Very Confident – Very Confident)
For the trials when the robot gave you this type of cue, respond to the questions in Table A.3.
For cues of this type, how often did you perform (or not perform) movements only to find out how the robot would react? (multiple choice: never, only in one trial, in a couple trials, in several trials, in many or all trials)
Please explain. (optional open response)
Do you have any other comments about this type of cue? (optional open response)

Table A.4: Cue Evaluation

Expert – I frequently design, build, program, and/or work with robots on my own)

f) Please rate your level of experience interacting with a Baxter robot (prior to this study): (multiple choice:
Complete novice – I did not know what a Baxter robot is before today
Beginner – I have seen and/or read about a Baxter robot before
Intermediate – I have interacted with a Baxter robot via robot demos or other activities
Proficient – I have some experience interacting with and programming a Baxter robot
Expert – I frequently program and/or work with a Baxter robot)

g) How often do you exercise? (multiple choice:
Never
Once a month
Every other week
Once a week
Several times a week)

h) How many times have you taken fitness or dance courses like yoga, Zumba, gymnastics, or ballet? (multiple choice:
Never
One to three times
Between four and ten times

Between eleven and one hundred times
More than one hundred times)

i) Hometown (where you grew up): city and country (open response)

A.1.3 Additional Results

Participant Heart Rate

Fig. A.1 shows the average participant heart rate separated for the four feedback conditions; we see a wide range of values from below 60 bpm to above 130 bpm due to individual physiological differences. Formative feedback has been qualitatively shown to reduce test anxiety (Ismail *et al.*, 2022). Thus, a proper measurement and analysis of heart-rate data could have the potential to test this observation. In the future, we recommend collecting a baseline resting heart rate so that such measurements can be properly normalized by participant and then analyzed.

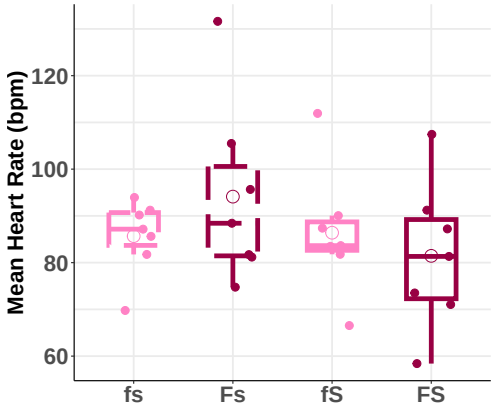


Figure A.1: Mean participant heart rate during the study separated by feedback conditions. This indicator was measured in beats per minute (bpm).

Complete Heatmaps for All Conditions

This section presents all the heatmaps generated across all three cue types and feedback types; see Figs. A.2, A.3, and A.4. Unlike the universal color scaling used in the figures in Chapter 3, the color intensities in these heatmaps have been scaled to the maximum value per cue type per feedback for better visibility of user actions.

Minimum, Mean, and Maximum Error Per Trial

To provide more insights into how formative and summative feedback affected user performance in the study, we calculated and plotted the minimum, average, and maximum

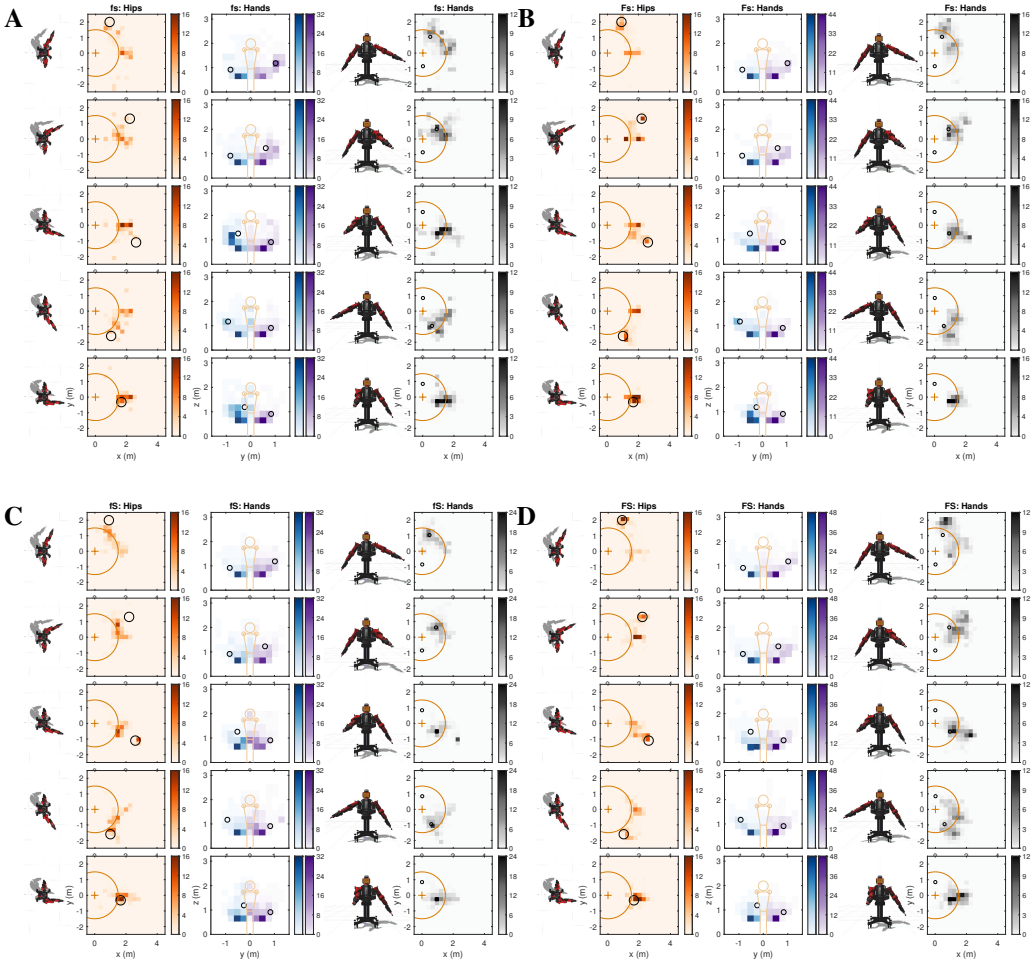


Figure A.2: Heatmap visualizations of participant activities for the location cue, separated into the four feedback conditions. Each subfigure comprises three columns of heatmaps. The first column represents the top view of the user position, which enables one to see how much time was spent by the user in different positions of the room. The second column provides an overview of the gesture performed by the user via a front view of the user's right and left hands with respect to their body. Finally, the last column represents a top view of the position of the user hand that was closest to the hand that the robot offered. If both arms of the robot were offered, this heatmap depicts both hands of the user. **(A)** shows the condition where no feedback was provided (fs). **(B)** is the case where only formative feedback was provided (Fs). **(C)** depicts summative feedback only (fS). **(D)** shows both types of feedback (FS).

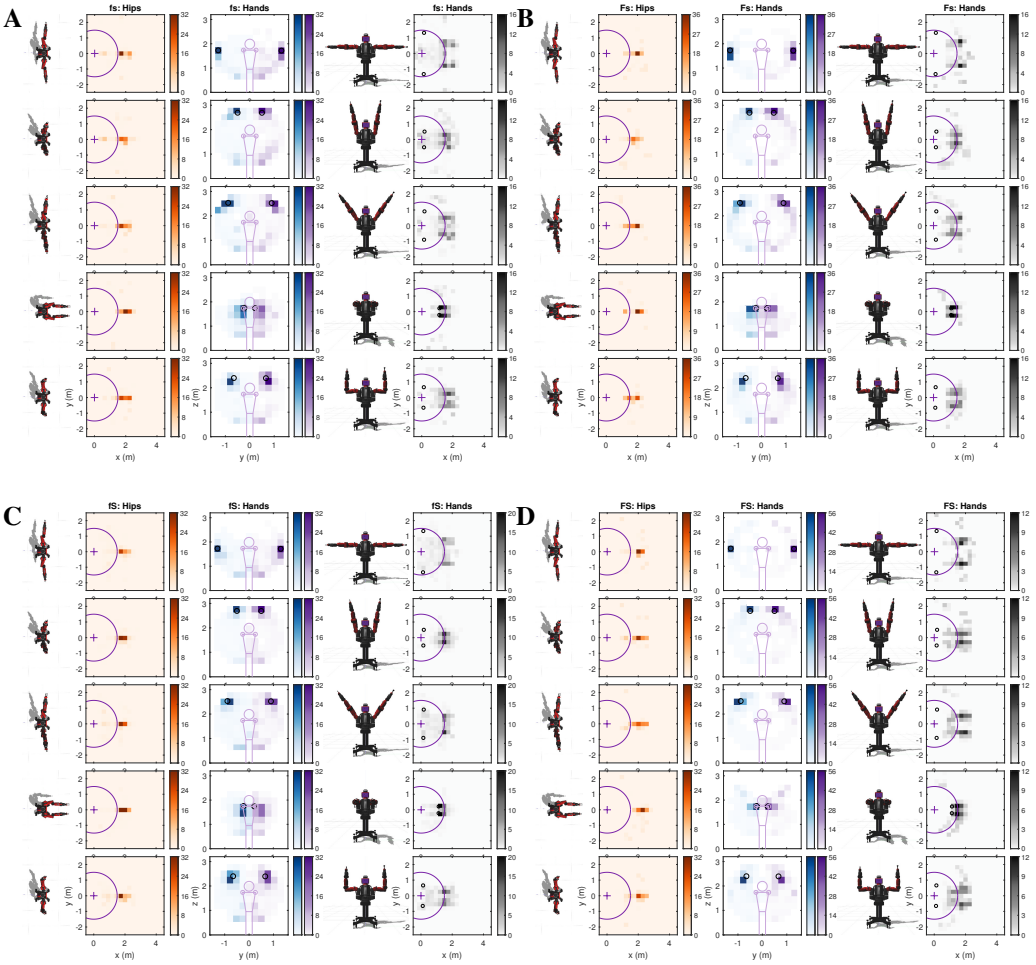


Figure A.3: Heatmap visualizations of participant activities for the pose cue, separated into the four feedback conditions. All details are the same as in Fig. A.2.

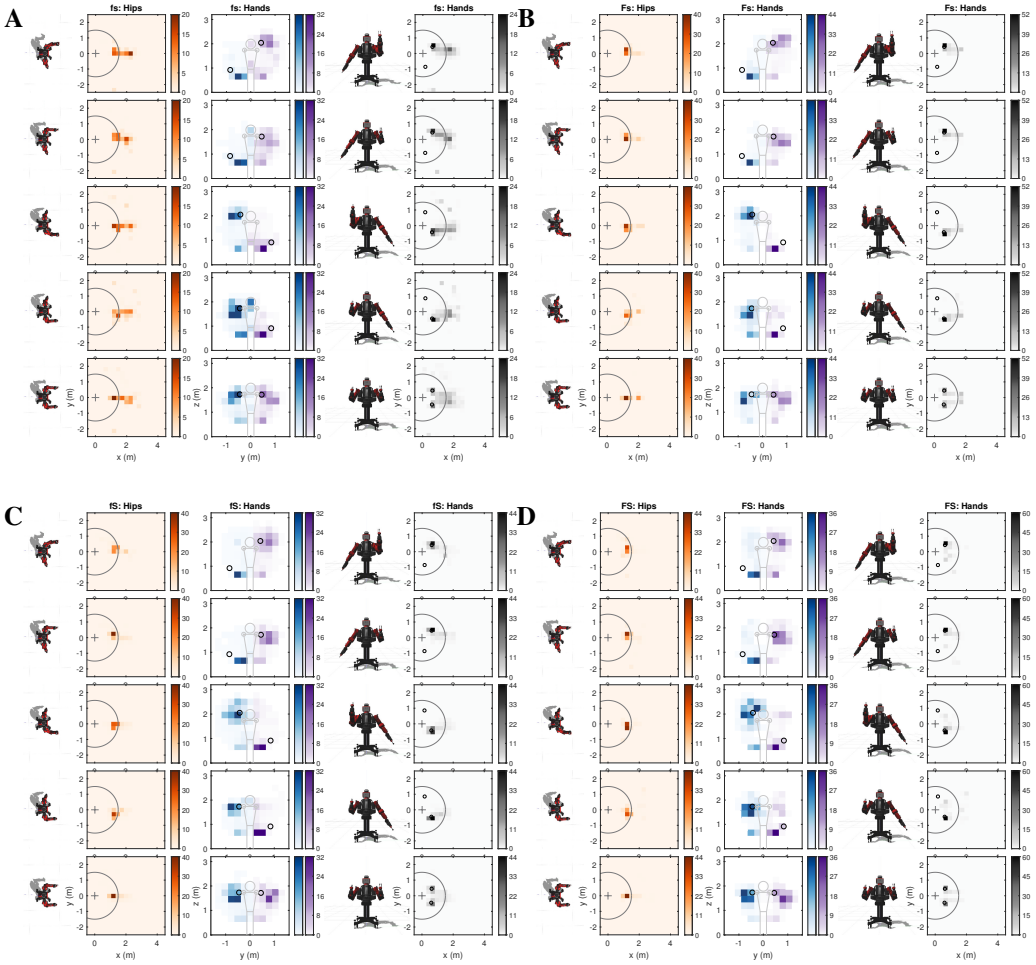


Figure A.4: Heatmap visualizations of participant activities for the contact cue, separated into the four feedback conditions. All details are the same as in Fig. A.2.

instantaneous error (distance from the desired behavior, measured in meters) that occurred in each trial in the study. We then calculated the average of each of these metrics for each user and each cue type, and we plotted the resulting data by feedback condition, as seen in Fig. A.5. The minimum error is particularly interesting because it shows whether the user ever did the desired behavior, even if they did not do it long enough to achieve a non-negligible score. Similarly, the maximum error typically occurred at the start of the trial; reducing the maximum error shows that participants were able to move toward performing the desired action before the trial even began, by correctly reading the robot's cue.

Location cue We did not find any significant differences between the feedback types for minimum, mean, or maximum error (Fig. A.5A) for the location cue. Visual inspection indicates that feedback might have slightly reduced the errors, but these trends are not statistically significant.

Pose cue As shown in Fig. A.5B, we evaluated the effects that both types of feedback had on the errors for the pose cue. First, there was a statistically significant main effect of formative feedback on the minimum error ($F(1,24) = 14, p = 0.00093, \eta_p^2 = 0.37$), mean error ($F(1,24) = 7.9, p = 0.0098, \eta_p^2 = 0.25$), and maximum error ($F(1,24) = 12, p = 0.0021, \eta_p^2 = 0.33$). The presence of formative feedback reduced all three types of errors: minimum ($t(24) = 3.8, p = 0.00093$), mean ($t(24) = 2.8, p = 0.0098$), and maximum ($t(24) = 3.5, p = 0.0021$). Second, summative feedback also had a significant effect on the minimum error ($F(1,24) = 8.6, p = 0.0073, \eta_p^2 = 0.26$) but had no effect on the mean or maximum errors. Specifically, the presence of summative feedback led to a decrease in minimum error ($t(24) = 2.9, p = 0.0073$). Thus, both types of feedback reduced user error in the pose cue.

Contact cue Finally, we evaluated the effects that both types of feedback had on the errors for the contact cue (Fig. A.5C). There was a significant two-way interaction between formative and summative feedback for mean error ($F(1,24) = 5.4, p = 0.029, \eta_p^2 = 0.18$). The presence of only summative feedback ($t(24) = 3, p = 0.04$) or both types of feedback ($t(24) = 3.4, p = 0.016$) significantly reduced the error compared to when no feedback was provided. Furthermore, there was a statistically significant main effect of summative feedback on the minimum error ($F(1,24) = 4.5, p = 0.044, \eta_p^2 = 0.16$); the presence of summative feedback reduced minimum user errors for the contact cue ($t(24) = 2.1, p = 0.044$). However, only formative feedback had a significant effect on maximum error ($F(1,24) = 7.8, p = 0.01, \eta_p^2 = 0.25$); the presence of formative feedback reduced the maximum error for the contact cue ($t(24) = 2.8, p = 0.01$). Thus, we see that both types of feedback significantly reduced some aspects of user error for this cue.

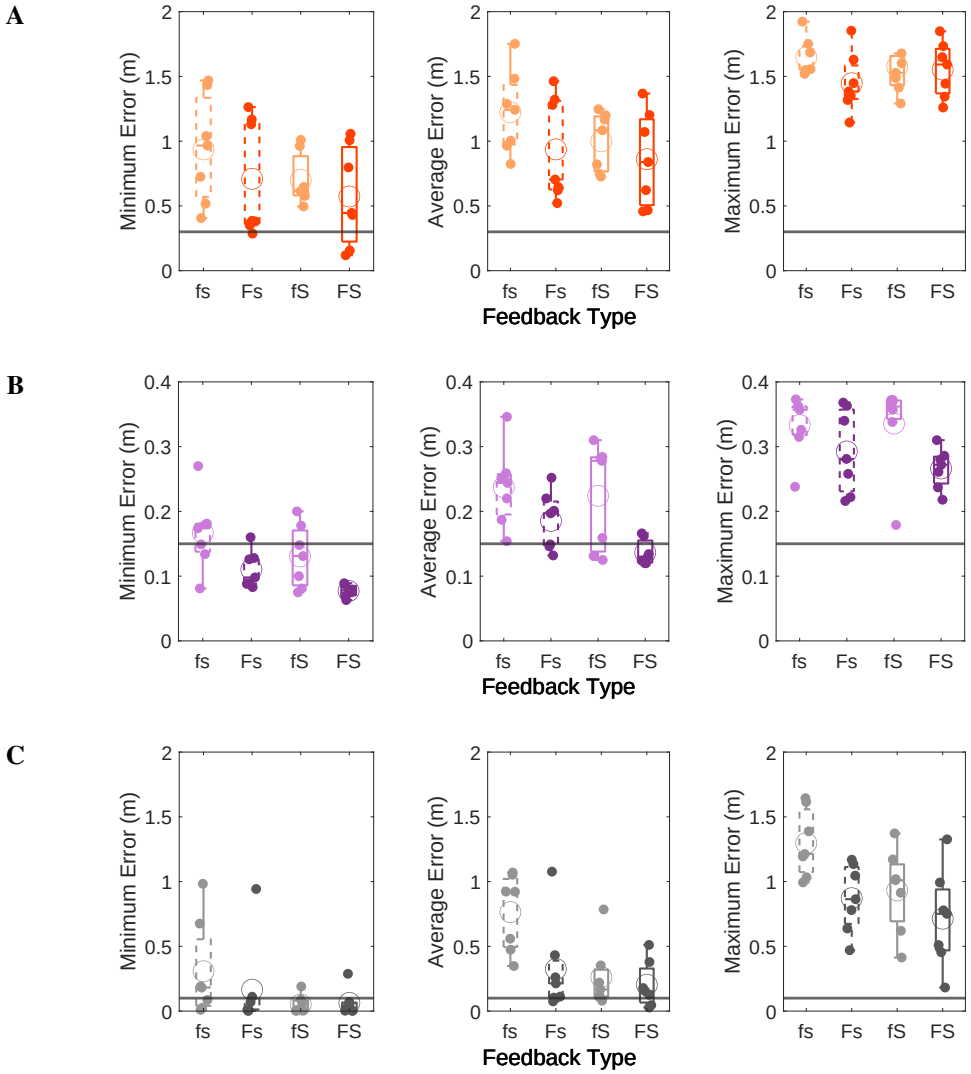


Figure A.5: Minimum, average, and maximum instantaneous errors for each cue across feedback conditions. Lower errors are better. The gray line indicates the threshold used to determine whether the user was performing the task correctly. **(A)** Errors for the location cue. There are no significant differences for this cue type. **(B)** Errors for the pose cue. The presence of formative feedback significantly reduced the minimum, average, and maximum errors, while summative feedback significantly reduced maximum error. **(C)** Errors for the contact cue. Interestingly, summative feedback significantly reduced minimum error, whereas formative feedback reduced maximum error.

A.2 Expert Perception Study

This section provides supplementary materials for the user study on expert perceptions of social robots and teleoperation for exercise, as reported in Chapter 4.

A.2.1 Administered Surveys

Opening / Closing Evaluation

This Robot Acceptance Survey (RAS) is a modified version of the survey developed by Weiss et al. (Weiss *et al.*, 2008). Due to the participant population of this study, it differs somewhat from the RAS version used in the validation study (Chapter 2) and the feedback study (Chapter 3).

Please rate how much you agree or disagree (continuous slider bar, strongly disagree – strongly agree) with each of the following statements.

I think using the robot is a **good** idea.
I am afraid to make **mistakes** while using the robot.
I am afraid to **break** the robot.
I am afraid that the robot would hurt my patient.
People would be **impressed** if I had a robot.
A robot would be **nice** to work with.
I am afraid that I could lose my **job** because of a robot.
I could **cooperate** with a robot.
I think robots are **easy** to use.
I **like** the presence of robots.
I feel **threatened** by robots.
A robot would be **useful** to me.
A robot could **help** me.
A robot could **support** me.
I consider a robot to be a **social** agent.
I feel **understood** by robots.
I feel **comfortable** while interacting with a robot.
I could work with a robot, **if** someone **helped** me.
I could work with a robot **without** any **help**.
I could work with a robot, if I had good initial **training**.
I **trust** robots.
I think my patient can **follow** the example of a robot.

Table A.5: Robot Acceptance Survey (RAS) (Weiss *et al.*, 2008) administered at the beginning and end of the study.

Exercise Robot Evaluation Questionnaire

Exercise Robot Demo

- Please watch the following video in its entirety: Narrated game play video of Roboga Game ((Fitter *et al.*, 2020))
- Please watch the following video in its entirety: Narrated game play video of Mimic Game ((Fitter *et al.*, 2020))
- Please watch the following video in its entirety: Narrated game play video of Agility Game ((Fitter *et al.*, 2020))
- Please watch the following video in its entirety: Narrated video of Location cue (Section 3.2.3)

Robot Demo Questionnaire:

- Please rate how much you agree or disagree (continuous slider bar, strongly disagree – strongly agree) with each of the following statements:
  I could cooperate with this robot.
  This robot would be nice to work with.
  I like the presence of this robot.
  I consider this robot to be a social agent.
  This robot would be useful in facilitating exercises.
  I would be interested in learning how to use this robot.
  I would like to use this robot as part of exercise therapy for my clients.
  I would recommend this robot to my clients.
  I think using this robot for exercise therapy is a good idea.
  I feel worried about using this robot.
- Which aspects of this robot do you like?
- How could this robot be improved?
- Please rate how much you agree or disagree (continuous slider bar, strongly disagree – strongly agree) with each of the following statements:
  An exercise robot should be able to talk.
  An exercise robot should be able to move its arms.
  An exercise robot should be able to show emotions.
  An exercise robot should be able to play videos on an additional screen.
  An exercise robot should be able to walk.
  An exercise robot should be able to play music.
  An exercise robot should be able to recognize the user.
- What other feedback do you have for us? (optional)

Teleoperation Demo

•Please watch the following video in its entirety: Narrated video of arm lift motion (Section 5.3)

•Please watch the following video in its entirety: Narrated video of initialization motion (Section 5.2.3)

•Please watch the following video in its entirety: Narrated video of flamenco motion (Section 5.3)

•Please watch the following video in its entirety: Narrated video of robot head and face teleoperation (Section 5.3)

Teleoperation Demo Questionnaire:

•Please rate how much you agree or disagree (continuous slider bar, strongly disagree – strongly agree) with each of the following statements:
I would like to teleoperate this robot.
Teleoperation is nice to work with.
I like the ability to customize the behavior of this robot.
Teleoperation could facilitate remote social interaction with my client.
Teleoperation could be used to create new exercises for this robot.
I would like to learn how to teleoperate this robot.
I would like to create new behaviors for this robot.
I would like to teach new exercises to my clients using this robot.
I think using teleoperation is a good idea.
I feel worried about using teleoperation.

•Which aspects of this teleoperation system do you like?

•How could this teleoperation system be improved?

•Please rate how much you agree or disagree (continuous slider bar, strongly disagree – strongly agree) with each of the following statements:

A teleoperation system for exercise therapy should allow me to talk to my client via the robot.
A teleoperation system for exercise therapy should allow me to control the robot's arms.
A teleoperation system for exercise therapy should allow me to move the robot's head.
A teleoperation system for exercise therapy should allow me to control the robot's facial expressions.
A teleoperation system for exercise therapy should allow me to play videos on an additional screen.
A teleoperation system for exercise therapy should allow me to control the robot's legs.
A teleoperation system for exercise therapy should allow me to play music.

- What other feedback do you have for us? (optional)

Demographics Questionnaire

After the closing Robot Acceptance Survey (Appendix Section A.2.1), the following demographics questionnaire was administered.

1. Age (open response)

2. Gender (multiple choice: male, female, other)

3. Highest level of education completed (open response)

4. Current Profession (open response)

5. What age group are your clients in generally? (open response)

6. What are the main reasons your patients come to you? Example: Stroke, Spinal Cord Injury etc. (open response)

7. Please rate your level of experience interacting with robots (prior to this study): (multiple choice:
Complete novice – I have never used a robot before today
Beginner – I have played with commercial robot toys
Intermediate – I have interacted with more sophisticated robots (like Pepper, NAO, Baxter, Kuka arm, Kinova arm, etc.)
Proficient – I have some experience interacting with and designing, building, and/or programming different kinds of robots
Expert – I frequently design, build, program, and/or work with robots on my own)

8. Please rate your level of experience interacting with a Baxter robot (prior to this study): (multiple choice:
   Complete novice – I did not know what a Baxter robot is before today
   Beginner – I have seen and/or read about a Baxter robot before
   Intermediate – I have interacted with a Baxter robot via robot demos or other activities
   Proficient – I have some experience interacting with and programming a Baxter robot
   Expert – I frequently program and/or work with a Baxter robot)

9. Country:(open response)
   Country (your hometown where you are from)
   Country (where you are currently practicing)

10. Thank you for your participation. Participants who are not employed by the Max Planck Society will be paid 30 euros per hour via a gift card to an online store of your choice such as Amazon. Which store and associated country do you prefer? Please provide a link if possible.

A.2.2 Semi-Structured Interview Questions

Types of Exercises
What types of exercises do you assign to clients?
How do you categorize the exercises?
Are there exercises that are common to a lot of clients?
Based on the diagnosis, what customized exercises do you assign?
How are self-practice exercises different from exercises conducted during a therapy session?
Metrics and Technology
What do you pay attention to while you are working with a client?
What measures do you normally use for therapy to quantify user performance?
What technology do you use currently in your therapy?
What technology would you like to have?
Interaction with Clients
How do you begin a therapy session?
What social interactions do you have with your client?
What feedback or guidance do you give to a client?
How do you end a therapy session?
Challenges
What are the challenges that you face?

Table A.6: Outline of questions for **phase 1** of the expert perception study. These questions focus on day-to-day interactions with clients, metrics and technology used and, finally, challenges faced.

Pre Robot Demo
As roboticists, we are curious about what ideas you have about a robot that could support you?
What is your physical image of such a robot?
Can you imagine a way in which a robot could support you or your clients in exercise therapy?
We think a robot can act as an exercise coach or partner. What do you think about this?
What exercise activities do you think a robot can do?
Now imagine that the robot is stationary humanoid robot with two arms and a face and is in a room. Specifically, we are thinking of this robot [Show picture].
What exercises or activities do you think this robot can do?
Post Robot Demo
What do you think of the robot?
What do you think of the exercise games?
What role do you think this robot can play in therapy?
What exercises would you like to perform with this robot?

Table A.7: Outline of questions for **phase 2** of the expert perception study. These questions aim to assess participant perception of social exercise robots before and after seeing example interaction videos with a Baxter robot.

Pre Teleoperation Demo
Do you think this kind of technology would be useful to you?
What role does this technology play in facilitating exercise?
If given this technology, what exercise would you like to perform using the system?
Post Teleoperation Demo
What do you think of the system?
Discuss pros/ cons. What concerns do you have about the robot?
Would you use this system?
Given such a teleoperated robot, what exercises would you like to create for the robot using teleoperation?
What we showed you now is simultaneous movement of the robot with the movement of the robot.
But, we can also record these motions and replay it later. What do you think about this?

Table A.8: Outline of questions for **phase 3** of the expert perception study. These questions aim to assess participant perception of social exercise robots before and after seeing example interaction videos with a Baxter robot.

A.2.3 Summary Statistics for RAS

RAS Item		Pre-Study		Post-Study	
	N	M	SD	M	SD
goodIdea	8	7.388	3.352	8.162	2.69
afraidMistakes	8	4.013	3.17	2.75	2.062
afraidBreak	8	2.987	2.277	2.388	2.589
PeopleImpressed	8	6.125	3.662	7.5	3.072
niceWorkWith	8	6.15	2.743	7.85	2.965
afraidJobLoss	8	0.787	0.892	0.688	0.948
cooperate	8	7.175	2.521	8.4	1.797
easyToUse	8	5.05	2.719	6.05	1.961
likePresence	8	5.875	2.958	7.45	2.343
threatened	8	0.45	0.682	1.325	2.518
robotUseful	8	7.45	3.311	8.275	1.848
couldHelp	8	7.763	2.588	7.938	2.108
couldSupport	8	7.688	2.341	8.387	1.602
socialAgent	8	5.675	3.157	7.45	2.082
feelUnderstoodbyRobot	8	3.962	2.735	4.575	3.4
feelComfortable	8	7.138	2.963	7.7	1.963
workIfHelped	8	7.4	3.237	7.737	1.655
workWithoutHelp	8	4.325	3.392	3.913	3.354
workWithTraining	8	8.287	3.078	8.95	1.564
trust	8	5.963	2.9	6.838	2.604
followExampleClient	6	5.833	2.837	7.35	3.195
afraidHurtPatient	6	6.617	3.059	3.4	2.585

Table A.9: Summary statistics for the items on the RAS administered to the experts.

A.3 Retargeting Study

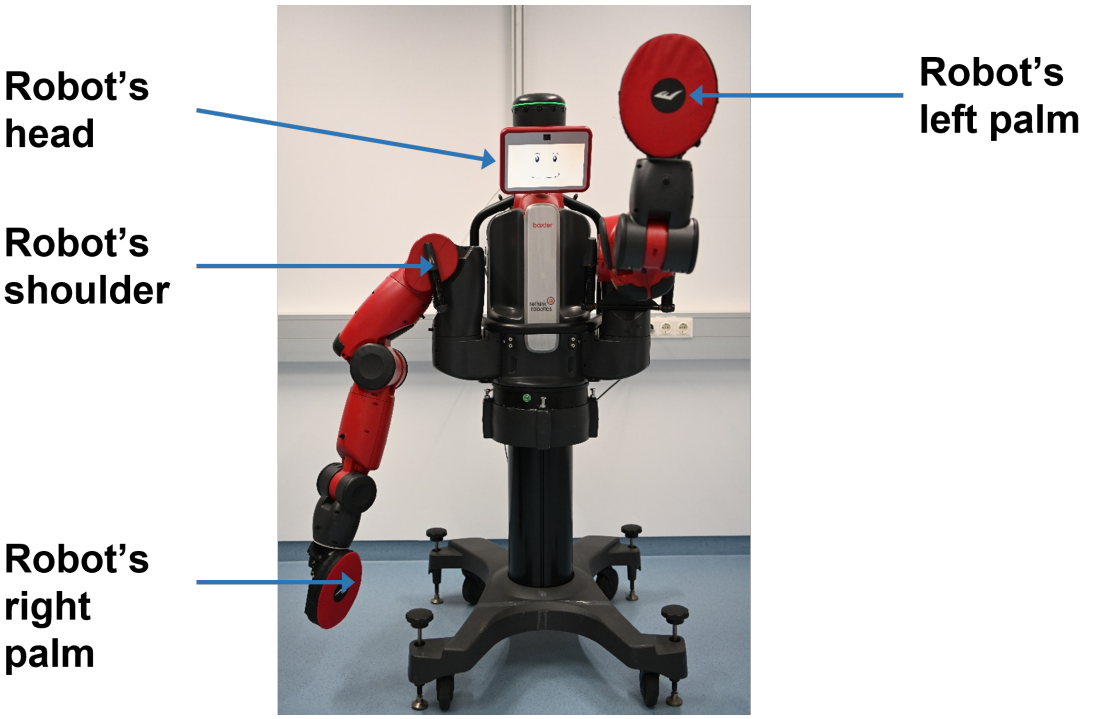
This section provides supplementary materials for the user study on kinematic retargeting reported in Chapter 5.

A.3.1 Study Instructions

Thank you for volunteering for this research study. It starts with a brief demographics survey.

The core part of the survey is **24 video comparisons**. Each of these pages will show you a video of a **person** performing a movement and a **robot** performing the same motion.

The robot in the video is a Rethink Robotics Baxter Research Robot with red boxing pads attached to its palms, as shown in the following photograph:



Your job is to **rate how similar you find the robot motion to be** in comparison to the human movement using the slider provided below each video. Optionally, you can also add comments to explain your rating.

Once you submit an answer, it is final and cannot be modified. Please do not use your browser's Back button.

Each video varies in duration from 6 to 24 seconds, which you must watch in full at least once. You are free to take a short break at any time. We estimate that this survey will take approximately 20 minutes to complete. If you are using **a mobile device** to

complete the survey, **please complete questions of this type in the landscape mode** to ensure that they display correctly.

At the end of the survey, you will be asked to reflect on the criteria you used to judge the similarity. For your data to be included, **you must answer all questions** in the survey. Please do not close this browser window until you have submitted your responses.

By proceeding, you confirm that you are in a quiet place with a stable internet connection.

A.3.2 Surveys Administered

Demographics Questionnaire

1.Age (open response)

2.Gender (multiple choice: male, female, other)

3.Profession (open response)

4.Is your education/background technical? (multiple choice: yes, no)

5.Please rate your level of expertise at human arm movements (e.g., via sports, dance, physical therapy): (multiple choice:
Complete novice – I have no personal or professional experience related to human arm movements
Beginner – I watch activities involving arm movements such as dance shows or sports via digital media
Intermediate – I directly observe or participate in activities that involve arm movements a few times a year
Proficient – I take part in activities that require specific arm movements at least once a month
Expert – I am a professional in a field that involves arm movements, such as sports, dance, or physical therapy)

6.Please rate your level of experience interacting with robots (prior to this study): (multiple choice:
Complete novice – I have never used a robot before today
Beginner – I have played with commercial robot toys
Intermediate – I have interacted with more sophisticated robots (like Pepper, NAO, Baxter, Kuka arm, Kinova arm, etc.)
Proficient – I have some experience interacting with and designing, building, and/or programming different kinds of robots
Expert – I frequently design, build, program, and/or work with robots on my own)

7. Please rate your level of experience interacting with a Baxter robot (prior to this study): (multiple choice:
   Complete novice – I did not know what a Baxter robot is before today
   Beginner – I have seen and/or read about a Baxter robot before
   Intermediate – I have interacted with a Baxter robot via robot demos or other activities
   Proficient – I have some experience interacting with and programming a Baxter robot
   Expert – I frequently program and/or work with a Baxter robot)

8. How often do you exercise? (multiple choice:
   Never
   Once a month
   Every other week
   Once a week
   Several times a week)

9. How many times have you taken fitness or dance courses like yoga, Zumba, gymnastics, or ballet? (multiple choice:
   Never
   One to three times
   Between four and ten times
   Between eleven and one hundred times
   More than one hundred times)

10. Hometown (where you grew up): city and country (open response)

Sample Video Comparison Question

1. Please watch this video (followed by video as seen in Fig. A.6). How similar is the robot motion to that of the human?: (multiple choice:
   Strongly Different
   Different
   Neither Similar nor Different
   Similar
   Strongly Similar)

2. Optionally, please explain your similarity rating.

Post Study Questionnaire: Part A

1. What aspects of the robot's motion did you consider when you compared the robot motion to that of the human?

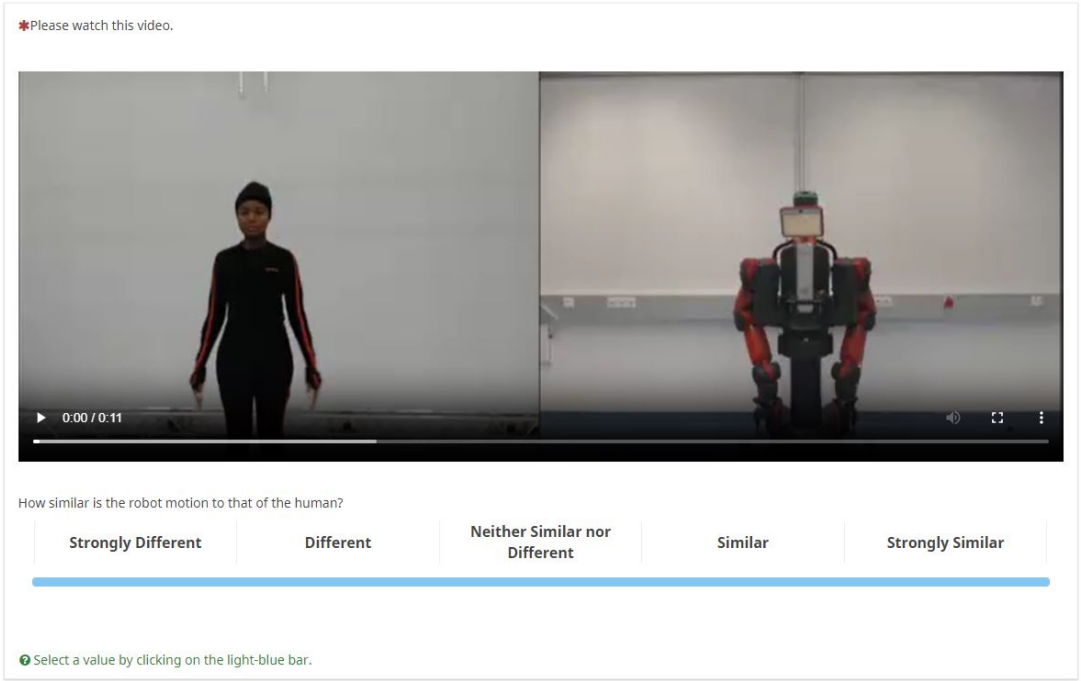


Figure A.6: A sample of the video comparison question as seen by the participant on LimeSurvey. The scale is a hybrid between a Likert scale and a Visual Analog Scale

2.What suggestions do you have for enabling the robot to match the human movement even better?

Post Study Questionnaire: Part B

1.Rank the following from most important (top) to the least important (bottom) following based on your methodology of comparing robot and human motion. (options:
Shape of arm
Orientation of hand
Position of hand
Smoothness of motion)

2.Is there something else that you also feel is important beyond the items you ranked in the previous question? If so, please enter it in the text box provided and choose where you would place your item in the ranking above using the radio buttons. If you feel the items mentioned in the previous question are sufficient, please choose none. (multiple choice with provided textbox for comment:
Above item 1
Between item 1 and item 2

Between item 2 and item 3
Between item 3 and item 4
Below item 4
None)

3. Would you like to provide any other feedback beyond what was asked in the above questions?

A.3.3 Additional Results

For the mock phase of the study, we showed four videos in a random order. There were two different movements, neither of which was included as a part of the main study videos. However, we had one video from each error combination. The participant ratings followed a trend similar to the actual study. The combination of both errors performed better.

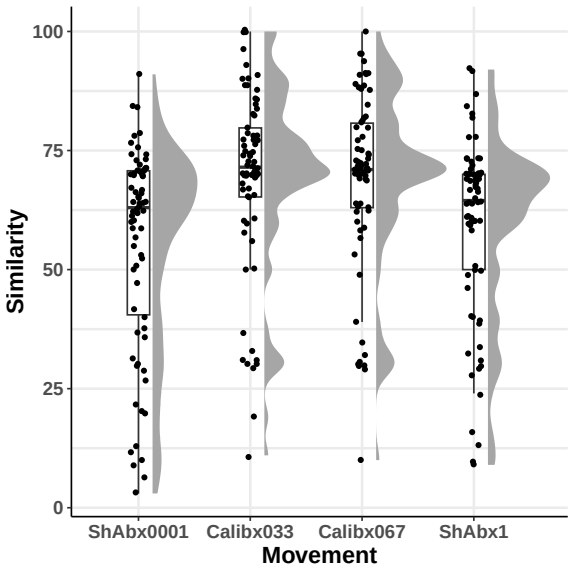


Figure A.7: Participant similarity ratings of mock phase. The general trends of the ratings are similar to those of the motions in the actual study, where the combinations of both errors (central two distributions) are rated higher than the videos where the algorithm optimized for almost only hand orientation error (left distribution) or only arm skeleton error (right distribution).

Symbols

α	Weight of skeleton error
$\vec{b}_c$	Vector from center of robot's offered hand to closest human hand
$\vec{b}_L$	Vector from center of robot's offered hand to left human hand
$\vec{b}_R$	Vector from center of robot's offered hand to right human hand
β	Weight of orientation error
d	Cohen's d or distance, depending on context
d_{Lj}	Shortest distance from joint j of left robot arm to the closest segment on the corresponding human arm
d_{Rj}	Shortest distance from joint j of right robot arm to the closest segment on the corresponding human arm
$\vec{d}_h$	Vector representing position of human joint
$\vec{d}_\tau$	Vector representing position of robot joint
$\vec{d}_{\tau_{sh}}$	Vector going through both human shoulder joints
$\vec{d}_{\tau_{shR}}$	Vector representing position of human right shoulder
$\vec{d}_{\tau_{shL}}$	Vector representing position of human left shoulder
D	Total path length travelled within the Robot Interaction Studio
ε_ℓ	Location cue error
$\bar{\varepsilon}_\ell$	Location cue error threshold
ε_p	Pose cue error
$\bar{\varepsilon}_p$	Pose cue error threshold
ε_c	Contact cue error
$\bar{\varepsilon}_c$	Contact cue error threshold
ε_s	Arm skeleton error
ε_R	Retargeting error
ε_o	Hand orientation error
$\bar{\varepsilon}_s$	Skeleton error normalization factor
€	Euro(s)
F	F statistic
h^H	Source or human kinematic chain for retargeting algorithm expressed in source frame
h^T	Source or human kinematic chain for retargeting algorithm expressed in target frame
H	Source kinematic chain expressed in world frame

j	Current robot joint
k	Distance traveled by participant in one time step
ℓ_h	Length of human arm
ℓ_{h_h}	Length of human arm segment
ℓ_τ	Length of robot arm
ℓ_{τ_h}	Length of robot arm segment
m	Number of joints of the source or human arm
n	Number of joints of the robot arm
N	Number of participants or the unit of analysis for statistical tests
η_p^2	Partial eta-squared
$\vec{o}^\tau$	Vector through the shoulders of the robot
$\vec{p}$	Desired position for the location cue
p	P-value
P	Initialization pose of OCRA
q_H^T	Quaternion representing rotation from source frame to target frame
q_h	Quaternion representing rotation of human joint
q_τ	Quaternion representing rotation of robot joint
q_ς	Current orientation of human arm end-effector frame
q_τ	Current orientation of robot arm end-effector frame
Q_d	Orientation difference between source end-effector frame and robot end-effector frame
R^2	Variation in similarity predictable from the errors
s_i	Distance from a joint on human arm to closest robot arm segment
S	Participant score location, pose and contact cue
S_p	Participant similarity rating of robot motion in comparison to human motion
$Scal_\tau$	Scaling operator to be applied to source chain to scale it to the length of the robot arm
t	Time
t_j	Distance from a joint on robot arm to closest human arm segment
T	Robot kinematic chain for retargeting algorithm expressed in target frame
τ^T	Target or robot kinematic chain for retargeting algorithm in robot frame
$Trans_\tau$	Translation to be applied on source chain to place it on robot's shoulder
θ_{sh}	Angle by which the source chain has to be rotated about the relevant axis to convert to robot frame
θ_d	Angle difference between robot and human end-effector frames
$\vec{u}_g$	Projection of the user's hip position on the ground
$\vec{u}_L$	User's current left-hand position

$\vec{u}_R$	User's current right-hand position
$\vec{v}^{\tau}$	Vector perpendicular to the vectors going through shoulders of robot and human
ζ_0	Constant component of equation relating similarity to error
ζ_1	Skeleton error coefficient of equation relating similarity to error
ζ_2	Orientation error coefficient of equation relating similarity to error

Abbreviations

2D	Two-dimensional
3D	Three-dimensional
ART	Aligned Rank Transform
ANOVA	Analysis of variance
bpm	Beats per minute
COVID	Coronavirus disease
CPU	Central processing unit
DoF	Degree of freedom
Fig.	Figure
fs	Formative feedback absent and summative feedback absent
Fs	Formative feedback present and summative feedback absent
fS	Formative feedback absent and summative feedback present
FS	Formative feedback present and summative feedback present
GigE	Gigabit Ethernet
GPU	Graphics processing unit
GUI	Graphical user interface
HI	Haptic Intelligence
HRI	Human-robot interaction
IMPRS-IS	International Max Planck Research School for Intelligent Systems
LfD	Learning from Demonstration
M	Mean
MPI-IS	Max Planck Institute for Intelligent Systems
OCRA	Optimization-based Customizable Retargeting Algorithm
ORQ	Overall Research Question
Ph.D.	Doctor of Philosophy
PoE	Power over Ethernet
Q-Q	Quantile-quantile
RAS	Robot Acceptance Survey
ROS	Robot Operating System
RQ	Research Question
SCI	Spinal Cord Injury
SSE	Sum of squares due to error of fit
SD	Standard deviation
SUS	System Usability Scale

TBI	Traumatic brain injury
TCP	Transfer Control Protocol
TLX	Task Load Index
UTAUT	Unified Theory of Acceptance and Use of Technology
VAS	Visual analog scale

List of Tables

3.1 Number of participants who identified each listed type of action as being provided by the robot during the study, separated by feedback condition. 49
3.2 Number of participants who wished for each listed type of action from the robot, separated by feedback condition. 49

4.1 Details of the different phases of the study presented in Fig. 4.1. In this table, **I** refers to an interview, **Q** refers to questionnaires or surveys, and **D** refers to video demos shown to the participant. The semi-structured interview commenced with an overview of the expert's job responsibilities, followed by an introductory survey. In the next phase, the discussion delved into the subject of exercise robots, and then transitioned to teleoperation. Finally, the participant completed both a closing survey and a demographics survey before debriefing with the experimenters. The detailed interview structure is provided in Appendix Section A.2.1 . . . 65
4.2 Participant demographics. 68

5.1 Pairwise frequencies of ranking data 101
5.2 Curve-fitting equations associated with the fits in Fig 5.17, along with the associated goodness of fit statistics. SSE is the sum of squares errors due to the fit. 104

A.1 Robot Acceptance Survey (RAS) (Weiss *et al.*, 2008). 117
A.2 System Usability Scale (SUS) (Brooke, 1996). 118
A.3 NASA Task Load Index (TLX) (Hart and Staveland, 1988). 118
A.4 Cue Evaluation 119
A.5 Robot Acceptance Survey (RAS) (Weiss *et al.*, 2008) administered at the beginning and end of the study. 126
A.6 Outline of questions for **phase 1** of the expert perception study. These questions focus on day-to-day interactions with clients, metrics and technology used and, finally, challenges faced. 131
A.7 Outline of questions for **phase 2** of the expert perception study. These questions aim to assess participant perception of social exercise robots before and after seeing example interaction videos with a Baxter robot. . 132
A.8 Outline of questions for **phase 3** of the expert perception study. These questions aim to assess participant perception of social exercise robots before and after seeing example interaction videos with a Baxter robot. . 132

A.9 Summary statistics for the items on the RAS administered to the experts. 133

List of Figures

2.1 Three high-resolution cameras show distinct views of what the minimally supervised user is doing in the Robot Interaction Studio. The cameras for the two upper views are stationary. The lower view comes from the camera attached to Baxter robot's face screen, which continually rotates to face the user based on data from the markerless motion-capture system. 8

2.2 The approximate locations of the thirteen cameras used in the Robot Interaction Studio. The three-green camera icons (A, B, C) show the placement of the high-resolution cameras that allow viewing of the interaction after it occurs, as shown in Fig. 2.1. The ten black camera icons (1–10) represent the cameras used for markerless motion capture, which have a lower resolution but a faster frame rate to allow for real-time performance. The coordinate frame shows the origin of the environment. . . 9

2.3 A screenshot of the Captury experimenter dashboard while a user takes part in the reported proof-of-concept study. Each small numbered image shows the view from one of the ten cameras used for real-time markerless motion capture. The lower-right image shows an expanded view from camera 8; note that other the locations of other cameras are annotated with a small reference frame in each camera's view. Captury also draws the human skeleton that it calculates on top of each video frame, demonstrating a good match with the user's actual pose in three dimensions. 10

2.4 Timeline showing the sequential presentation of orange, purple, and gray cues during the experiment. The robot begins with a white neutral face, holding both arms down at its sides. Each cue starts with a characteristic sound; the robot's face takes on the cue's color, and then its arms move to show the cue. When the robot reaches its final position, it nods its head to indicate that the user should respond. After 30 seconds, the arms and face return to neutral, and the next cue begins after a short pause. Cues always cycle through in the depicted color order, with each cue chosen randomly from the five available options within the color category. . . . 12

2.5 **(A)** Q-Q plot for the total distance travelled dependent variable that shows that all points fall along the reference line. **(B)** Q-Q plot for the percent time spent in workspace variable that shows that all points fall along the reference line for this dependent variable as well. 16

2.6 **(A)** Average distance participants traveled per trial of the different cue types across the floor of the Robot Interaction Studio during each cue type. **(B)** Average percentage of the trial time that each user spent inside the robot's reachable workspace during trials of each color. A value of 100% corresponds to approximately 30 seconds. Significance notations: $\star\star$ $p < 0.01$, $\star\star\star$ $p < 0.001$, and $\star\star\star\star$ $p < 0.0001$. 18

2.7 Heatmaps showing all five variants of the orange cue. The first column shows the top view of the robot's task cue instruction. The second column shows heatmap visualizations of user position in the workspace of the Robot Interaction Studio. The black circle represents the position that the robot was pointing to. The third column depicts heatmap visualizations of the front view of the user's hands. The black circles represent hand positions corresponding to the robot end-effector positions. These visualizations show that participants often either mimicked the robot or held their hands at their sides. The next column shows front views of the robot's instructions for the variants of the orange cue. The final column shows the top view of the user's hand position. The black circles are top-view positions of the robot's end effectors. This interesting set of plots shows the tendency of users to hold the robot's end-effectors. 19

2.8 Heatmaps showing all five variants of the purple cue. The first column shows the top view of the robot's task cue instruction. The second column shows heatmap visualizations of user position in the workspace of the Robot Interaction Studio; participants did not move much during these cues. The third column depicts heatmap visualizations of the front view of the user's hands. The black circles represent hand positions corresponding to the robot end-effector positions. These visualizations show that almost all participants mimicked the robot over several trials. The next column shows front views of the robot's instructions for the variants of the purple cue. The final column shows the top view of the user's hand positions. 20

2.9 Heatmaps showing all five variants of the gray cues. The first column shows the top view of the robot's task cue instruction. The second column shows heatmap visualizations of user position in the workspace of the Robot Interaction Studio. Participants generally walked into the robot's workspace. The third column depicts heatmap visualizations of the front view of the user's hands. The black circles represent hand positions corresponding to the robot end-effector positions. These visualizations show that participant pose is similar to that of the robot. The next column shows front views of the robot's instructions for the variants of the gray cues. The final column shows the top view of the user's hand position(s). These show that participants often made contact with the robot's offered hand, as seen by the darker squares near the annotated end-effector positions. 21

2.10 Robot Acceptance Survey categories based on the UTAUT model before (pre) and after (post) the study. Positive questions are colored green, and negative questions are colored pink. The paired t-tests revealed no significant results. **(A)** Participant responses regarding performance expectancy of robot. **(B)** Participant responses to questions assessing their attitudes toward technology. **(C)** Participant responses about effort expectancy. **(D)** Participant responses to self-efficacy measures. 22

2.11 Robot Acceptance Survey categories based on the concept of object-centered sociality before (pre) and after (post) the study. Positive questions are colored green, and negative questions are colored pink. Paired t-tests revealed no significant differences. **(A)** Participant responses to forms of grouping category items. **(B)** Participant responses regarding attachment toward robots. **(C)** Feelings of reciprocity reported by participants. **(D)** Participant responses to cultural perception of robots. 23

2.12 Items on the System Usability Scale. **(A)** The ratings on the individual items are quite average. Although participants felt that they would require assistance to use the system, they felt that it was well-integrated. **(B)** The overall SUS score is relatively low as a result of the lack of instructions. 23

3.1 **(A)** The three cue categories (location, pose, and contact), which each had five variants. **(B)** The formative feedback adopted by the robot for each cue type. The location cue (top row) is modeled after pointing gestures. The robot points to a location on the ground within the Robot Interaction Studio for the user to move to. For formative feedback, the robot points to the user's current incorrect location and then to the desired location. The five variants of the pose cue (second row) show the symmetric arm pose the user should imitate. For formative feedback, the robot shows an intermediate arm pose between the desired pose and a neutral pose based on the magnitude of the user's current pose error. Finally, for the contact cue (third row), the robot signals for contact by holding up one or two of its hands similar to a high five. For formative feedback, the robot wiggles its forearm (elbow joint) for a duration proportional to the current distance from the user's hands to the robot's offered hand(s). **(C)** The way in which the robot conveyed summative feedback at the end of each trial, including presentation of a red, yellow, or green face with an animated facial expression and a matching auditory cue. 29

3.2 **(A)** Sample timeline of three successive cues when both formative and summative feedback were provided (FS); the graphics above the timeline depict the robot facial expressions (Fitter and Kuchenbecker, 2016a) and arm gestures. At the end of each trial, the robot displays the summative feedback for several seconds before returning to its neutral pose. **(B)** User error over time for a sample trial of the location cue. The participant did not figure out this task and thus never achieved an error below the threshold (indicated with a dotted line), resulting in a final score of 0 seconds. The highlights in the line plot depict the seven instances of formative feedback. **(C)** Screenshots from one instance of formative feedback for the location cue. The user incorrectly attempts to hold the robot's hand; the robot points to the user's feet and then points to the correct location. **(D)** User error over time for a sample trial of the pose cue with only one instance of formative feedback; the user then copied the robot's arm pose. **(E)** Example formative feedback instance for the pose cue. The user starts out not performing the task. The robot provides formative feedback by moving its arms based on the error metric. The user then mimics the robot. **(F)** User error over time for a sample trial of the contact cue. The user repeatedly touches the robot's hand and then moves their hand back, causing oscillations in the error signal. Formative feedback is provided only when the user has not performed the task correctly for at least five seconds, which occurs twice. **(G)** Example formative feedback for the contact cue. The user can be seen mimicking the robot. The robot then performs a hand wiggle as formative feedback, and the user touches the robot's palm. All sample trials shown in this figure were taken from a participant who received both types of feedback and gave consent to share video. 32

3.3 **(A)** Participant self-reported experience with robots. **(B)** Participant self-reported experience with a Baxter robot. **(C)** Participant self-reported exercise frequency. **(D)** Participant self-reported experience with exercise classes. 37

3.4 **(A)** Boxplots of mean participant score for each condition (no feedback (fs), only formative feedback (Fs), only summative feedback (fS), or both types of feedback (FS)) and cue types. Higher scores are better, and the maximum is 30 s; the background colors depict the three levels of summative feedback. Participants performed better when the robot provided either type of feedback. **(B)** Boxplots of mean participant scores separated by level of formative feedback. Formative feedback significantly increased participant scores. **(C)** Boxplots of mean participant scores separated by level of summative feedback. Summative feedback also significantly increased participant scores. **(D)** Boxplots of mean participant scores separated by cue type. Participants earned higher scores in the pose cue and the contact cue compared to the location cue. For all boxplots, the dots mark the data points (one per participant per cue), the central line shows the median, the box shows the lower and upper quartiles, and the whiskers show the range up to 1.5 times the interquartile range. Significance notation: $\star$ $p < 0.05$, $\star\star$ $p < 0.01$, $\star\star\star$ $p < 0.001$, and $\star\star\star\star$ $p < 0.0001$. 40

3.5 **(A)** Grades of participant comprehension of each task cue based on structured grading of written explanations (seven participants in each feedback condition). **(B)** Grades separated by level of formative feedback (fourteen participants in each formative feedback level with responses for all three cue types). Formative feedback significantly improved understanding, with more participants scoring 2 or 3 points out of 3. **(C)** Grades separated by level of summative feedback (fourteen participants in each summative feedback level with responses for all three cue types). Summative feedback did not affect user understanding. **(D)** Grades separated by cue type (all 28 participants for each cue type). Participants understood the pose and contact cues better than the location cue. Significance notation: $\star$ $p < 0.05$ and $\star\star$ $p < 0.01$. 41

3.6 **(A)** Overhead view of a sample location cue trial from the condition with both formative and summative feedback (FS). The user walks around the room, leaving and entering the robot's workspace and occasionally looking at the robot's face. **(B)** Sample pose cue trial. The user stands in one position to perform the task. **(C)** Sample trial of contact cue. The user walks into the robot's workspace to touch its hands. **(D)** Mean distance travelled by each participant for each type of cue, organized by the four feedback conditions (fs, Fs, fS, and FS). **(E)** Mean attention level that the user paid to the robot. **(F)** Users paid significantly more attention to the robot during the pose cue compared to the location and contact cues. **(G)** Mean percentage of trial time that the participants spent within the workspace of the robot. **(H)** Summative feedback significantly increased the time participants spent in the robot's workspace for the contact cue. Though formative feedback also had a significant interaction with cue type for this metric, post-hoc results revealed no interesting observations. Significance notation: $\star\star$ $p < 0.01$ and $\star\star\star\star$ $p < 0.0001$. 42

3.7 User activities across different feedback conditions of the location cue. **(A)** Heatmap visualizations of user position in the workspace of the Robot Interaction Studio. Across all conditions, users can be seen spending time at various locations around the room. When feedback is provided, the users tended to spend more time near the target position depicted by the black circle. **(B)** Heatmap visualization of the front view of the user's hands for the FS condition. Participants sometimes mimicked the robot's arm pose, depicted by the black circles, and they usually kept their arms at their sides as they walked to and stood at the correct location. **(C)** Heatmap visualization of the overhead view of the user's closer hand in the horizontal plane for the FS condition. This heatmap resembles the FS condition of the user position heatmap because users often walked to the desired location. **(D)** shows the Wasserstein metric representing how much participant position in the room deviated from the correct location. **(E)** A sample location cue and the face associated with the cue seen by participants across all conditions. 44

3.8 User activities across different feedback conditions of the pose cue. **(A)** Heatmap visualizations of user position in the workspace of the Robot Interaction Studio for the FS condition. Users tend to stand in front of the robot. **(B)** Heatmap visualizations of the front view of the user's hand positions. In general, participants across all conditions figured out the desired behavior of mimicking the robot's arm pose for this cue. However, their imitation of the pose was better when feedback was provided. **(C)** shows the overhead view of the user's hand(s) in the horizontal plane for the condition where both types of feedback were provided. Since participants generally stood in front of the robot, and since all poses are symmetric, one can see both the left and right hands. **(D)** Wasserstein metric representing how much participant hand positions deviated from the correct hand positions. The data presented here is the grand mean of the Wasserstein metric between actual and desired hand positions for both left and right hands. We look at the front view in this case, since the focus is on the arm pose. **(E)** Wasserstein metric separated by level of formative feedback. The presence of formative feedback had a positive impact on pose cue performance. Participants mimicked the robot more accurately when they had formative feedback to help correct their pose. **(F)** A sample pose cue and the face associated with the cue seen by participants across all conditions. 46

3.9 User activities across different feedback conditions of the contact cue. **(A)** Heatmap visualizations of user position in the room for the FS condition. Participants generally stood directly in front of the robot to perform this task. **(B)** Heatmap visualization of the front view of the user's hands relative to their body for the FS condition. From this angle, it appears as though the user is mimicking the robot's pose since they often hold their arms in a similar way to touch the robot's hand(s). **(C)** Heatmap visualizations of an overhead view of the user's hand(s) for all conditions. The first four rows show the position of the user hand that was closest to the robot's offered end-effector; both robot hands were presented in the fifth variant of this cue, so that row depicts both hands of the user. This task cue was the easiest of all cues. Almost all participants figured out the solution to this cue when feedback was provided: they entered the robot's workspace and touched the robot's offered hand(s). **(D)** Wasserstein metric representing how much participant hand positions in the room deviated from the correct hand positions. There was a significant interaction between formative and summative feedback. All feedback conditions were significantly different from the fs condition. The FS condition was significantly better than all other conditions, and only formative feedback (Fs) was better than only summative feedback (fS). **(E)** A sample contact cue and the face associated with the cue seen by participants across all conditions. 48

3.10 Participant self-reported frequency of exploring how the robot would react to different movements. 50

3.11 Participant confidence in their explanation of each of the robot's cue types. 50

3.12 Self-reported total NASA-TLX workload results across the three cue types and the four feedback conditions (fs, Fs, fS, and FS). Table A.3 shows the six questions that constitute this survey; higher values indicate greater workload. Though there was a significant three-way interaction, post-hoc results were not significant. 51

3.13 System Usability Scale results across the four feedback conditions (fs, Fs, fS, and FS). **(A)** Plots of the responses to the ten individual questions that make up this survey (Table A.2). Positive questions are colored green, and negative questions are colored pink. **(B)** Total score. Responses to negative questions were inverted before summing with the positive responses to yield the total score shown. 51

3.14 Robot Acceptance Survey categories based on the UTAUT model across the four feedback conditions (fs, Fs, fS, and FS). Positive questions are colored green, and negative questions are colored pink. **(A)** Participant responses regarding performance expectancy of robot. They felt that the robot was less useful, less helpful, and less supportive after the study. **(B)** Participant responses to questions assessing their attitudes toward technology. Overall, after the study they were less confident that using that robot was a good idea and were less afraid to break the robot after the study. **(C)** Participant responses about effort expectancy. They felt less able to cooperate with the robot after the study. However, the study had no impact on ease of use. **(D)** Participant responses to self-efficacy measures. Summative feedback made participants feel like they could work with the robot if external help was provided. There is also some indication that the presence of formative feedback made them think they could work with the robot without help. 54

3.15 Robot Acceptance Survey categories based on the concept of object-centered sociality across the four feedback conditions (fs, Fs, fS, and FS). Positive questions are colored green, and negative questions are colored pink. **(A)** Participant responses to forms of grouping category items. There are no effects of the feedback types or the study on either liking the presence of the robot or feeling threatened. **(B)** Participant responses regarding attachment toward robots. There are no effects of the study or feedback on feelings of trust or the user's desire to follow the example of the robot. **(C)** Feelings of reciprocity reported by participants. There are no effects of the study or feedback on feelings of social agency of the robot or being understood by the robot. **(D)** Participant responses to cultural perception of robots. After finishing the study, participants felt that others would be less impressed if they owned a robot. They also were less afraid that they might lose their job because of a robot. 56

4.1 A visual representation of the experimental progression, highlighting the three primary phases, the approximate time allocation for each stage, and the administered questionnaires. 64

4.2 Expert feedback about exercise robots. Positive questions are colored green, and negative questions are colored pink. Expert responses are diverse but generally trending positive. 69

4.3 Expert responses about the features that they would like on a social exercise robot. Experts particularly like movement of arms, the ability to talk and recognize the user. 70

4.4 Expert feedback about teleoperation. Positive questions are colored green, and negative questions are colored pink. The experts were very positive about teleoperation systems. 70

4.5 Expert responses to features they would like on a teleoperation system. Experts particularly like the ability to move the arms and head of the robots. They also like the ability to talk to clients during teleoperation. . 71

4.6 Robot Acceptance Survey categories based on the UTAUT model before (pre) and after (post) the study. Positive questions are colored green, and negative questions are colored pink. **(A)** Expert responses regarding performance expectancy of robot. There were no significant differences. **(B)** Expert responses to questions assessing their attitudes toward technology. Though there were no significant differences here, their attitudes towards robots improved after the study. **(C)** Expert responses about effort expectancy. Experts felt significantly more positive about cooperating with the robot after the study. **(D)** Expert responses to self-efficacy measures. There were no significant differences in this category. 73

4.7 Robot Acceptance Survey categories based on the concept of object-centered sociality. Positive questions are colored green, and negative questions are colored pink. **(A)** Expert responses to forms of grouping category items. Though not significant, they like the presence of the robot a bit more after the study. **(B)** Expert responses regarding attachment toward robots. Experts were generally more positive about robots after the study **(C)** Feelings of reciprocity reported by experts. There were no significant differences sin this category. **(D)** Expert responses to cultural perception of robots. After finishing the study, experts significantly felt that people would be impressed if they had a robot. 74

5.1 Flowchart of OCRA. The operator's arm pose is obtained from a motion-capture system. A set of homogeneous transformations are then applied to this pose before performing the minimization to obtain the desired robot joint angles. 79

5.2 Sample screenshots of the four movement videos shown in the study for the $\alpha = 0.67$ condition. The two-armed robot tries to match the human operator's hand orientations while prioritizing matching their arm skeletons. 85

5.3 Participant similarity ratings of attention-check videos. The dashed red line shows the threshold above which participants were considered for exclusion. The larger red dots show the ratings that are above this threshold. 89

5.4 **(A)** Distribution of participant home countries. Participants hailed from 18 different countries. **(B)** Participant experience with arm movements, robots and Baxter. The participants exhibited a diverse range of experience in fields related to arm movements. They also had differing levels of expertise with robots. Finally, more than 50% of participants had never seen a Baxter robot before this study. 90

5.5 Q-Q plot for the similarity ratings shows kurtosis for almost all combinations of movements and weights. It is particularly severe for the dance sequence (row 3) and for bent arm touch (row 2). 92
5.6 The various errors relevant to the algorithm. The blue to green gradient in these plots represents the increasing skeleton error weight and decreasing orientation error weight. The bright blue curve represents almost pure orientation error whereas the bright green represents pure skeleton error. **(A)** Commanded skeleton error for the four different weights of the dance sequence movement. The blue curve has the highest skeleton error whereas the green curve has a very low skeleton error. **(B)** Commanded orientation error for the four different weights of the dance sequence movement. The green curve has the highest orientation error. **(C)** Commanded retargeting error for the four different weights of the dance sequence movement. The retargeting error is lower in magnitude because of the normalization and the weights. The two bumps in the left arm's retargeting error are the result of collision with the robot's pedestal during the left arm's curvilinear motion. 93
5.7 Skeleton error grand means across all motions and weight sets. The patterns of statistically significant differences are the same for both the commanded and actual errors. 94
5.8 Orientation error grand means across all motions and weight sets. There were differences in the post-hoc comparison results between the commanded and actual errors. 95
5.9 Retargeting error grand means across all motions and weight sets. This error is the weighted sum of the normalized skeleton and orientation errors. The significant errors reflect some of the differences in the skeleton and orientation errors. 96
5.10 Pipeline duration grand means across all motions and weight sets. The higher the skeleton error component, the higher the latency, as skeleton error is more computationally intensive to compute than hand orientation error. 97
5.11 Screenshots of the simulated Sawyer robot performing the human's right-arm motion for the four motion sequences. All simulations have a skeleton error weight of $\alpha = 0.67$. 97
5.12 Skeleton errors for Baxter and Sawyer across all motions and weight sets. The solid line shows Baxter, and the dotted line shows Sawyer. Statistically significant differences are similar to those of the benchmarking tests. The skeleton error functions similarly even for a robot that is highly different from the human arm such as Sawyer. 98
5.13 Hand orientation errors for Baxter and Sawyer across all motions and weight sets. The solid lines show Baxter, and the dotted lines show Sawyer 99

5.14 Retargeting errors for Baxter and Sawyer across all motions and weight sets. The solid lines show Baxter, and the dotted lines show Sawyer . . 99

5.15 **(A)** Rain cloud plots of participant ratings for the movements shown in the user study, each with four weights for skeleton error. Significant pairwise differences between medians (the central bars in the boxes) are indicated by lines with stars: Significance notation: $\star$ $p < 0.05$, $\star\star$ $p < 0.01$,$\star\star\star$ $p < 0.001$, and $\star\star\star\star$ $p < 0.0001$. **(B)** Participant similarity ratings presented as a stacked bar plot. This layout shows that a large percentage of participants found the robot motion to be strongly similar to that of the human operator. 100

5.16 Participant rank- ings of rating criteria 101

5.17 **(A)** Residual and contour plot for the arm lift motion. **(B)** Residual and contour plot for the bent arm touch motion. **(C)** Residual and contour plot for the dance sequence motion. **(D)** Residual and contour plot for the front clap motion. **Note:** Axis limits of the plots differ between different movement types to enhance visibility. 103

6.1 System architecture. The operator workstation handles Xsens motion tracking and runs the AFFDEX algorithm for facial expression tracking. The robot workstation is ROS compatible and is mainly responsible for the kinematic retargeting algorithm and markerless motion capture of the user. Visual, audio, and haptic data also flow from the robot-user interaction to the operator. The two systems exchange data via a TCP (Transmission Control Protocol) connection. 113

A.1 Mean participant heart rate during the study separated by feedback conditions. This indicator was measured in beats per minute (bpm). 120

A.2 Heatmap visualizations of participant activities for the location cue, separated into the four feedback conditions. Each subfigure comprises three columns of heatmaps. The first column represents the top view of the user position, which enables one to see how much time was spent by the user in different positions of the room. The second column provides an overview of the gesture performed by the user via a front view of the user's right and left hands with respect to their body. Finally, the last column represents a top view of the position of the user hand that was closest to the hand that the robot offered. If both arms of the robot were offered, this heatmap depicts both hands of the user. **(A)** shows the condition where no feedback was provided (fs). **(B)** is the case where only formative feedback was provided (Fs). **(C)** depicts summative feedback only (fS). **(D)** shows both types of feedback (FS). 121

A.3 Heatmap visualizations of participant activities for the pose cue, separated into the four feedback conditions. All details are the same as in Fig. A.2. 122
A.4 Heatmap visualizations of participant activities for the contact cue, separated into the four feedback conditions. All details are the same as in Fig. A.2. 123
A.5 Minimum, average, and maximum instantaneous errors for each cue across feedback conditions. Lower errors are better. The gray line indicates the threshold used to determine whether the user was performing the task correctly. **(A)** Errors for the location cue. There are no significant differences for this cue type. **(B)** Errors for the pose cue. The presence of formative feedback significantly reduced the minimum, average, and maximum errors, while summative feedback significantly reduced maximum error. **(C)** Errors for the contact cue. Interestingly, summative feedback significantly reduced minimum error, whereas formative feedback reduced maximum error. 125
A.6 A sample of the video comparison question as seen by the participant on LimeSurvey. The scale is a hybrid between a Likert scale and a Visual Analog Scale 137
A.7 Participant similarity ratings of mock phase. The general trends of the ratings are similar to those of the motions in the actual study, where the combinations of both errors (central two distributions) are rated higher than the videos where the algorithm optimized for almost only hand orientation error (left distribution) or only arm skeleton error (right distribution). 138

Bibliography

Adams, W. C. (2015). Conducting semi-structured interviews. *Handbook of practical program evaluation*, pages 492–505.

Admoni, H. and Scassellati, B. (2017). Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, **6**(1), 25–63.

Akalin, N., Kristoffersson, A., and Loutfi, A. (2019). The influence of feedback type in robot-assisted training. *Multimodal Technologies and Interaction*, **3**(4), 67.

Akber, S. M. A., Kazmi, S. N., Mohsin, S. M., and Szczęsna, A. (2023). Deep learning-based motion style transfer tools, techniques and future challenges. *Sensors*, **23**(5), 2597.

Alt, H. and Godau, M. (1995). Computing the Fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, **5**(01n02), 75–91.

Alzayat, A., Hancock, M., and Nacenta, M. A. (2019). Quantitative measurement of tool embodiment for virtual reality input alternatives. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1–11.

Andrade, H. and Cizek, G. J. (2010). *Handbook of Formative Assessment*. Routledge.

Argyle, M., Cook, M., and Cramer, D. (1994). Gaze and Mutual Gaze. *The British Journal of Psychiatry*, **165**(6), 848–850.

Aristidou, A., Lasenby, J., Chrysanthou, Y., and Shamir, A. (2018). Inverse kinematics techniques in computer graphics: A survey. *Computer Graphics forum*, **6**, 35–58.

Ashapkina, M., Alpatov, A., Sablina, V., and Kolpakov, A. (2019). Metric for exercise recognition for telemedicine systems. In *Proceedings of the 8th Mediterranean Conference on Embedded Computing (MECO)*, pages 1–4. IEEE.

Ayusawa, K. and Yoshida, E. (2017). Motion retargeting for humanoid robots based on simultaneous morphing parameter identification and motion optimization. *IEEE Transactions on Robotics*, **33**(6), 1343–1357.

Babinčáková, M., Ganajová, M., Sotáková, I., and Bernard, P. (2020). Influence of formative assessment classroom techniques (facts) on student's outcomes in chemistry at secondary school. *Journal of Baltic Science Education*, **19**(1).

Bajones, M., Fischinger, D., Weiss, A., Puente, P. D. L., Wolf, D., Vincze, M., Körtner, T., Weninger, M., Papoutsakis, K., Michel, D., *et al.* (2019). Results of field trials with a mobile service robot for older adults in 16 private households. *ACM Transactions on Human-Robot Interaction*, **9**(2), 1–27.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, **37**(2), 122.

Bandura, A. (1986). The explanatory and predictive scope of self-efficacy theory. *Journal of Social and Clinical Psychology*, **4**(3), 359.

Bartneck, C., Belpaeme, T., Eyssel, F., Kanda, T., Keijsers, M., and Šabanović, S. (2020). *Human-Robot Interaction: An Introduction*. Cambridge University Press.

Beard, J. R. and Bloom, D. E. (2015). Towards a comprehensive public health response to population ageing. *The Lancet*, **385**(9968), 658–661.

Bisio, A., Sciutti, A., Nori, F., Metta, G., Fadiga, L., Sandini, G., and Pozzo, T. (2014). Motor contagion during human-human and human-robot interaction. *PlOS ONE*, **9**(8), e106172.

Blake, R. and Shiffrar, M. (2007). Perception of human motion. *Annual Review of Psychology*, **58**, 47–73.

Brooke, J. (1996). SUS: a 'quick and dirty' usability scale. *Usability Evaluation in Industry*, **189**.

Burgoon, J. K., Manusov, V., and Guerrero, L. K. (2021). *Nonverbal Communication*. Routledge.

Campbell, J. I. and Thompson, V. A. (2012). MorePower 6.0 for ANOVA with relational confidence intervals and Bayesian analysis. *Behavior Research Methods*, **44**, 1255–1265.

Céspedes Gómez, N., Irfan, B., Senft, E., Cifuentes, C. A., Gutierrez, L. F., Rincon-Roncancio, M., Belpaeme, T., and Munera, M. (2021). A socially assistive robot for long-term cardiac rehabilitation in the real world. *Frontiers in Neurorobotics*, **15**, 21.

Cetina, K. K. (1997). Sociality with objects: Social relations in postsocial knowledge societies. *Theory, Culture & Society*, **14**(4), 1–30.

Chen, Z., Jiao, J., and Hu, K. (2021). Formative assessment as an online instruction intervention: Student engagement, outcomes, and perceptions. *International Journal of Distance Education Technologies (IJDET)*, **19**(1), 50–65.

Childs, E. and de Wit, H. (2014). Regular exercise is associated with emotional resilience to acute stress in healthy adults. *Frontiers in Physiology*, **5**, 161.

Childs, J. D., Benz, L. N., Arellano, A., Briggs, A. A., and Walker, M. J. (2022). Challenging assumptions about the future supply and demand of physical therapists in the united states. *Physical Therapy*, **102**(1), pzab239.

Christensen, A., Bond, S., and McKenna, J. (2022). The COVID-19 conundrum: Keeping safe while becoming inactive. A rapid review of physical activity, sedentary behaviour, and exercise in adults by gender and age. *PLoS ONE*, **17**(1), e0263053.

Cifuentes, C. A., Pinto, M. J., Céspedes, N., and Múnera, M. (2020). Social robots in therapy and care. *Current Robotics Reports*, **1**, 59–74.

Clabaugh, C., Mahajan, K., Jain, S., Pakkar, R., Becerra, D., Shi, Z., Deng, E., Lee, R., Ragusa, G., and Matarić, M. (2019). Long-term personalization of an in-home socially assistive robot for children with autism spectrum disorders. *Frontiers in Robotics and AI*, page 110.

Clark, J. M. and Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, **3**, 149–210.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.

Connors, C. B. (2021). Summative and formative assessments: an educational polarity. *Kappa Delta Pi Record*, **57**(2), 70–74.

Corder, G. W. and Foreman, D. I. (2011). *Nonparametric Statistics for Non-statisticians*. John Wiley & Sons, Inc.

Cordo, P., Lutsep, H., Cordo, L., Wright, W. G., Cacciatore, T., and Skoss, R. (2009). Assisted movement with enhanced sensation (ames): coupling motor and sensory to remediate motor deficits in chronic stroke patients. *Neurorehabilitation and Neural Repair*, **23**(1), 67–77.

Dang, L. M., Min, K., Wang, H., Piran, M. J., Lee, C. H., and Moon, H. (2020). Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition*, **108**, 107561.

Dariush, B., Gienger, M., Arumbakkam, A., Zhu, Y., Jian, B., Fujimura, K., and Goerick, C. (2009). Online transfer of human motion to humanoids. *International Journal of Humanoid Robotics*, **6**(02), 265–289.

Darvish, K., Tirupachuri, Y., Romualdi, G., Rapetti, L., Ferigo, D., Chavez, F. J. A., and Pucci, D. (2019). Whole-body geometric retargeting for humanoid robots. In *Proceedings of the IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*, pages 679–686.

Darvish, K., Penco, L., Ramos, J., Cisneros, R., Pratt, J., Yoshida, E., Ivaldi, S., and Pucci, D. (2023). Teleoperation of humanoid robots: A survey. *IEEE Transactions on Robotics*, pages 1–22.

Davison, D. P., Wijnen, F. M., Charisi, V., van der Meij, J., Evers, V., and Reidsma, D. (2020). Working with a social robot in school: a long-term real-world unsupervised deployment. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 63–72.

Dawe, J., Sutherland, C., Barco, A., and Broadbent, E. (2019). Can social robots help children in healthcare contexts? a scoping review. *BMJ paediatrics open*, **3**(1).

de Wit, J., Brandse, A., Krahmer, E., and Vogt, P. (2020). Varied human-like gestures for social robots: Investigating the effects on children's engagement and language learning. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 359–367.

Dixson, D. D. and Worrell, F. C. (2016). Formative and summative assessment in the classroom. *Theory into Practice*, **55**(2), 153–159.

Dolin, J. and Evans, R. (2017). *Transforming assessment: Through an interplay between practice, research and policy*, volume 4. Springer.

Elbeleidy, S., Rosen, D., Liu, D., Shick, A., and Williams, T. (2021). Analyzing teleoperation interface usage of robots in therapy for children with autism. In *Interaction Design and Children*, pages 112–118.

Elbeleidy, S., Mott, T., Liu, D., and Williams, T. (2022). Practical considerations for deploying robot teleoperation in therapy and telehealth. In *Proceedings of the 31st International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 977–984. IEEE.

Elkin, L. A., Kay, M., Higgins, J. J., and Wobbrock, J. O. (2021). An aligned rank transform procedure for multifactor contrast tests. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*, pages 754–768.

Feil-Seifer, D. and Matarić, M. J. (2005). Defining socially assistive robotics. In *Proceedings of the IEEE International Conference on Rehabilitation Robotics (ICORR)*, pages 465–468.

Feingold-Polak, R., Barzel, O., and Levy-Tzedek, S. (2021). A robot goes to rehab: a novel gamified system for long-term stroke rehabilitation using a socially assistive robot—methodology and usability testing. *Journal of NeuroEngineering and Rehabilitation*, **18**(1), 1–18.

Fitter, N. T. and Kuchenbecker, K. J. (2016a). Designing and assessing expressive open-source faces for the Baxter robot. In *Proceedings of the 8th International Conference on Social Robotics*, volume 9979 of *Lecture Notes in Artificial Intelligence*, pages 340–350. Springer International Publishing.

Fitter, N. T. and Kuchenbecker, K. J. (2016b). Designing and assessing expressive open-source faces for the Baxter robot. In *Proceedings of the 8th International Conference on Social Robotics (ICSR)*, volume 9979 of *Lecture Notes in Artificial Intelligence*, pages 340–350. Springer International Publishing.

Fitter, N. T. and Kuchenbecker, K. J. (2019). How does it feel to clap hands with a robot? *International Journal of Social Robotics*, **12**(1), 113–127.

Fitter, N. T. and Kuchenbecker, K. J. (2020). Synchronicity trumps mischief in rhythmic human-robot social-physical interaction. In *Robotics Research*, volume 10 of *Springer Proceedings in Advanced Robotics*, pages 269–284, Cham. Springer.

Fitter, N. T., Mohan, M., Kuchenbecker, K. J., and Johnson, M. J. (2020). Exercising with Baxter: Preliminary support for assistive social-physical human-robot interaction. *Journal of NeuroEngineering and Rehabilitation*, **17**(19).

Fitter, N. T., Mohan, M., Preston, R. C., Johnson, M. J., and Kuchenbecker, K. J. (2023). How should robots exercise with people? robot-mediated exergames win with music, social analogues, and gameplay clarity. *Frontiers in Robotics and AI*. Under review.

Fleming, K. C., Evans, J. M., and Chutka, D. S. (2003). Caregiver and clinician shortages in an aging nation. In *Mayo Clinic Proceedings*, volume 78(8), pages 1026–1040. Elsevier.

Fraune, M. R., Leite, I., Karatas, N., Amirova, A., Legeleux, A., Sandygulova, A., Neerincx, A., Dilip Tikas, G., Gunes, H., **Mohan**, t., Abbasi, N. I., Shenoy, S., Scassellati, B., de Visser, E. J., and Komatsu, T. (2022). Lessons learned about designing and conducting studies from HRI experts. *Frontiers in Robotics and AI*, **8**.

Fréchet, M. (1957). Sur la distance de deux lois de probabilité. *Comptes Rendus Hebdomadaires des Seances de L Academie des Sciences*, **6**, 689–692.

Freesound (2020). Open source sounds. `https://freesound.org/browse/`. Accessed: 2023-06-27.

Frennert, S., Eftring, H., and Östlund, B. (2017). Case report: Implications of doing research on socially assistive robots in real homes. *International Journal of Social Robotics*, **9**(3), 401–415.

Frost, E. R. (2016). Feedback distortion: The shortcomings of model answers as formative feedback. *Journal of Legal Education*, **65**(4), 938–965.

Gallace, A. and Spence, C. (2010). The science of interpersonal touch: An overview. *Neuroscience & Biobehavioral Reviews*, **34**(2), 246–259.

Gao, L., Zhang, G., Yu, B., Qiao, Z., and Wang, J. (2020). Wearable human motion posture capture and medical health monitoring based on wireless sensor networks. *Measurement*, **166**, 108252.

Gibson-Moore, H. (2019). UK chief medical officers' physical activity guidelines 2019: What's new and how can we get people more active? *Nutrition Bulletin*, **44**(4), 320–328.

GmbH, T. C. (2020). Captury Live. `https://captury.com/captury-live/`. Accessed: 2023-07-05.

Goodall, C. (1991). Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society: Series B (Methodological)*, **53**(2), 285–321.

Görer, B., Salah, A. A., and Akın, H. L. (2013). A robotic fitness coach for the elderly. In *Proceedings of the International Joint Conference on Ambient Intelligence*, pages 124–139. Springer.

Gosine, R. R., Damodaran, H., and Deutsch, J. E. (2015). Formative evaluation and preliminary validation of kinect open source stepping game. In *Proceedings of the International Conference on Virtual Rehabilitation (ICVR)*, pages 92–99. IEEE.

Guneysu, A. and Arnrich, B. (2017). Socially assistive child-robot interaction in physical exercise coaching. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*.

Ham, J., Cuijpers, R. H., and Cabibihan, J.-J. (2015). Combining robotic persuasive strategies: The persuasive power of a storytelling robot that uses gazing and gestures. *International Journal of Social Robotics*, **7**, 479–487.

Hart, J. C., Francis, G. K., and Kauffman, L. H. (1994). Visualizing quaternion rotation. *ACM Transactions on Graphics (TOG)*, **13**(3), 256–276.

Hart, S. G. and Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In *Advances in Psychology*, volume 52, pages 139–183. Elsevier.

Hattie, J. and Timperley, H. (2007). The power of feedback. *Review of Educational Research*, **77**(1), 81–112.

Hedlund, E., Johnson, M., and Gombolay, M. (2021). The effects of a robot's performance on human teachers for learning from demonstration tasks. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 207–215. ACM.

Henschel, A., Laban, G., and Cross, E. S. (2021). What makes a robot social? a review of social robots from science fiction to a home or hospital near you. *Current Robotics Reports*, **2**, 9–19.

Horak, F., King, L., and Mancini, M. (2015). Role of body-worn movement monitor technology for balance and gait rehabilitation. *Physical therapy*, **95**(3), 461–470.

Huggins-Daines, D., Kumar, M., Chan, A., Black, A. W., Ravishankar, M., and Rudnicky, A. I. (2006). Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, volume 1. IEEE.

Hyakutake, K., Morishita, T., Saita, K., Fukuda, H., Shiota, E., Higaki, Y., Inoue, T., and Uehara, Y. (2019). Effects of home-based robotic therapy involving the single-joint hybrid assistive limb robotic suit in the chronic phase of stroke: a pilot study. *BioMed Research International*, **2019**.

iMotions A/S (2020). iMotions Human Behaviour Research Software. `https://imotions.com/products/`. Accessed: 2023-07-16.

International Telecommunication Union (2019). Methodology for the subjective assessment of the quality of television pictures version bt.500-14. `https://www.itu.int/rec/R-REC-BT.500`. Accessed: March 29, 2023.

Ismail, S. M., Rahul, D. R., Patra, I., and Rezvani, E. (2022). Formative vs. summative assessment: impacts on academic motivation, attitude toward learning, test anxiety, and self-regulation skill. *Language Testing in Asia*, **12**(1), 40.

Jain, S., Thiagarajan, B., Shi, Z., Clabaugh, C., and Matarić, M. J. (2020). Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders. *Science Robotics*, **5**(39), eaaz3791.

Ju, Z., Yang, C., Li, Z., Cheng, L., and Ma, H. (2014). Teleoperation of humanoid Baxter robot using haptic feedback. In *Proceedings of the IEEE International Conference on Multisensor Fusion and Information Integration for Intelligent Systems (MFI)*, pages 1–6.

Kachouie, R., Sedighadeli, S., Khosla, R., and Chu, M.-T. (2014). Socially assistive robots in elderly care: A mixed-method systematic literature review. *International Journal of Human-Computer Interaction*, **30**(5), 369–393.

Kanda, T., Sato, R., Saiwaki, N., and Ishiguro, H. (2007). A two-month field trial in an elementary school for long-term human–robot interaction. *IEEE Transactions on Robotics*, **23**(5), 962–971.

Kaplish, A. and Yamane, K. (2019). Motion retargeting and control for teleoperated physical human-robot interaction. In *Proceedings of the IEEE-RAS 19th International Conference on Humanoid Robots*, pages 723–730. IEEE.

Kashdan, T. B., Gallagher, M. W., Silvia, P. J., Winterstein, B. P., Breen, W. E., Terhar, D., and Steger, M. F. (2009). The curiosity and exploration inventory-ii: Development, factor structure, and psychometrics. *Journal of Research in Personality*, **43**(6), 987–998.

Klaic, M., Fong, J., Crocher, V., Davies, K., Brock, K., Sutton, E., Oetomo, D., Tan, Y., and Galea, M. P. (2022). Application of the extended technology acceptance model to explore clinician likelihood to use robotics in rehabilitation. *Disability and rehabilitation: Assistive Technology*, pages 1–8.

Koenemann, J., Burget, F., and Bennewitz, M. (2014). Real-time imitation of human whole-body motions by humanoids. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2806–2812.

Kujala, U. M. (2009). Evidence on the effects of exercise therapy in the treatment of chronic disease. *British Journal of Sports Medicine*, **43**(8), 550–555.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, page 863.

Lambercy, O., Lehner, R., Chua, K. S., Wee, S. K., Rajeswaran, D. K., Kuah, C. W. K., Ang, W. T., Liang, P., Campolo, D., Hussain, A., *et al.* (2021). Neurorehabilitation from a distance: can intelligent technology support decentralized access to quality therapy? *Frontiers in Robotics and AI*, **8**, 126.

Lambert, A., Norouzi, N., Bruder, G., and Welch, G. (2020). A systematic review of ten years of research on human interaction with social robots. *International Journal of Human–Computer Interaction*, **36**(19), 1804–1817.

Landmann, C., Landi, S. M., Grafton, S. T., and Della-Maggiore, V. (2011). fMRI supports the sensorimotor theory of motor resonance. *PLOS ONE*, **6**(11), e26859.

Latikka, R., Turja, T., and Oksanen, A. (2019). Self-efficacy and acceptance of robots. *Computers in Human Behavior*, **93**, 157–163.

Lau, A. M. S. (2016). 'Formative good, summative bad?'–a review of the dichotomy in assessment literature. *Journal of Further and Higher Education*, **40**(4), 509–525.

Lee, P. H. and Yu, P. L. (2013). An r package for analyzing and modeling ranking data. *BMC medical research methodology*, **13**(1), 1–11.

Lentini, G., Settimi, A., Caporale, D., Garabini, M., Grioli, G., Pallottino, L., Catalano, M. G., and Bicchi, A. (2019). Alter-Ego: A Mobile Robot With a Functionally Anthropomorphic Upper Body Designed for Physical Interaction. *IEEE Robotics & Automation Magazine*, **26**(4), 94–107.

Lenz, C. and Behnke, S. (2023). Bimanual telemanipulation with force and haptic feedback through an anthropomorphic avatar system. *Robotics and Autonomous Systems*, **161**, 104338.

Lewis, L., Metzler, T., and Cook, L. (2016). Evaluating human-robot interaction using a robot exercise instructor at a senior living community. In *Proceedings of the International Conference on Intelligent Robotics and Applications*, pages 15–25. Springer.

Liang, Y., Li, W., Wang, Y., Xiong, R., Mao, Y., and Zhang, J. (2021). Dynamic movement primitive based motion retargeting for dual-arm sign language motions. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pages 8195–8201.

Liu, H. and Wang, L. (2018). Gesture recognition for human-robot collaboration: A review. *International Journal of Industrial Ergonomics*, **68**, 355–367.

Lyons, E. J. and Hatkevich, C. (2013). Prevalence of behavior changing strategies in fitness video games: theory-based content analysis. *Journal of Medical Internet Research*, **15**(5), e81.

Manns, M., Otto, M., and Mauer, M. (2016). Measuring motion capture data quality for data driven human motion synthesis. *Procedia CIRP*, **41**, 945–950.

Martelaro, N. (2016). Wizard-of-Oz interfaces as a step towards autonomous HRI. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium Series*.

Matarić, M. J. and Scassellati, B. (2016). Socially assistive robotics. In *Springer Handbook of Robotics*, pages 1973–1994. Springer.

Matejka, J., Glueck, M., Grossman, T., and Fitzmaurice, G. (2016). The effect of visual appearance on the performance of continuous sliders and visual analogue scales. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, pages 5421–5432, San jose, USA.

Matsusaka, Y., Fujii, H., Okano, T., and Hara, I. (2009). Health exercise demonstration robot TAIZO and effects of using voice command in robot-human collaborative demonstration. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 472–477.

McCabe, J. P., Henniger, D., Perkins, J., Skelly, M., Tatsuoka, C., and Pundik, S. (2019). Feasibility and clinical experience of implementing a myoelectric upper limb orthosis in the rehabilitation of chronic stroke patients: a clinical case series report. *PlOS ONE*, **14**(4), e0215311.

McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J., and Kaliouby, R. e. (2016). Affdex sdk: A cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, page 3723–3726, New York, NY, USA. Association for Computing Machinery.

McMahan, W., Gewirtz, J., Standish, D., Martin, P., Kunkel, J., Lilavois, M., Wedmid, A., Lee, D. I., and Kuchenbecker, K. J. (2011). Tool contact acceleration feedback for telerobotic surgery. *IEEE Transactions on Haptics*, **4**(3), 210–220.

McMillan, J. H. (2013). *SAGE Handbook of Research on Classroom Assessment*. Sage.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal About Thought*. The University of Chicago Press.

Michaelis, J. E. and Di Canio, D. (2022). Embodied geometric reasoning with a robot: The impact of robot gestures on student reasoning about geometrical conjectures. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Michaelis, J. E. and Mutlu, B. (2018). Reading socially: Transforming the in-home reading experience with a learning-companion robot. *Science Robotics*, **3**(21), eaat5999.

Midway, S., Robertson, M., Flinn, S., and Kaller, M. (2020). Comparing multiple comparisons: practical guidance for choosing the best multiple comparisons test. *PeerJ*, **8**, e10387.

Mogboh, V. E. and Okoye, A. C. (2019). Formative and summative assessment: Trends and practices in basic education. *Journal of Education and Practice*, **10**(27), 39–45.

Mohan, M. and Kuchenbecker, K. J. (2019). A design tool for therapeutic social-physical human-robot interactions. Workshop paper (3 pages) presented at the HRI Pioneers Workshop.

Mohan, M. and Kuchenbecker, K. J. (2023). OCRA: An optimization-based customizable retargeting algorithm for teleoperation. Workshop paper (3 pages) presented at the ICRA Workshop Toward Robot Avatars.

Mohan, M., Nunez, C. M., and Kuchenbecker, K. J. (2020). How does real-time feedback affect communicative actions in social-physical human-robot interaction? Workshop paper (2 pages) presented at the ROMAN Workshop on Quality of Interaction in Socially Assistive Robots (QISAR).

Mohan, M., Mat Husin, H., and Kuchenbecker, K. J. (2021a). Evaluation of a teleoperated robotic exercise coach. Workshop paper (4 pages) presented at the HRI Workshop on Workshop YOUR study design! Participatory critique and refinement of participants' studies.

Mohan, M., Nunez, C. M., and Kuchenbecker, K. J. (2021b). Robot interaction studio: A platform for unsupervised HRI. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Xian, China.

Mohan, M., Nunez, C. M., and Kuchenbecker, K. J. (2023). Formative and summative feedback in minimally supervised human-robot interaction. *Under Review*.

Morasso, P., Casadio, M., Giannoni, P., Masia, L., Sanguineti, V., Squeri, V., and Vergaro, E. (2009). Desirable features of a humanoid robot-therapist. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2418–2421.

Morrow Jr., J. R., Mood, D., Disch, J., and Kang, M. (2015). *Measurement and Evaluation in Human Performance*. Human Kinetics, 5 edition.

Movella Inc. (2018). Xsens MVN Link. `https://www.movella.com/products/motion-capture/xsens-mvn-link`. Accessed: 2023-03-27.

Nash, S. G. (1984). Newton-type minimization via the Lanczos method. *Journal on Numerical Analysis*, **21**(4), 770–788.

Niemeyer, G., Preusche, C., Stramigioli, S., and Lee, D. (2016). Telerobotics. In *Springer handbook of robotics*, pages 1085–1108. Springer.

Norris, S. (2011). Three hierarchical positions of deictic gesture in relation to spoken language: a multimodal interaction analysis. *Visual Communication*, **10**(2), 129–147.

Nourbakhsh, I. R., Bobenage, J., Grange, S., Lutz, R., Meyer, R., and Soto, A. (1999). An affective mobile robot educator with a full-time job. *Artificial Intelligence*, **114**(1-2), 95–124.

OBS Project (2020). OBS Studio. `https://obsproject.com/`. Accessed: 2023-08-10.

Okamura, A. M., Matarić, M. J., and Christensen, H. I. (2010). Medical and health-care robotics. *IEEE Robotics & Automation Magazine*, **17**(3), 26–37.

Open Robotics (2020a). ROS actionlib. `http://wiki.ros.org/actionlib`. Accessed: 2023-08-10.

Open Robotics (2020b). ROS Melodic Morenia. `http://wiki.ros.org/melodic`. Accessed: 2023-08-10.

Open Robotics (2022). ROS Noetic Ninjemys. `http://wiki.ros.org/noetic`. Accessed: 2023-08-10.

Ortenzi, V., Cosgun, A., Pardi, T., Chan, W. P., Croft, E., and Kulić, D. (2021). Object handovers: a review for robotics. *IEEE Transactions on Robotics*, **37**(6), 1855–1873.

Over, H. (2020). The social function of imitation in development. *Annual Review of Developmental Psychology*, **2**, 93–109.

Ozemek, C., Lavie, C. J., and Rognmo, Ø. (2019). Global physical activity levels – need for intervention. *Progress in Cardiovascular Diseases*, **62**(2), 102–107.

O'Reilly, M., Caulfield, B., Ward, T., Johnston, W., and Doherty, C. (2018). Wearable inertial sensor systems for lower limb exercise detection and evaluation: a systematic review. *Sports Medicine*, **48**(5), 1221–1246.

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A., Tzionas, D., and Black, M. J. (2019). Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985.

Piercy, K. L., Troiano, R. P., Ballard, R. M., Carlson, S. A., Fulton, J. E., Galuska, D. A., George, S. M., and Olson, R. D. (2018). The physical activity guidelines for Americans. *JAMA*, **320**(19), 2020–2028.

Polar Electro (2020). Polar OH1. `https://www.polar.com/us-en/products/accessories/oh1-optical-heart-rate-sensor`. Accessed: 2023-07-06.

Pulido, J. C., Suarez-Mejias, C., Gonzalez, J. C., Ruiz, A. D., Ferri, P. F., Sahuquillo, M. E. M., De Vargas, C. E. R., Infante-Cossio, P., Calderon, C. L. P., and Fernandez, F. (2019). A socially assistive robotic platform for upper-limb rehabilitation: a longitudinal study with pediatric patients. *IEEE Robotics & Automation Magazine*, **26**(2), 24–39.

Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., and Ng, A. Y. (2009). ROS: an open-source Robot Operating System. In *Companion of the Proceedings of the ICRA Workshop on Open Source Software*, volume 3, page 5. Kobe, Japan.

Rakita, D., Mutlu, B., and Gleicher, M. (2017). A motion retargeting method for effective mimicry-based teleoperation of robot arms. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 361–370.

Ranzani, R., Eicher, L., Viggiano, F., Engelbrecht, B., Held, J. P., Lambercy, O., and Gassert, R. (2021). Towards a platform for robot-assisted minimally-supervised therapy of hand function: design and pilot usability evaluation. *Frontiers in Bioengineering and Biotechnology*, **9**.

Rethink Robotics (2023). Sawyer robot. `https://www.rethinkrobotics.com/sawyer`. Accessed: 2023-08-09.

Riek, L. D. (2012). Wizard of Oz studies in HRI: A systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, **1**(1), 119–136.

Robotics, R. (2013). Rethink Robotics Baxter robot. `https://robots.ieee.org/robots/baxter/`. Accessed: 2023-03-27.

Rüschendorf, L. (1985). The Wasserstein distance and approximation theorems. *Probability Theory and Related Fields*, **70**(1), 117–129.

Rybarczyk, Y., Deters, J. K., Gonzalo, A. A., Esparza, D., Gonzalez, M., Villarreal, S., and Nunes, I. L. (2017). Recognition of physiotherapeutic exercises through dtw and low-cost vision-based motion capture. In *Proceedings of the International Conference on Applied Human Factors and Ergonomics*, pages 348–360. Springer.

Šabanović, S. (2010). Robots in society, society in robots: Mutual shaping of society and technology as a framework for social robot design. *International Journal of Social Robotics*, **2**(4), 439–450.

Samarakoon, S. B. P., Muthugala, M. V. J., and Jayasekara, A. B. P. (2022). A review on human–robot proxemics. *Electronics*, **11**(16), 2490.

Saunderson, S. and Nejat, G. (2019). How robots influence humans: A survey of nonverbal communication in social human–robot interaction. *International Journal of Social Robotics*, **11**, 575–608.

Sauppé, A. and Mutlu, B. (2014). Robot deictics: How gesture and context shape referential communication. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 342–349.

Scassellati, B. and Vázquez, M. (2020). The potential of socially assistive robots during infectious disease outbreaks. *Science Robotics*, **5**(44), eabc9014.

Scassellati, B., Boccanfuso, L., Huang, C.-M., Mademtzi, M., Qin, M., Salomons, N., Ventola, P., and Shic, F. (2018). Improving social skills in children with asd using a long-term, in-home social robot. *Science Robotics*, **3**(21), eaat7544.

Schildkamp, K., van der Kleij, F. M., Heitink, M. C., Kippers, W. B., and Veldkamp, B. P. (2020). Formative assessment: A systematic review of critical teacher prerequisites for classroom practice. *International Journal of Educational Research*, **103**, 101602.

Schultz, C., Gaurav, S., Monfort, M., Zhang, L., and Ziebart, B. D. (2017). Goal-predictive robotic teleoperation from noisy sensors. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 5377–5383.

Shao, M., Alves, S. F. D. R., Ismail, O., Zhang, X., Nejat, G., and Benhabib, B. (2019). You are doing great! only one rep left: an affect-aware social robot for exercising. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 3811–3817.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**(3/4), 591–611.

Shiomi, M., Kanda, T., Ishiguro, H., and Hagita, N. (2006). Interactive humanoid robots for a science museum. In *Proceedings of the ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, pages 305–312.

Sierotowicz, M., Connan, M., and Castellini, C. (2020). Human-in-the-loop assessment of an ultralight, low-cost body posture tracking device. *Sensors*, **20**(3), 890.

Simi Reality Motion Systems GmbH (2023). Simi Shape 3D. `http://www.simi.com/en/products/movement-analysis/markerless-motion-capture.html`. Accessed: 2023-18-11.

Soares-Miranda, L., Sattelmair, J., Chaves, P., Duncan, G. E., Siscovick, D. S., Stein, P. K., and Mozaffarian, D. (2014). Physical activity and heart rate variability in older adults: the cardiovascular health study. *Circulation*, **129**(21), 2100–2110.

Sousa, P., Oliveira, J. L., Reis, L. P., and Gouyon, F. (2011). Humanized robot dancing: humanoid motion retargeting based in a metrical representation of human dance styles. In *Progress in Artificial Intelligence: Proceedings of the 15th Portuguese Conference on Artificial Intelligence, EPIA*, pages 392–406, Lisbon, Portugal. Springer.

Stanton, C., Bogdanovych, A., and Ratanasena, E. (2012). Teleoperation of a humanoid robot using full-body motion capture, example movements, and machine learning. In *Proceedings of the Australasian Conference on Robotics and Automation*, volume 8, page 51.

StatCounter (2022). Screen resolution stats worldwide. `https://gs.statcounter.com/screen-resolution-stats`. Accessed: 2023-07-31.

Stiggins, R. and DuFour, R. (2009). Maximizing the power of formative assessments. *Phi Delta Kappan*, **90**(9), 640–644.

Swift-Spong, K., Short, E., Wade, E., and Matarić, M. J. (2015). Effects of comparative feedback from a socially assistive robot on self-efficacy in post-stroke rehabilitation. In *Proceedings of the IEEE International Conference on Rehabilitation Robotics (ICORR)*, pages 764–769.

Tanguy, P., Rémy-Néris, O., *et al.* (2016). Computational architecture of a robot coach for physical exercises in kinaesthetic rehabilitation. In *Proceedings of the 25th International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1138–1143. IEEE.

Theia Markerless, Inc. (2022). Theia 3D. `https://www.theiamarkerless.ca/`. Accessed: 2023-07-16.

Tomasello, M., Carpenter, M., and Liszkowski, U. (2007). A new look at infant pointing. *Child Development*, **78**(3), 705–722.

Tosun, T., Mead, R., and Stengel, R. (2014). A general method for kinematic retargeting: Adapting poses between humans and robots. In *Proceedings of the ASME International Mechanical Engineering Congress and Exposition (IMECE)*.

Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, **27**(3), 425–478.

Vicon Motion Systems Ltd. (2023). Vicon nexus. `https://www.vicon.com/`. Accessed: 2020-10-15.

Vilk, J. and Fitter, N. T. (2020). Comedians in cafes getting data: evaluating timing and adaptivity in real-world robot comedy performance. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 223–231.

Vossen, S., Ham, J., and Midden, C. (2010). What makes social feedback from a robot work? disentangling the effect of speech, physical appearance and evaluation. In *International Conference on Persuasive Technology*, pages 52–57. Springer.

Wang, S. and Christensen, H. I. (2018). Tritonbot: First lessons learned from deployment of a long-term autonomy tour guide robot. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 158–165.

Wang, S., Zuo, X., Wang, R., and Yang, R. (2019). A generative human-robot motion retargeting approach using a single rgbd sensor. *IEEE Access*, **7**, 51499–51512.

Wang, W., Yu, W., Qin, X., Wang, H., Hong, J., and Feng, Y. (2015a). Human motion tracking control for humanoid robot based on the optimized motion retargeting. In *Proceedings of the 12th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, pages 392–399.

Wang, Z., Williamson, R. A., and Meltzoff, A. N. (2015b). Imitation as a mechanism in cognitive development: A cross-cultural investigation of 4-year-old children's rule learning. *Frontiers in Psychology*, **6**, 562.

Wang, Z., Reed, I., and Fey, A. M. (2018). Toward intuitive teleoperation in surgery: Human-centric evaluation of teleoperation algorithms for robotic needle steering. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pages 5799–5806. IEEE.

Weiss, A., Bernhaupt, R., Tscheligi, M., Wollherr, D., Kuhnlenz, K., and Buss, M. (2008). A methodological variation for acceptance evaluation of human-robot interaction in public places. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 713–718.

Winkle, K., Caleb-Solly, P., Turton, A., and Bremner, P. (2018). Social robots for engagement in rehabilitative therapies: Design implications from a study with therapists. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 289–297.

Wobbrock, J. O., Findlater, L., Gergle, D., and Higgins, J. J. (2011). The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 143–146.

Wolf, S. L., Sahu, K., Bay, R. C., Buchanan, S., Reiss, A., Linder, S., Rosenfeldt, A., and Alberts, J. (2015). The HAAPI (Home Arm Assistance Progression Initiative) trial: a novel robotics delivery approach in stroke rehabilitation. *Neurorehabilitation and Neural Repair*, **29**(10), 958–968.

Wolfert, P., Robinson, N., and Belpaeme, T. (2022). A review of evaluation practices of gesture generation in embodied conversational agents. *IEEE Transactions on Human-Machine Systems*.

Yoganandhan, A., Kanna, G. R., Subhash, S., and Jothi, J. H. (2021). Retrospective and prospective application of robots and artificial intelligence in global pandemic and epidemic diseases. *Vacunas (English Edition)*, **22**(2), 98–105.

Zabala, U., Rodriguez, I., Martínez-Otzeta, J. M., and Lazkano, E. (2021). Expressing robot personality through talking body language. *Applied Sciences*, **11**(10), 4639.

Zhang, H., Austin, H., Buchanan, S., Herman, R., Koeneman, J., and He, J. (2011). Feasibility studies of robot-assisted stroke rehabilitation at clinic and home settings using RUPERT. In *Proceedings of the IEEE International Conference on Rehabilitation Robotics (ICORR)*, pages 1–6.

Zhang, H., Li, W., Liu, J., Chen, Z., Cui, Y., Wang, Y., and Xiong, R. (2022). Kinematic motion retargeting via neural latent optimization for learning sign language. *IEEE Robotics and Automation Letters*, **7**(2), 4582–4589.

Zhao, L., Liu, Y., Wang, K., Liang, P., and Li, R. (2016). An intuitive human robot interface for tele-operation. In *Proceedings of the International Conference on Real-time Computing and Robotics (RCAR)*, pages 454–459. IEEE.

Zimmerman, M., Bagchi, S., Marvel, J., and Nguyen, V. (2022). An analysis of metrics and methods in research from human-robot interaction conferences, 2015–2021. In *Proceedings of the 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 644–648.

Zinina, A., Zaidelman, L., Arinkin, N., and Kotov, A. (2020). Non-verbal behavior of the robot companion: a contribution to the likeability. *Procedia Computer Science*, **169**, 800–806.