Systematic user evaluation method

- For Human-Robot Interaction -

Project Report

Cecilie Pallisgaard Jensen

Aalborg University
Electronics and IT

Here you can write something about which tools and software you have used for typesetting the document, running simulations and creating figures. If you do not know what to write, either leave this page blank or have a look at the colophon in some of your books.

Electronics and IT
Aalborg University
http://www.aau.dk

AALBORG UNIVERSITY
STUDENT REPORT

Title:
Systematic user evaluation method

Theme:
Human-Robot Interaction

Project Period:
Fall Semester 2021 & Spring Semester 2022

Project Group:
989a

Participant(s):
Cecilie Pallisgaard Jensen

Supervisor(s):
Sara Nielsen
Rodrigo Ordonez

Copies: 0

Page Numbers: 129

Date of Completion:
June 1, 2022

Abstract:

This project investigates systematic user evaluation methods in HRI. Through a literature review it was found that researchers in the field of HRI use a lot of different types of data-collection methods, but do not always provide sufficient information about the analyses. Therefore, it was decided to compare different data-collection methods, with the purpose of analysing what they have in common, how they differ, how they can supplement each other, and the time-resources spend on preparation, data collection ans analysis. This was done through a user evaluation in a collaborative beer-pong scenario. The methods used were: subjective, psychophysical, and quantitative measures. The analysis of the user evaluation first go through the results from these methods separately. The comparison of the methods indicated that they had several measures of the interaction in common. Furthermore, it was found that they supplement each other well. In terms of time resources, the questionnaires had the biggest preparation time, the methods were somewhat equal in terms of time spent on collecting the data, and that the subjective analysis took the longest.

Contents

Chapter 1

Introduction

The purpose of this project is to answer the following research question:

What aspects of robot design could benefit from systematic user evaluation methods, to improve human-robot interaction?

In the aims of answering the question, the project will be divided into two subsequent parts: A literature review and a user evaluation. In the literature review different papers from Human-Robot Interaction journals and conferences will be investigated, with the purpose of finding which elements of a systematic user evaluation should be the scope of the rest of the project. In the user evaluation part, research questions relevant to chosen scope, will be presented. Thereafter, decisions concerning the user evaluation will be discussed and determined. Then an analysis of the collected data is presented. The project will conclude with a discussion of the findings throughout the project, as well as a conclusion consisting of guidelines for other researchers user evaluations, and further works within the topic of systematic user evaluation methods within Human-Robot Interaction.

Part I

Literature Review

Chapter 1

The process of the Literature Review

As mentioned in Chapter .1 this project is attempting to answer the research question; *what aspects of robot design could benefit from systematic user evaluation methods, to improve Human-Robot Interaction (HRI)*. The first step toward answering this question, is to investigate which methods are currently used when evaluating HRI. To find these methods a literature review is done.

This chapter will first of all go through the process of the literature review: how papers were found, what type of papers were of interest, and which criteria were used. From this a set of final papers will be chosen for the literature review, which will then be presented.

1.1 Step 1: the broad search

To find papers for the literature review, two databases have been used; ACM Digital Library and IEEE Explore. More specifically it was chosen to search within different journals and conferences published by ACM an IEEE. Within ACM Digital Library the journal *"Transactions of Human-Robot Interaction (THRI)"* were used. Joined from ACM and IEEE the conference *"ACM/IEEE International Conference on Human-Robot Interaction"* were also used to find literature. Within IEEE Explore the journal *"IEEE Transactions on robotics"* were used. For the literature review the following types of papers will be included: articles, short papers, proceedings, journal papers and abstracts.

Within the two journals and the conference the key-words; User study, user evaluation, human-robot interaction and social robots, were used in the following way:

"User study" OR "User Evaluation" AND "Human-Robot Interaction" OR "HRI"

AND "Social Robots"

In Table 1.1 the amount of results from this search can be seen. Furthermore it was chosen to limit the year-range to the last 10 years: from 2011-2021. The number of papers found within this time-range can also be seen in Table 1.1. This search was done on the 18th of October 2021.

Search engine	Initial Number of Results	After Year criterion Number of Results
ACM Transactions on Human-Robot Interaction	85	51
ACM/IEEE International Conference on Human-Robot Interaction	896	805
IEEE Transactions on Robotics	7	5
Total	988	861

Table 1.1: The initial results of the search explained previously, as well as the results after including the criterion concerning the year.

1.2 Step 2: Reducing the papers

To reduce the 861 papers further, another criterion is introduced. As the interest of the literature review is evaluation methods in robotics and the interaction between humans and robots, the next criterion is that the papers must evaluate on an interaction between humans and robots. To reduce the papers, first of all another key word is introduced to the search, after the keyword "Social Robots": AND "Evaluation method". This led to removing 550 papers from the literature review. An overview of the papers in the two journals and the conference, can be seen in Table 1.2. These results were found on the 18th of October 2021.

1.3 Step 3: Reduction based on citations

To find relevant papers for the literature review a new criterion is introduced. This criterion is that the papers must have at least one citation. In Table 1.3 the results of introducing this criterion can be seen in the column "Step 3". This inclusion criterion was introduced on the 19th of October 2021.

Furthermore it is decided to confine the criterion by only including papers with five or more citations. The results of this can be seen in Table 1.3, in the column "Step 3.1".

Search engine	Step 1	Step 2
ACM Transactions on Human-Robot Interaction	51	29
ACM/IEEE International Conference on Human-Robot Interaction	805	278
IEEE Transactions on Robotics	5	4
Total	861	311

Table 1.2: An overview of the differences of results from when the year criterion was introduced, to when the search term "Evaluation method" was introduced.

Search engine	Step 2	Step 3	Step 3.1
ACM Transactions on Human-Robot Interaction	29	21	11
ACM/IEEE International Conference on Human-Robot Interaction	278	195	90
IEEE Transactions on Robotics	4	3	2
Total	311	219	103

Table 1.3: The results after introducing the citation criterion, compared to the last step.

As mentioned in Section 1.1 a year-range of which the papers must have been published, to be taken into account in the literature review, was defined. To investigate how many of the papers, with five or more citations, have been published in the different years of the range, Table 1.4 is presented.

Year	Number of papers: With 5 or more citations
2011	8
2012	7
2013	2
2014	9
2015	10
2016	5
2017	22
2018	26
2019	8
2020	6
2021	0

Table 1.4: Overview of the publication years and the number of papers within those years.

It can be seen from Table 1.4 that all the years within the range, except for 2021, have published papers with five or more citations.

1.4 Step 4: The Abstracts

To reduce the papers further, the abstracts will be read. To find the papers that should be included, the next criterion is introduced. This criterion is that minimum one person have been used in the evaluation. The participants does not have to interact with a robot themselves, but can also evaluate another persons interaction with a robot. To find the information about whether one or more people were used in the evaluation, the abstracts of the 103 papers were read.

From reading the abstracts it was found that 8 of the papers did not involve tests with people. 2 of these were tutorials, which invites researchers to workshops involving different topics within the field of HRI, to find which topics should be investigated and how to do it. 2 describes abilities of robots, and improvements of these. 1 investigates the use of haptic feedback to help visually impaired people, but did not involve user testing. 1 investigates how people would prefer the driving style of an automated vehicle, and how this corresponds to their own driving style. 1 investigates the relationship between main characters and their sidekicks

from popular books, to see how these can be used in robot design. The last paper were excluded as it investigates how consent should be used in HRI. Furthermore 1 were excluded as it was a video.

Besides this, 2 papers were found to be literature reviews. Even though literature reviews do not involve user testing on people, the criterion will not be imposed on the literature reviews, as they provide insights on the structural process of literature review.

This leaves 92 papers.

1.5 Step 5: Grouping of papers

As mentioned in the beginning of this chapter, the purpose of the literature review is to find which methods have been used to evaluate the interactions between humans and robots. To take another step towards reducing the 92 papers even further, the papers were grouped based on the information present in the abstracts. This resulted in 8 different groups. The number of papers in the different groups, as well as the given name of the groups, can be seen in Table 1.5.

Group	Number of Papers
Validating the robot	1
Comparing different robots/robot designs	26
Difference in people, effects on the perception of robot(s)	1
Investigating factors of robotics	4
How robot(s) can help people learn	8
Peoples perception of robot(s)	12
Robot learning/making new features for robot(s)	38
User test setup is of interest	2

Table 1.5: The different groups of papers and the number of papers in each group.

As can be seen in Table 1.5 the first group is named *Validating the robot*, and only one paper is in this group. In this case validating is when tests has already been done on the robot, but to validate the robot these results have been tested in another

setting to confirm that the results are alike. The second group is named *Comparing different robots/robot designs*. As can be deduced from the name the papers in this group either compare different robots or robot designs, for example ways of controlling or operating a robot. As can be seen from Table 1.5 this group holds 26 papers. The third group is named *Difference in people, effects on the perception of robot(s)* which only one paper fits in to. The paper in this group investigates how different demographic measures, such as age and gender, effect their acceptance of a robot. The fourth group named *Investigating factors of robotics* consists of 4 papers. They all investigate important factors of designing robots, e.g. how to ensure privacy in a robot that helps elderly have contact with their families. The fifth group named *How robot(s) can help people learn* holds 8 papers, and evolve around how robots can help people learn different things, such as reading or changing people mindset in a positive manner. The sixth group, *Peoples perception of robot(s)*, which holds 12 papers, evolve around measuring people's perception of a robot/robots. The seventh group, *Robot learning/making new features for robot(s)*, which is the biggest group of 38 papers, holds papers that investigate new features of robot design, or how a robot should learn new things to enhance HRI. The eighth and last group *User test setup is of interest*, with 2 papers, evolve around investigating the setup which is used to test HRI.

Even though grouping of papers are not usually used when doing systematic literature reviews, this process has helped getting a better overview of the 93 papers, which can help the process of introducing a new inclusion criteria for the review.

1.6 Step 6: Reduction of papers based on the change in people or robots

To reduce the papers, a new inclusion criterion is introduced. As the theme is to investigate how interactions between humans and robots have been evaluated in the literature, the inclusion criterion is that the papers have to investigate either a change in people or robots. This step can be done by looking at the groups explained in the previous section. As can be seen Table 1.5 three of the groups can be excluded; *Validating the robot*, *Investigating factors of robotics* and *User test setup is of interest*. This excludes 7 papers from the literature review, leaving 86 papers.

1.6.1 Step 6.1: Reduction of the papers based on the interest within the topic of HRI

Furthermore, another inclusion criterion is introduced. This criterion is based on the interest within the topic of Human-Robot Interaction. Within the topic it is of big interest to look at how different designs of robots change and/or enhance the

interaction between humans and robots and how people perceive robots. Therefore, the papers which evolve around these topics will be included in the literature review. As can be seen in Table 1.5 two groups of papers meet this criteria: *Peoples perception of robot(s)* and *Robot learning/making new features for robot(s)*. This leaves 50 papers.

1.7 Step 7: The predictions in the papers

As the methods used in the papers are of interest in the literature review, it was deemed important that what they were investigating were clearly stated in the papers. Therefore, it was decided to go through the 50 papers, with the purpose of finding clear definitions of what they were investigating. This was done by searching for the following terms: "Hypothe" and "Predict", as these would give all results concerning: hypothesis, hypotheses, hypothesise, hypothesised, predict, prediction and predicted. Most of the papers, who had clear definitions of their purpose used a conjugation of "hypothesis", but it was found that some papers used a conjugation of "predict" when stating their hypothesis, and to ensure that these papers would not be excluded based on this wording, predict was introduced. Besides, another criterion was that the predictions/hypotheses were to be defined before the results sections of the experiments/user studis/user evaluations. This led to excluding 24 papers, which leaves 26 papers.

1.7.1 Step 7.1: Reading the methods of the papers

For this step the methods of the 26 papers were read. This was done to get an overview of the methods used. From reading these it was found that 2 of the papers did not qualify for the literature review. One were excluded as the hypothesis in the paper, was not a prediction used in the evaluation of the robot. The other was excluded as it did not involve an interaction between humans and robots, and furthermore there were no user evaluation. This leaves 24 papers.

Furthermore one paper did not use the user evaluation to enhance the interaction, rather it was used to test whether a system could do a similar evaluation. The participants in the user study of the paper were to watch a video of a child interacting with a robot while playing chess. The participants assignment were to answer whether the children were engaged in the game of chess or not. They used the information from the user study to extract features of engagement, and developed a system which could do the same task of measuring engagement [Sanghvi *et al.*, 2011]. This system was not evaluated by users. Because of this, this papers is also excluded from the literature review. Which leaves 23 papers for further analysis.

1.8 Step 8: Knowledge extraction

From looking through Schulz *et al.* [2019] it was found that their literature review were based on 27 papers. And the review by Venture & Kulić [2019] were based upon 53 different papers. Therefore this literature review, on methods used to evaluate human robot interaction, will be based upon the 23 papers.

The first step is to find what the evaluation methods used in the 23 papers have in common, and what separates them, for this, the experiment sections, results/analysis, and the limitations on the methods used will be read, and information relevant to their evaluation method will be extracted to be used in the literature review.

In the process of categorising information from the papers, it was found that one paper did not fit into the literature review. Basu *et al.* [2018] was excluded after reading the sections more thoroughly, as it became apparent that the paper does not involve interactions between people and robots nor data based on observations of others interacting with robots. The paper investigates how queries can be used to evaluate trajectories of cars.
By excluding this the total number of papers in the literature review is 22.

Chapter 2

Literature review about User Evaluation Methods used in HRI

Chapter I.1 describes the process of finding papers for the literature review. As explained previously, the first step was a broad search. The initial result across the three journals and conferences showed that almost 900 papers use keywords about, amongst others, user evaluations or user studies. This initial result indicates a wide use of user evaluations in the field of HRI. However, as researchers investigate different aspects within the field, the literature review will show the use of different methods in user evaluations. The use of different methods can sometimes be problematic since results from different papers are hard to compare [Bartneck *et al.*, 2009].

Therefore, the field of HRI could benefit from a systematic user evaluation method to investigate various aspects.

The first step toward developing a systematic way of evaluating robots is investigating the use of methods in user evaluations of robots in recent years. This is investigated through a literature review based on 22 papers investigating different aspects of robotics with user evaluations. Figure 2.1 shows a recap of the process of finding the 22 papers.

In this literature review, the definition of a user evaluation is;

a study where people either observe a robot/robots or have an interaction with a robot/robots, from which researchers can extract information about the robot or the participants in the study.

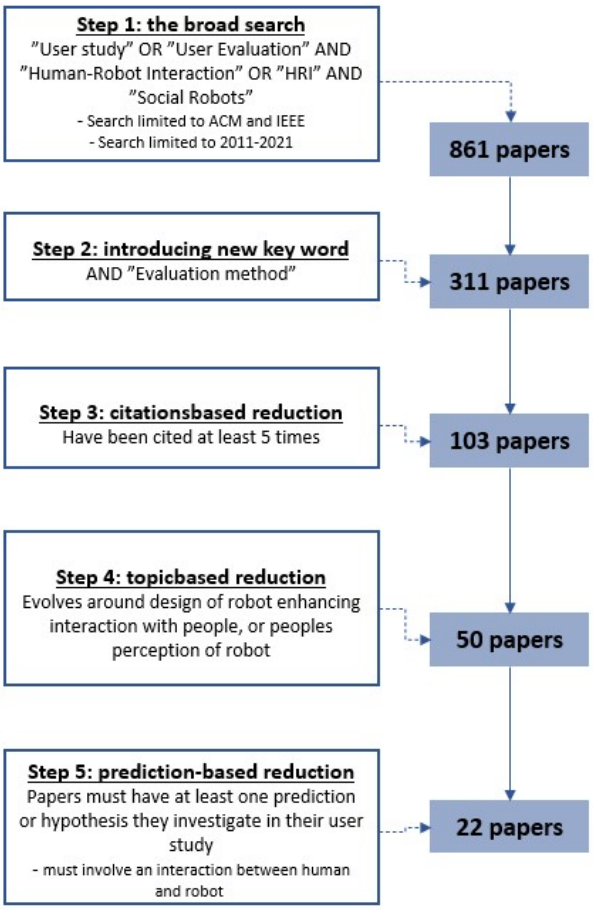

Figure 2.1: An overview of the process of finding the papers for the literature review. The process can also be found in Chapter I.1.

2.1 The use of iterations

In this literature review, iterations count how many evaluations are in the papers. These will show the differences between the use of evaluations. The use of iterations in the papers is quite different. However, four different categories are deduced: 1) One user evaluation, 2) A simulation and a user evaluation, 3) Several user evaluations, and 4) Human-Human interaction followed by Human-Robot interaction. The first category encompasses the papers that only describe one user evaluation. The second category consists of the papers that start with a system simulation and then investigate the system in a single user evaluation. The third category consists of the papers that do more than one user evaluation to investigate their hypothesis/hypotheses. The fourth category consists of the papers that first investigate a human-human interaction and use results to do a user evaluation of human-robot interaction. Table 2.1 shows the categories and the papers divided into the categories.

Group	Number of papers	References for the papers
One user evaluation	11	Javed *et al.* [2019], Rouanet *et al.* [2011], St-Onge *et al.* [2019], Kruse *et al.* [2014], Fitter *et al.* [2018], Menne & Lugrin [2017], Totsuka *et al.* [2017], Rakita *et al.* [2018], Gielniak & Thomaz [2011], Jayaraman *et al.* [2018], St. Clair & Mataric [2015]
A simulation and one user evaluation	3	Mavrogiannis *et al.* [2018], Kwon *et al.* [2020]
Several user evaluations	7	Jacq *et al.* [2016], Gielniak & Thomaz [2012], Kwon *et al.* [2018], Dragan & Srinivasa [2014], Oudah *et al.* [2015], Nikolaidis *et al.* [2018], Hanheide *et al.* [2017]
Human-Human then Human-Robot	1	Murakami *et al.* [2014], Doering *et al.* [2019]

Table 2.1: An overview of the papers in the four different groups of iterations.

2.1.1 One user evaluation

As can be seen from Table 2.1 11 of the papers only have one iteration of evaluations. Six of these papers applied the user evaluation to test a system they developed [St-Onge *et al.*, 2019; Kruse *et al.*, 2014; Totsuka *et al.*, 2017; Rakita *et al.*, 2018; Gielniak & Thomaz, 2011; St. Clair & Mataric, 2015]. All of these compared their system to another system in their user evaluations, except for St-Onge *et al.* [2019]. Of the remaining five papers in the category, four used the user evaluation to compare two or more groups. Menne & Lugrin [2017] compared facial expressions and self-reported feelings when observing friendly and torturous interactions

with a robot. Rouanet *et al.* [2011] compared four interfaces for interacting with a robot. Fitter *et al.* [2018] compared a personalised robot with a non-personalised robot. Javed *et al.* [2019] compared two robots and two groups of children with each other. In the paper by [Jayaraman *et al.*, 2018], they did not develop a system, nor did they compare different systems. They investigated trust in Autonomous Vehicles (AVs) in a virtual reality environment to find how the driving styles of an AV affected the participants' trust in the AV.

Consequently, nine of the 11 papers in this group use their user evaluation to compare systems, groups of people, robots, or interaction methods. Of these nine, Tables 10.1 and 10.2 show that the papers by Menne & Lugrin [2017]; Totsuka *et al.* [2017]; Gielniak & Thomaz [2011]; Kruse *et al.* [2014]; Javed *et al.* [2019] and St. Clair & Mataric [2015] all used a within-subject design to investigate comparisons. Common for these papers is that all of them had two conditions concerning the design of the robot [Javed *et al.*, 2019], the system of the robot [Totsuka *et al.*, 2017; Kruse *et al.*, 2014; St. Clair & Mataric, 2015; Gielniak & Thomaz, 2011], or the interaction with the robot [Menne & Lugrin, 2017]. Besides, the two groups concerning the robot Gielniak & Thomaz [2011], that investigate human-like motion, also had a condition with videos of a human making the motions. However, in the last part of the study, where participants were to choose, i.e., their preference for a certain robot motion, only the robot categories were considered. An advantage of the within-subject design is that it minimises the interpersonal differences in the analysis of the results. However, a disadvantage of using the within-subject design can be the carry-over effect, where participants get better at solving a task during the experiment. One paper in which this could affect the results is the paper by St. Clair & Mataric [2015] where the experimental setup is a pseudo-herding task with a communicative and a non-communicative robot. The participants had to solve the same tasks with the two different robot conditions. To minimise the carry-over effect, they counterbalanced the order in which the participants collaborated with the two different robot categories. Another thing these papers have in common is that all of them collect qualitative data from the participants. Menne & Lugrin [2017] collect data concerning self-reported positive and negative feelings, and Javed *et al.* [2019] collect changes in the children interacting with the robots (collected from parents). While the rest of the papers collect qualitative data concerning the robot, for example, St. Clair & Mataric [2015] collects participant-reported data about the robots as a teammate, and Totsuka *et al.* [2017] collects participant-reported data about the appropriateness of the robots they compared.

The papers by Rakita *et al.* [2018]; Rouanet *et al.* [2011]; Fitter *et al.* [2018] all used between-subject designs. Common for these three is that the participants in the different groups of the user evaluation all had to solve the same task in the different conditions of the user evaluation. The between-subject design, therefore, minimises the likelihood of a carry-over effect. However, the between-subject de-

sign also means that there can be interpersonal differences between the participants in the different conditions. As shown in Table 10.1, these three all used qualitative measures in their user evaluations. Rakita *et al.* [2018] used qualitative measures to measure the participants' perception of the robot concerning fluency, intelligence, trust, and understanding of the goal of the robot. Rouanet *et al.* [2011] used the qualitative measures to gather the participants' evaluations of the interfaces they used to interact with the robot and the task they solved. Fitter *et al.* [2018] used the qualitative measures to gather information on the perception of teleoperated robots, the perception of ownership, and the perception of self-presence. Consequently, the three papers all gathered information on the participants' perception of the robots they interacted with, which could lead to some interpersonal differences between the conditions in their between-subject designs.

In the paper by Javed *et al.* [2019], they hypothesise that long-term exposure to robots, designed for the purpose, will help children manage their negative feelings. They say that the children participating were allowed to visit the robots eight times; however, they only present the results from the first visit. Furthermore, it does not seem that they published a new paper in the journal (ACM Transaction on Human-Robot Interaction) with the complete data-set and analysis [1]. Therefore, the paper has only one iteration regarding this literature review. Another paper only presenting preliminary results is the paper by Menne & Lugrin [2017]. The paper is an abstract. They researched muscular activity on the human face when people watched videos of two different interactions with a robot; a friendly and a torturous interaction. They define the smallest unit of observable muscular activity on the human face as an Action Unit. In their analysis, they investigate two of these Action Units. However, they mention in the following discussion that they intend to investigate other Action Units using the data from the user evaluation. The ACM digital library was searched to find whether a paper had been published with a complete analysis. However, this does not seem to be the case [2]. From the abstract, it does not seem that their further work would involve another user study, which suggests that the category of the paper would not change. The study by St-Onge *et al.* [2019] that investigated robots' ability to recognise expressive motion started by building a data-set with the help of a dancer and a choreographer to train the system they developed. Subsequently, they made a user study with dancers. The study was a one-session evaluation consisting of three coherent parts: 1) Participants observed and commented on the pre-designed motion sequences of the robots. 2) Participants had 10 minutes to create expressive dance sequences for three to six of the pre-defined emotions, which were used to train the robots to recognise the emotions from expressive dancing. 3) 3-6 minutes of improvised dancing to which the robots reacted live. As these three parts were all done in one

[1] This search was done on the 20th of December 2021

[2] This search was done on the 20th of December 2021

session, the paper is categorised as One user evaluation.

2.1.2 A simulation and one user evaluation

The second category consisted of two papers, as these have both a simulation and a user evaluation. In the paper by Mavrogiannis *et al.* [2018] they start by developing a system for multi-agent planning. They test the system in a simulation where they compare the system with other common multi-agent planners [Mavrogiannis *et al.*, 2018]. They used their simulation to confirm that their system made simpler trajectories than the other systems. After the simulation, they did a user evaluation of their system. In the user evaluation, the participants watched 15 videos of a simulated workplace with five agents. The participants had to answer how two agents would pass each other (left or right) for each video. The purpose of the user evaluation was to find the correlation between entanglement and participants' ability to predict the trajectories of two agents and how fast they predicted this correctly. In the paper by Kwon *et al.* [2020], they develop a system to anticipate the risk-aware behaviour of people. They have three iterations in their study. First, they investigate risk-aware choices made by participants in an online survey. They used these results in a simulation to find how well the system they developed anticipated the participants' choices. In the second study, involving both a robot and humans, participants watched videos of towers built from cups by a human and a robot. Of the towers, 5 were examples of inefficient but stable towers(five were successful), and 5 were examples of efficient but unstable towers (one was successful). The participants were to answer which type of tower they would build with the robot. They then investigated how well a robot using their risk-aware system predicted the participants' choices compared to a robot using a noisy rational system. After finding that the risk-aware robot more frequently predicted the participants' actions correctly than the noisy rational system, they set up a collaborative user evaluation between a robot and a human. The participants and the robots were to collaborate on a cup-stacking task. The participants had ten trial rounds of stacking the cups. The robot was trained to predict the actions of the participants five times using the risk-aware system and five times using a noisy rational system. Subsequently, the participants did the cup-stacking task twice, once with each robot system, and evaluated the collaboration. The user evaluation confirmed that the risk-aware system anticipated participants' actions better than the robot using the noisy rational system.

Of the two papers in this category, Kwon *et al.* [2020] compared their developed system with another system in their user evaluation, and Mavrogiannis *et al.* [2018] compared their system with another system in their simulation. Furthermore, it can be seen from Tables 10.2 and 10.3 that both papers used a within-subject design in their user evaluation. In the paper, by Kwon *et al.* [2020] the within-subject

design might result in a carry-over effect since the participants' task is to build towers while collaborating with the two different robots. However, it could be argued that the carry-over effect is minimised as the participants do ten training trials before the final two trials where the collaborations with the robots take place.

2.1.3 Several user evaluations

Table 2.1 shows that seven of the papers had several user studies. The first one, by Jacq *et al.* [2016], investigates how a teacher-apprentice relationship can enhance children's handwriting skills when the child is the teacher, and the robot is the apprentice. Their first two iterations of their user evaluations had one participant each, where the tasks were tailored to the childrens' struggles with handwriting. The third user evaluation had eight participants whom all had the same tasks. Furthermore, the children in the third study had the possibility of evaluating the robot with thumbs up or down when the robot had written something based on the children's demonstrations. Even though the three different user evaluations are described subsequently, it does not seem that the results from the first two evaluations were used in the third. Gielniak & Thomaz [2012] investigated exaggerated motion in a storytelling scenario. Exaggeration is when trajectories emphasise a motion [Gielniak & Thomaz, 2012]. They conducted two user studies: The first one explored whether exaggerated motions would improve the recollection of the story details. In the second, they investigated what participants looked at on the robot while it told a story using exaggerated or non-exaggerated motion. The results from the first user study were not used in the second user study. Kwon *et al.* [2018] investigated how to express robot incapability. They did two preliminary user evaluations: First, they investigated the robot's timing of the attempt to solve a task and the rewind motions that best express robot incapability. The second investigated whether repetition of the attempt to solve a task helps express incapability. These preliminary studies found that the timing should be moderate or fast to best express incapability. Moreover, that repetition of the attempt helped to express incapability. These findings were used in the main user study. They investigated whether the timing and repetition helped participants see the robot's goal and why the robot was incapable of achieving the goal. Dragan & Srinivasa [2014] investigated familiarisation of robot motion. They did three user studies: In the first, they investigated how familiarisation helps predict robot motion. This study found that familiarisation helped participants recognise and predict the robot's motion. In the second, they investigated whether familiarisation helped participants predict the less natural motion of a robot and whether this motion would be as predictable as a more natural one. They compared the results with the first two studies and found that the prediction of a less natural motion was not as good as that of a more natural motion. They made a follow-up where they found that number of

repetitions did not help the predictability of the less natural motion. They used the findings of repetition and familiarisation in their third user study. This study investigated whether familiarisation would improve the comfort of working side-by-side with a robot. Oudah *et al.* [2015] investigated how cheap talk could improve the outcome of a collaborative task. They did two subsequent user evaluations. In the first one, they investigated which learning algorithm they should use for their robot. Here participants were divided into three groups: In one of the groups, participants were paired with other participants, and in the other two, participants were paired with two different learning algorithms. In this user evaluation, the participants were not allowed to communicate with their assigned partner during the task. From this user evaluation, they found that the algorithm Gabe-S++ best fitted their needs. For the second user evaluation, they developed two different cheap-talk systems for the learning algorithm: One which only provided feedback and one that provided both feedback and planning. The participants in the second user evaluation were also divided into three groups: In one, participants were paired with other participants, and in the other two, participants were paired with one of the cheap talk systems. All participants were allowed to communicate with their assigned partners to investigate the effects of cheap talk in collaborative tasks. The results of the second user evaluation were compared to the results of the first user evaluation. Nikolaidis *et al.* [2018] had five subsequent user evaluations. In the first, they investigated the priory of people in a table-turning task with a robot. The participants were to choose how to turn the table but did not get information on how the robot would prefer to turn the table. In the task, the robot's preferred way was the turn making the robot face the door. The participants chose the way they wanted to turn it, and if that were not the same way as the robot preferred, the table would not turn. In some cases, the robot would issue a verbal command of which way to turn the table, and in others, it would not. This study found that the participants' adaptability improved when the robot issued a verbal command. Nikolaidis *et al.* [2018] use adaptability as a measure of whether the participants would change the way they wanted to turn the table. In the second user evaluation, they investigated the robot issuing a state-conveying verbal command (*"I know the best way"*). The participants interacted with the robot twice, once without any verbal command and once with the state-conveying verbal command. They found that participants were more adaptable after the robot issued a state-conveying command. In the third user evaluation, they compared the state-conveying command to a compliance command (*"let's turn this way"*). They found that participants were more likely to adapt when the robot issued a compliance command, but they also tended to doubt the robot when it issued the state-conveying command. Because of the doubt that the participants had in the robot, a follow-up investigated whether the participants would be less doubtful and more adaptable if the state-conveying command were phrased differently (*"I think I know the best way..."*). Comparing

these results to the results of the state-conveying condition of the third user evaluation, they found that this phrase improved the adaptability but that participants still had doubts about the robot. Therefore, they did a second follow-up, where they investigated whether an explanation of the state-conveying command would help (*"My camera needs to face the door"*) and compared it to the third user evaluation and the first follow-up. Consequently, they used the results from previous user evaluations to investigate this type of collaboration further. In the paper by Hanheide *et al.* [2017], they investigate the robot's ability to move to different locations for different purposes. Besides this long-term investigation, they also did a usability test with 13 of the residents of the care house. The usability test results were not used in the long-term investigation and vice versa.

Different experiment designs have been used in the category of papers with several user evaluations. Jacq *et al.* [2016] and Hanheide *et al.* [2017] both used within-subject design in all of their user evaluations. Furthermore, neither of the papers describes comparisons of their system with other systems; thus, participants did not have to solve the same tasks under different conditions. Gielniak & Thomaz [2012] and Oudah *et al.* [2015] both used between-subject design in all of their user evaluations. Gielniak & Thomaz [2012], who investigated exaggerated motions, divided their participants into four different groups that all heard the same story from a robot. Oudah *et al.* [2015] investigated cheap-talk, and their participants were divided into three different groups in both of their user evaluations. In these studies, there could have been a possibility of a carry-over effect had the design been within-subject. Based on this, the between-subject design is arguably the better choice, even with some interpersonal differences. Kwon *et al.* [2018] used a mixed design in all three of their user evaluations. The robot's task type was a between-subjects variable in the two preliminary user evaluations. In the first preliminary user evaluation, the within-subject variable was timing, meaning that all participants were exposed to the three timing conditions but only for one of the robot tasks. In the second preliminary user evaluation, the within-subject variable was repetition, meaning that all participants were exposed to all three repetition conditions, but again, only for one of the robot tasks. The first part of their main user evaluation measured goal recognition and cause of incapability recognition as a between-subject variable. The last part of the main user evaluation measures the participants' perception of the methods the robots used as a within-subject variable. Dragan & Srinivasa [2014] and Nikolaidis *et al.* [2018] both used different designs in the different user evaluations they reported. Dragan & Srinivasa [2014] investigated the impact of familiarisation and naturalness of robot motion when predicting robot motion. They used a within-subject design for their two first studies. Their third user evaluation was a mixed design with familiarisation as the within-subject variable and naturalness of the motion as the between-subject variable. They had two conditions: their proposed method of predictable

motion and a less natural alternative. Nikolaidis *et al.* [2018] investigated the impact of communication from a robot in a collaborative task and used both within- and between-subject designs. The first user evaluation compared no communication with communication as within-subject variables, resulting in all participants solving a table-turning task twice. In the second, they compared no communication with a state-conveying command from the robot as a within-subject variable. The participants were to solve a table-turning task twice. In the third, they compared no communication with a state conveying and a compliance command from the robot as a between-subject variable. Furthermore, they used within-subject design in their two follow-up user evaluations, investigating the phrasing of the state-conveying command, but compared it the results of their third study's state-conveying command condition in a between-subjects manner.

2.1.4 Human-Human interaction followed by Human-Robot interaction

The fourth category of iterations consists of two papers. The paper by Murakami *et al.* [2014] starts with a human-human investigation of walking side-by-side. They investigate how a person not knowing the destination adapts to another person who does know the destination. They use these results to implement the person's behaviour of not knowing the destination into a robot. They compare their implemented method of walking side-by-side with a human to a velocity-based system. Consequently, they use the results of the human-human trial in the human-robot trial. Doering *et al.* [2019] investigates curiosity in a robot shopkeeper. They start with a human-human study where one person role-played as a customer and the other as a shopkeeper. They used this data to train a curious robot. They named this training the *offline learning phase*. Subsequently, they did a simulated evaluation of the curious robot. In this, they simulated customer interactions. The purpose was to investigate the effects of *online learning* in the curious robot. They had two conditions: 1) the robot did not adapt its responses based on the customer. 2) the robot did adapt its responses based on the customer. They evaluated this by analysing how well the robot predicted the customer and found that the adaptable robot had better prediction rates. They used this in the design of the curious robot in their user evaluation. In the human-robot interaction user evaluation, the participants role-played as customers in a camera shop with a robot shopkeeper. They compared a curious learner robot with an appropriate learner robot. As can be seen in Tables 10.2 and 10.3 both Murakami *et al.* [2014] and Doering *et al.* [2019] used within-subject design in their user evaluations Murakami *et al.* [2014] used a within-subject design in both their human-human experiment and in their human-robot experiment. In their human-human experiment, the participants took turns knowing the destination. The participants interacted with the proposed and velocity-based systems in their human-robot user evaluation.

2.2 The use of the data

Tables 10.1 to 10.3, found in Appendix 10.1, show that 19 of the papers use both qualitative and quantitative measures in their experiments, two use only quantitative measures, and one uses only qualitative measures. All the papers that used qualitative measurements used them to get self-reported information about either the robot they investigated or the setup in which the robot was investigated. Four of the papers using qualitative measures explain that they use items from other questionnaires; however, four different questionnaires. 15 of the 20 papers use Likert scales to measure qualitative items, one uses semantic differential scales, and four use other types of questionnaires or interviews. Of the 16 using scales to measure qualitative items, nine also use other types of questionnaires or interviews. When using scales, the reliability and internal consistency of the scales are essential. A way to measure is a Cronbach's Alpha. Of the 16 papers using scales, only three of them report a Cronbach's Alpha.

In terms of the analysis of the gathered data, four papers use t-tests to test for significant differences, of which one uses a one-tailed t-test. Eleven use Analysis Of Variance (ANOVA) to test for significant differences. Three use linear models, of which one is logistic regression. Eight use other types of analysis; for example, Totsuka *et al.* [2017] uses percentages and frequency. Other types of analysis could also be behaviours or statements by the participants, such as the analysis by Hanheide *et al.* [2017]. Four of the papers did not mention which methods they used. As can be seen from Tables 10.1 to 10.3, some of the papers used various methods to analyse their data, both in terms of several different statistical methods but also combinations of statistical methods and more qualitative approaches.

2.2.1 The use of Analysis of Variance (ANOVA)

When using ANOVA to analyse data, there are several assumptions the data needs to fulfill. Two of these are that the data is normally distributed, and the variance must be equal between the test factors. The assumption concerning equal variance can be investigated using Levene's test of Homogeneity in Variance, and the normality assumption can be investigated using Shapiro Wilk's test of Normality. For the normality assumption, it is, however, argued that if the sample size is big enough, e.g., above 30, the ANOVA will be robust enough even if the assumption is violated [Pallant, 2011, p. 204]. Only one of the papers using ANOVA describes using a non-parametric alternative. That is St-Onge *et al.* [2019], that counted 27 participants in a within-subjects design, see Table 10.1. Looking further into this guideline and the other papers using ANOVA to analyse their data, it is apparent that out of the 11 papers using ANOVA, only 4 have more than 30 participants. Three of these use between-subject designs. Two of the papers have under 30 participants in each condition, and for the third, two of four conditions have under

30 participants [Rouanet *et al.*, 2011]. As nothing is mentioned concerning the assumptions, it is presumed that the data does not violate the assumptions. When using ANOVA to test for significant differences, the F-value and the p-value show whether the test's null hypothesis can be rejected. The F-value offers convincing evidence to reject the null hypothesis if the F-value is high. Suppose, at the same time, the p-value is low, e.g., below 0.05. In that case, it gives further evidence against H0 [Agresti, 2018, p. 334], emphasising the documentation of both the F-value and the p-value when working with ANOVA. Unfortunately, 3 of the 11 papers using ANOVA do not report the F-value.

2.2.2 The use of Linear Models

When Logistic Regression is used to analyse data, the chi-squared from the model as well as the p-value must be reported [Pallant, 2011, p. 178]. The only paper in this literature review using logistic regression is Dragan & Srinivasa [2014]. They report the result from the logistic regression, from their first study, as follows: $\chi^2(1, 300) = 8.53, p = .0035$. They tested whether familiarisation impacts recognising motion, which explains the 1 in the parentheses. The 300 comes from 25 subjects with 2 different familiarisation conditions and 6 different test situations of robot motion. Therefore, they appropriately reported their logistic regression. As can be seen from Tables 10.1 and 10.2, the papers by Jayaraman *et al.* [2018] and Mavrogiannis *et al.* [2018] also used linear models. Mavrogiannis *et al.* [2018] used the linear model to find the relationship between a variable they call the *Complexity Index* and the time the participants spent choosing the right answer to which way two agents would pass each other. Mavrogiannis *et al.* [2018] explains the *Complexity Index* as "a proxy to quantify the complexity of multi-agent planning". Jayaraman *et al.* [2018] used the linear model to find the relationship between their variable of trust in AVs and the quantitative measures from their data, see Table 10.1. Neither of the papers reports the degrees of freedom in their models.

2.2.3 The use of T-test

As can be seen in Tables 10.1 to 10.3 four papers used t-test to analyse their data. When reporting the results of a t-test, it should include the test statistic, the degrees of freedom, p-value, the mean, and the standard deviation of the test [Pallant, 2011, p.240]. The test statistic tells us how far a point estimate is from the null hypothesis; ergo, the higher the test statistic, the better evidence against the null hypothesis. The p-value tells us the probability of the test-statistic being at least as large as the observed test-statistic [Agresti, 2018, p.155-156]. Therefore, it is important to report both the test statistic and the p-value. As for the ANOVA, the t-test also has assumptions the data need to fulfill to be adequate for the test. These assumptions also relate to normality and equal variance. None of the papers report

anything about these assumptions. In the result sections of Menne & Lugrin [2017] and Mavrogiannis *et al.* [2018], it is apparent that they report the test statistic, the degrees of freedom, and the p-value. However, both are missing presentations of the mean and the standard deviation. Furthermore, Javed *et al.* [2019] only presents the p-value. Kwon *et al.* [2020] reports the test statistic, degrees of freedom, and p-value, but only the mean and standard deviation in one of their user studies. In terms of participants Table 10.1 and 10.2 show that Javed *et al.* [2019] only had 18 participants, and that Kwon *et al.* [2020] only had 10 in their second user evaluation.

2.3 The interactions and the robots

This section will discuss the different types of interactions used in the papers, the different types of robots used, and how the setting of the user evaluation fits the context of the investigation. This literature review defines the setting as the setup of the user evaluation and the context as possible usage areas of the robots.

Some of the papers defined the robot they used as humanoid or non-humanoid. For the papers that did not, a google search was done on the robots to see descriptions or pictures of the robots. [Bartneck *et al.*, 2020, p. 49] describes that anthropomorphic designs of robots include different human-like characteristics: the appearance of the robot and the behaviour of the robot. An example of a human-like appearance is the robot DRC-Hubo+ used in Rakita *et al.* [2018] which has a face and two arms. An example of human-like behaviour is the robot Mini used in Javed *et al.* [2019]. The Mini robot is an iPod-based robot that expresses emotions in the study to teach children how to express negative emotions appropriately towards sensory input. Therefore, it arguably has anthropomorphic characteristics. These human-like characteristics will determine which type of robot the researchers use. See Table 2.2 for an overview of the robot types investigated in papers.

2.3.1 Settings outside the laboratory

Table 2.2 shows that five did their user evaluations outside of the laboratory and that these settings, to some extent, have characteristics resembling contexts the robots could be enrolled in [Totsuka *et al.*, 2017; St-Onge *et al.*, 2019; Hanheide *et al.*, 2017; Murakami *et al.*, 2014; Doering *et al.*, 2019]. Totsuka *et al.* [2017] investigated if utterances from a robot based on the visual scene would make the robot a better walking partner. They did this in an outdoor setting with two different visual scenes: a park and a parking lot. It can be argued that this setting resembles the possible contexts in which this type of robot could be used as the purpose was to investigate robot utterances compared to visual scenes. They chose to do their user evaluations in two different outdoors visual scenes. Murakami *et al.* [2014] investigated the design of motion of a robot not knowing the destina-

Reference	Robot used	Type of robot	Setting/context	Interaction
Menne & Lugrin [2017]	Reeti	Anthropomorphic Robot	Lab	Watched videos of others interacting with robot
Totsuka *et al.* [2017]	TEROOS-M	Human-like face	Outside: Parkinglot Park-area	Walked with robot on shoulder
Rakita *et al.* [2018]	DRC-Hubo+	Humanoid	Lab	Controlled robot while completing tasks
Gielniak & Thomaz [2011]	Simon	Upper-torso humanoid	Lab	Watched videos of robot motion, were to mimic the motions
Jayaraman *et al.* [2018]	Simulated Autonomous Vehicle (AV)	Non-humanoid	Lab: Virtual reality	Were to cross a virtual road, with AV approaching
Rouanet *et al.* [2011]	NAO	Humanoid	Lab: Designed as living room	Used assigned interface to point robot to football related objects
Kruse *et al.* [2014]	PR2	Humanoid	Lab: Designed as work environment	Walked in the work environment, were to interfere with the robot walking
Fitter *et al.* [2018]	Beam+	Non-humanoid Tablet on a wheel base	Lab: Robot in obstacle course	Teleoperating robot through an obstacle course
Javed *et al.* [2019]	Mini and Romo	Mini: Humanoid Romo: Ipod-based with humanoid behaviours	Lab With a table with the sensory stations	Went through the sensory stations, with the robots showing appropriate negative feelings
St-Onge *et al.* [2019]	Zooid	Non-humanoid	Dance studio	Dancing while zooid reacted live
Oudah *et al.* [2015]	NAO	Humanoid	Lab	Cheap-talk from robot partner, while solving game-tasks
Hanheide *et al.* [2017]	SCITOS G5	Human-like head	Care-house facility	Solved predefined tasks on info-terminal
Dragan & Srinivasa [2014]	HERB	Humanoid	First two: Online Third: Lab	First two: watched videos of robot solving tasks Third: standing next to robot while it solves tasks
Kwon *et al.* [2020]	Fetch	Industrial	First two: Online Third: Lab	First two: Participant filled online surveys to measure risk-aware behaviour, robot prediction simulated. Third: participants collaborated with robot in cup-stacking task
Mavrogiannis *et al.* [2018]	No explanation	Not known	Online	Participants watched videos of multi-agent trajectories, were to predict how two agents would pass each other
St. Clair & Mataric [2015]	Pioneer	Non-humanoid	Lab	Participants collaborated in a pseudo-herding task
Murakami *et al.* [2014]	Pioneer 3	Humanoid	Hallway	Participants walked side-by-side with a robot not knowing the destination
Nikolaidis *et al.* [2018]	HERB	Humanoid	Online	Table turning task with robot issuing different verbal commands
Jacq *et al.* [2016]	NAO	Humanoid	1: Therapist's office 2: Lab 3: therapist's office	All children taught robot handwriting
Gielniak & Thomaz [2012]	Simon	Upper-torso Humanoid	Lab	First watched the robot telling a story. Last changing motions to fit preferences
Doering *et al.* [2019]	3-DOF head two 4-DOF arms Wheel base	Humanoid	Camera shop	Participants interacted with robot, and answered questions from robot
Kwon *et al.* [2018]	Simulated PR2	Humanoid	Online	Watched different videos where a robot was incapable of solving a task

Table 2.2: The Table shows which robots the researcher used in their user evaluation, what type they are, the setting of their user evaluations, and which interactions the participants had with the robots in the user evaluations.

tion. They tested this in a hallway setting with two different destination options. As explained in Section2.1 Murakami *et al.* [2014] first did a human-human trial, then developed a system based on these results, and at last, tested their system in a human-robot interaction scenario; this suggests that the study was a "Proof of concept" study. Arguably, the setting was sufficient to prove the system worked. However, in terms of resembling possible contexts of use, the study is limited in its setting. As they test the robot in the same hallway in all conditions, it can be hard to generalise the results for other settings than that specific hallway. Both Totsuka *et al.* [2017] and Murakami *et al.* [2014] used robots with human-like features in their user evaluations, and both robots talked to the participants. This seems to be a good choice as the robots are companions to humans on a walk in both contexts. Hanheide *et al.* [2017], and Doering *et al.* [2019] both did user evaluations where the robots' purpose, to some extent, was to guide people. Hanheide *et al.* [2017] did a long-term evaluation of an info-terminal as well as a usability test in a care-house facility. The purpose of the info-terminal was that, e.g. residents, and visitors, could use the info-terminal to obtain information about, e.g., the menu at the facility or the weather forecast. The robot in their study had a humanoid face and a wheel-base, and the info-terminal itself was Ipad-based, which fits the requirements of the robot in this type of interaction quite well. However, a humanoid robot without an Ipad where the interaction was verbal between human and robot could also have fitted the requirements well. Doering *et al.* [2019] investigated how a robot shopkeeper should act with a customer. They chose to set up a camera-shop environment to do their user evaluation. The robot would ask the participants questions about cameras and gather information based on the answers to guide the participants in choosing the right camera. The setting fits quite well when the robot's context is limited to camera shops. In the user evaluation, they used a humanoid robot which fits the context quite well, as it, to some extent, could resemble a human shopkeeper. St-Onge *et al.* [2019] investigated robots' expressive emotions in a dance studio where they reacted live to performances by dancers. They used robots whose appearances were non-humanoid, but since they express emotions, their behaviour is anthropomorphic. Robots' expressive emotions can be relevant in different areas; an example of this could be robots reacting live to a person telling a story with expressive motions. Therefore, the setting is quite limited compared to possible application areas of expressive emotions.

2.3.2 Lab settings

Table 2.2 show that 11 of the papers, solely did their user evaluations in laboratory settings. In three of these papers, the participants observed the robots, of which two of them used videos of the robots [Menne & Lugrin, 2017; Gielniak & Thomaz, 2011, 2012]. Menne & Lugrin [2017] showed videos of two types of interactions

with a humanoid robot and recorded the facial muscular activity of the participants as they watched the videos, and collected self-reported emotions from the participants. No argument against the choice of the robot can be made as they mention that previous research has been done in this field of HRI, where the robot resembled a dinosaur, and that they wanted to investigate whether the feelings are similar when the robot is humanoid. As their purpose was to investigate whether the facial muscular activity could elicit emotions towards interactions with robots, it can be argued that the setting needs to be controlled. Therefore, the lab setting fits the study quite well. In both Gielniak & Thomaz [2011] and Gielniak & Thomaz [2012] the theme is robot motion. In Gielniak & Thomaz [2011] participants watched videos of robot motions and were to mimic the motions. They hypothesised that the more human-like robot motion is, the easier it is to mimic. As the purpose is to mimic motions, the lab setting is a good choice, as the researchers can control the user evaluation. Furthermore, the lab setting makes it possible to, e.g., record the participants when they mimic the robot. Gielniak & Thomaz [2012] investigates the differences between exaggerated and unexaggerated motions in a storytelling setting. After the robot told the story, the participants were to change the motions to fit their preferences. Given that the participants were to change the robot's motions in the last part of the user evaluation, the lab setting makes sense. However, for the storytelling part, other alternative settings could be used. For example, it could be tested with a bigger audience when the participants had to fill out a questionnaire about the story they heard after the presentation. This setting could also have resulted in more participants for their between-subject design.

Rakita *et al.* [2018] and Fitter *et al.* [2018] used the lab setting to investigate aspects relating to controlling a robot. Rakita *et al.* [2018] developed a method to control a robot' motion directly and compared it to other robot control methods. To investigate this, they used a humanoid robot which seems to be a good choice, as the robot is to perform motions done by a human. Fitter *et al.* [2018] investigated teleoperated robots and the effects of personalisation of such a robot. The robot was in a lab with an obstacle course. There are plenty of valuable contexts for such a robot, e.g., educational purposes. Therefore, it could be argued that the setting in the user evaluation could be made to resemble such contexts more. However, it seems to be a preliminary study on the effects of personalisation in teleoperated robots. With this in mind, a more controlled setting such as a lab fits for a proof of concept. Furthermore, using a tablet-based robot arguably makes personalisation easier.

Oudah *et al.* [2015], and St. Clair & Mataric [2015] both investigated the effects of communication in collaborative tasks in a laboratory setting. Oudah *et al.* [2015] did a preliminary study in which participants were not allowed to communicate with their assigned partner. Arguably, studies with this type of restriction need to be in a more controlled setting that a lab can provide to ensure that participants

follow the rules of the user evaluation. As mentioned in Section 2.1 in the main user evaluation, the participants were allowed to communicate with their assigned partner. Arguments can be made for and against choosing a lab setting when communication is allowed. As the participants were allowed to communicate with their assigned partners, the user evaluation was less restricted. Therefore, the user evaluation could also have emerged participants into a setting that fits this type of collaborative robot's application area. However, as it seems to be an exploratory investigation of the effects of cheap talk, the lab setting maybe provides more control over the user evaluation. Further works of cheap talk in collaborative tasks between robots and humans could benefit from a user evaluation done in situ. As mentioned before, the communicative nature of this evaluation makes the choice of the humanoid robot fit quite well into the context. St. Clair & Mataric [2015] investigated the effects of communication in a collaborative pseudo-herding task between a person and robot in a lab. Regarding the choice of the task, it could be argued that the pseudo-herding limits the generalisability of the results. As for the study by Oudah *et al.* [2015], the study by St. Clair & Mataric [2015] is also done in a lab setting. This choice makes the user evaluation easier to control by the experimenters, which could be an advantage for the study.

The remaining four studies in which the user evaluations were carried out in a lab did, to some extent, adapt the lab to fit their needs better. Jayaraman *et al.* [2018] emerged their participants into a virtual reality environment within their lab to investigate pedestrians' trust in autonomous vehicles. A virtual reality environment is arguably a good setting, especially considering that the setting can resemble the desired context quite easily, and settings are easier to change. Furthermore, if the user evaluation was possible to carry out outside of virtual reality, more ethical considerations should also be made, such as the safety of the participants. Rouanet *et al.* [2011] investigated interfaces to interact with robots. They chose a task where participants were to teach a robot about football by pointing the robot's attention toward football-related objects. They chose to adapt the lab setting to resemble a living room area. This choice seems reasonable, as it can be argued to be a good solution to have some of the advantages of an in situ evaluation and a more controlled lab evaluation. The participants might feel more at ease when the setting is homier, and the researchers still have the possibility of controlling the evaluation in the ways they want. Kruse *et al.* [2014] adapted the lab to resemble an office environment with people sitting at desks working during the user evaluation. Considering that the surroundings did not have to disturb the user evaluation, this adaptation was a good choice to ensure that the people working at the desks did not switch their attention toward the user evaluation, which could disturb the participant. Suppose the user evaluation was in a real office. In that case, people working in that office might direct their attention toward the robot if they were not used to being around robots. Javed *et al.* [2019] who investi-

gated how to teach children to have proper responses to negative sensory inputs, investigated this in a lab, where they set up the sensory stations. As their purpose was to teach these responses to children diagnosed with autism spectrum disorder, the setting could have been improved by being in, e.g., a therapist's office. Doing the user evaluation in a therapist's office would possibly resemble the application area of the robot better. Furthermore, it could be a more relaxing setting. Some of the children in the user evaluation were accompanied by their/a therapists. In terms of the choice of the robots, the humanoid robots fit the purpose quite well, as emotions/behaviours are anthropomorphic characteristics.

2.3.3 Online user evaluations

From Table 2.2 it can be seen that three of the papers solely did their user evaluations online [Mavrogiannis *et al.*, 2018; Nikolaidis *et al.*, 2018; Kwon *et al.*, 2018]. Mavrogiannis *et al.* [2018] investigated multi-agent trajectory prediction. As explained in Section 2.1, Mavrogiannis *et al.* [2018] compared their developed system in a simulation and used the online survey with videos to confirm that their system made predictable entangled trajectories. The videos they showed their participants were simulated work environments, and they did not explain which robot they simulated or whether it was a humanoid or not. It can be argued that predicting trajectories of robots is relevant to both humanoids and non-humanoids and that it could have been an advantage to do two different settings, one with a humanoid and one with a non-humanoid. They used simulated videos of multi-agent trajectories in their user evaluation. It could have been an advantage to use recordings of real-world scenarios where predictable multi-agent trajectories could be relevant such as Public places, office areas, and factories. As explained in Section 2.1, Nikolaidis *et al.* [2018] investigates different types of commands from a robot in a collaborative task. Table 2.2 shows that for this purpose, they used the humanoid robot HERB. This choice is reasonable as the robot gives verbal commands to a human, which is an anthropomorphic characteristic. The table-turning task they chose to use in their user evaluation also seems to resemble the context of a collaborative task quite well. Aside from doing an online user evaluation of the collaborative task, it could also have been an advantage to do a real-life investigation of collaborative table-turning done in a lab. Kwon *et al.* [2018] investigated how to express robot incapability so that people can elicit both why the robot is incapable and what its goal was. Table 2.2 shows that they used a simulated PR2 humanoid robot. Their paper shows that the videos they used in their online user evaluation were also simulated. However, the user evaluation setting could have benefited from videos from real-life settings. Moreover, it can also be argued that expressing robot incapability could be relevant to non-humanoid robots.

2.3.4 User evaluations in different settings

Some of the researchers used different settings for their user evaluations. Dragan & Srinivasa [2014] and Kwon *et al.* [2020] did their first two user evaluations online and their third in a lab. Jacq *et al.* [2016] did two of their user evaluations in therapist offices and one in a Lab. Dragan & Srinivasa [2014] used their online user evaluations to find information concerning familiarisation and naturalness of motion. The results were used in the final user evaluation to investigate how familiarisation and naturalness of robot motion affect comfort. Kwon *et al.* [2020] used the two online user evaluations to investigate the risk-aware behaviour of people to train the system they developed to predict risk-aware behaviour. They compared the system to a prediction method called noisy rational in a simulation based on the second online user evaluation. Their third user evaluation was a collaborative cup-stacking task between the two robots and the participants in a lab. In terms of the two first user evaluations by Dragan & Srinivasa [2014] and Kwon *et al.* [2020], it can be argued that the choice of doing these online is an appropriate setting to collect preliminary information and to use this information in a user evaluation in another setting. Dragan & Srinivasa [2014] chose a lab setting for their third user evaluation. Other options could have been to adapt the lab to resemble an industrial setting or a real industrial setting where robots are being used daily or might be implemented in the future to resemble the possible contexts better. As both Kwon *et al.* [2020] and Dragan & Srinivasa [2014] used different settings in their user evaluations, it is interesting to see how the data they collected differs in the three user evaluations. Table 2.3 shows an overview of the data they collected in their user evaluations.

	First online	Second online	Third lab
Kwon *et al.* [2020]	**Action Distribution** Accelerate vs. Stop	**Action Distribution** Stable vs. Unstable	**Efficiency:** time **Safety:** trajectory length **Four statements** about collaboration with robot
Dragan & Srinivasa [2014]	**Objective predictability:** Choosing the motion of the robot which matches their expectation **Subjective predictability:** Ranking of statements concerning the motion trajectory of the robot Subjective measures of utility of robot and motion attributes	Same as first	**Objective comfort:** how far from robot **Subjective comfort:** rating of one question about comfort in working side-by-side with the robot

Table 2.3: Table with an overview of the data collected by Kwon *et al.* [2020] and Dragan & Srinivasa [2014], in their three user evaluations.

Table 2.3 shows that both Kwon *et al.* [2020] and Dragan & Srinivasa [2014] col-

lected the same types of data in both of their online user evaluations. It also shows that the data collected in the user evaluations in the lab settings differ from the data collected in the online user evaluations. As mentioned in 2.1 Kwon *et al.* [2020] use the data collected in the first user evaluation in the training of the risk-aware robot system they developed. The data from the second user evaluation were used to compare the predictions of the risk-aware robot system with a noisy rational robot system. The Action Distribution measures how often participants choose to stop or accelerate in the first user evaluation. In the second it is a measure of how often the participants choose a stable or unstable tower. In both user evaluations, these action distributions are dependent on the time limit given for making a choice and the success rate of the different choices. The third user evaluation compares the two systems in a collaborative task. They compare them using three different measures: 1) Efficiency as the time spent on the collaborative tasks, 2) Safety as a measure of the trajectory length of the robot, where longer trajectories are safer than short trajectories, and 3) self-reported measures from the participants concerning their enjoyment in collaborating with the robot, how well they thought the robot understood their behaviour, the robot predicted the cups they would reach for, and how efficient they perceived the robot. Kwon *et al.* [2020] collected data in the online user evaluations objectively, whereas, in the third user evaluation, they used both objective and subjective measures. As mentioned in Section 2.1 Dragan & Srinivasa [2014] use the results of familiarisation and repetition from the two online user evaluations to design their user evaluation in the lab. Even though they investigate three different things in their three user evaluations, it can be seen from Table 2.3 that the procedure concerning measurements are alike. The two online user evaluations collect the same data, making them comparable. They compared to find the differences between motions created with their method and less natural motion. In the first two user evaluations, the objective predictability is measured by the participants first describing what motion they think the robot will make and subsequently revealing which of three different trajectories best matched their expectations. This was a measure of the accuracy of the motions compared to the participants' predictions. In their third user evaluation in the lab, the objective measure was the distance the participant chose to the robot. In all three user evaluations, the subjective measure evolved around the robot. The subjective measurement in the third user evaluation was asked both before and after the participants were familiarised with the robot motion. Before familiarisation) *"I would feel comfortable working side by side with the robot on a close-proximity task like cleaning up the dining room table."* After familiarisation) *Added to the original sentence* *"if it moved in the way I saw"*. Even though the purposes of the three user evaluations in Dragan & Srinivasa [2014] are different, they use the same structure of data gathering: First objective measures, followed by subjective measures. Quite the contrary to Kwon *et al.* [2020] who collect different data in their user evaluation in

the lab, compared to the two online user evaluations. Arguably, Kwon *et al.* [2020] could have used an action distribution, similar to the one in the second online user evaluation in their user evaluation in the lab as well, as participants in the labs also had to choose between building a stable or an unstable tower. They could have used this to get more evidence of the robots' abilities to predict the participants' choices and have compared these prediction rates with the self-reported measures from the participants. This could also have been compared to the results from the second user evaluation, where participants were only to choose a tower once, giving the robots only one prediction per participant, whereas, in the third user evaluation, the robots trained five times each for each participant.

Jacq *et al.* [2016] did their first and third user evaluation in a therapist's office and their second user evaluation in a lab. Furthermore, they used the humanoid robot NAO. Using NAO in their user evaluation seems to be a reasonable choice as the topic of the user evaluations was for children to teach a robot handwriting, which is arguably an anthropomorphic characteristic. The first two user evaluations only had one participant each, with particular problems in terms of handwriting. The first of these was done in a therapist's office, where the child had already received handwriting lessons. The second was done in a lab. Arguably it should also have been done in a therapist's office or the participant's home if the parents were the ones who usually taught the child handwriting skills. The third user evaluation was done in a therapist's office. The setting was not tailored to the individual child in this user evaluation, contrary to the first two user evaluations. Arguably, the setting could have been in the children's school, which could have resembled a context of using robots as teaching assistants.

2.4 The authors self-reported limitations of their studies

Some of the papers reported methodological, technological and analytical limitations. These limitations have been divided into five categories: 1) Generalisability Doering *et al.* [2019]; Kwon *et al.* [2018]; Totsuka *et al.* [2017]; Menne & Lugrin [2017]; Murakami *et al.* [2014], 2) Focusing on limited possibilities Kwon *et al.* [2018], 3) Lack of realism Jayaraman *et al.* [2018]; Oudah *et al.* [2015], 4) Sampling Javed *et al.* [2019], and 5) Investigating long-term interaction in a short-term setting Kwon *et al.* [2020].

2.4.1 Lack of generalisability

The limitations resulting in a lack of generalisability are the limitations that would influence the results of the papers in other settings. Doering *et al.* [2019] investigated curiosity in a robot shopkeeper in a camera shop. They report that the participants were instructed only to ask questions relevant to cameras resulting in

not testing the system they developed for the robot on a wider variety of topics. As explained in Section 2.1.2, they first trained their robot with an offline training method using the information from the human-human trial, and after that trained their robot with an online training method using the information provided by the participants. They report that a limitation of the robot is that it cannot act entirely based on the online training provided through interaction with humans. Totsuka *et al.* [2017] reports a limitation regarding the method they used in their experiments. The participants walked with the robot on their shoulders in two different surroundings in the experiment. One of these, a garden, was also used when they trained the robot to deliver utterances about the visual scene. Because of this, they report that the positive results of the system they developed for this purpose are hard to generalise as the scene is very specific in the experiment. Kwon *et al.* [2018] investigated how to express robot incapability. They report that some limitations to their approach are that incapability was only investigated for a limited amount of tasks. As mentioned in Section 2.1 the paper by Menne & Lugrin [2017] is an abstract, and therefore they had not analysed the data for other Action Units than pleasantness and unpleasantness. They report this as a limitation of the analysis in their abstract. This limitation of their analysis concerns a lack of generalisability. Murakami *et al.* [2014] investigated how robots should behave when walking side-by-side with a person. The system was based on an indoor setting, and it was tested in the same indoor setting in human-robot interactions. They mention that this might be a limitation since their developed system may be too simple for more advanced settings. St-Onge *et al.* [2019] reports that the time constraint they used in their user evaluation may have limited the amount of data they gathered. As it is hard to generalise the data they gathered from the constraint performance, this can be argued to be a limitation concerning generalisability.

2.4.2 Focusing on limited possibilities

When authors are focused on limited possibilities, the limitations concern restrictions of other factors which could have been interesting to investigate in their user evaluations. Kwon *et al.* [2018] also reports that robot incapability could be expressed in other ways, e.g., verbally. Dragan & Srinivasa [2014] who investigated the impact of familiarisation, mentions that there are many factors of familiarisation. However, they only investigated three possible factors of familiarisation.

2.4.3 Lack of realism

Lack of realism in the user evaluations restricts the results to specific settings. Jayaraman *et al.* [2018] did their experiment in a Virtual Reality Environment. In this environment, the behaviour of the Autonomous Vehicle was based on pedestrian behaviours. There would only be one person at a time and the crosswalk that the

participants were to cross was unidirectional. Therefore they report that a limitation of their experiment is that this setup might not elicit actual behaviours at a real crosswalk. Oudah *et al.* [2015] tested a system they developed whose purpose was to cheap-talk to a person who was to complete a task. They found that a robot providing cheap talk consisting of both feedback and planning made a better collaboration between the robot and a human partner. However, the system they developed did not consider what the person said during the collaboration. St-Onge *et al.* [2019] reported that the robots were sensitive to the body orientation of the dancers, which is a limitation concerning lack of realism, as body orientations will also vary in the real world.

2.4.4 Sampling

Sampling limitations are when the sampling in the user evaluations does not necessarily represent the population. To investigate how robots can help children express their negative feelings Javed *et al.* [2019] used two different groups of children: Traditionally Developing children and children with Autism Spectrum Disorder. Their analysis found heterogeneity in the two groups, which could explain the result of no significant differences in engagement between the two groups of children.

2.4.5 Investigating long-term interaction in a short-term setting

When authors investigate aspects of Human-Robot interaction, a possibility is that the data they collect is restricted to short-term usage. Kwon *et al.* [2020] developed a system to recognise risk-aware humans to make better Human-Robot Collaborations. They tested a collaboration in a cup-stacking experiment. They mention a limitation to their approach: the collaboration investigated a short-term scenario; however, modelling human behaviours could benefit from a long-term collaboration.

The effects of the limitations

These limitations reported by the authors have different effects on this literature review. In Section 2.3 it was described how the settings of the user evaluations fit the contexts that the authors wish to investigate. When the authors themselves report limitations concerning the generalisability of their user evaluations, in some cases, it revolves around the setting they investigate. Both Murakami *et al.* [2014] and Totsuka *et al.* [2017] report using the same settings in their user evaluation as they used in the training of their robots. Murakami *et al.* [2014] uses the same setting in the human-human trial they used to train the system they developed, as they did in the human-robot trial. Therefore, the results only account for this

specific setting, making the possibilities of using their results in later investigations more limited. Totsuka *et al.* [2017] only tested one of their training settings in their human-robot user evaluations. Therefore, contexts of using their system are more extensive than they are for Murakami *et al.* [2014]. Even though the specific setting of the user evaluation is important, it is also very important that the settings also resemble a wide variety of possible contexts.

St-Onge *et al.* [2019] reports a limitation concerning body orientation. Furthermore, they report that the time constraint in their experiment may have limited the amount of data they gathered — these two limitations combined lead to whether they gathered enough data from the dancers. Taking into account that St-Onge *et al.* [2019] is categorised as a One User Evaluation paper, these limitations could have been minimised by doing two preliminary investigations: One to investigate the constraints of the robots and one to investigate different time constraints' effect on their data. If this was done, it could lead to more assurance in their collected data. Another study that could have benefited from preliminary studies is Javed *et al.* [2019]. They report that their non-significant results may be because their two groups of children are too alike, even though the children in one group are diagnosed with Autism Spectrum Disorder. If they had used preliminary user evaluations to make a screening of possible participants, maybe they could have found two groups of children who reacted differently to their sensory stations.

2.4.6 Limitations not reported by the authors

As explained earlier in the paper by Murakami *et al.* [2014], they first investigated human-human interaction of walking side-by-side with only one of the agents knowing the destination. They did this to develop a system for a robot not knowing the destination when walking side-by-side with a human. However, the human-human interaction offered only two different destinations, and the people participating took turns knowing the destination. Consequently, the participants of the human-human trial would know that the destination could be one of two after the first rounds. Arguably, this is a limitation, as the people could have changed their behaviour of walking side-by-side based on their knowledge of the destination. To ensure that the destinations were unknown, they could have recruited more participants who only went through one or two rounds. Gielniak & Thomaz [2012], who investigated exaggerated motions in a storytelling setup, used a measure of fill-in-the-blank questions about the story the participants heard. Their description of the experiment states that when using this type of measure, to find differences in exaggerated and unexaggerated motions, it is required to either have a high failure rate of correct answers or a large number of participants. They used this measure in the first user evaluation, where they had 54 participants. Besides this, they show the percentages of correct answers to these questions in their results section. However,

they do not comment on whether they had enough participants or if there were enough wrong answers to the fill-in-the-blank questions. Therefore, the question still stands: do they fulfill the needs required to use this type of measurement?

2.5 Discussion

This section discusses aspects of the process leading up to the literature review and the findings.

2.5.1 The process of finding papers for the literature review

Chapter I.1 describes the process leading up to the literature review. Several different inclusion criteria were chosen; one was that the papers should have at least five citations. This criterion was introduced in an attempt to ensure that other authors, to some extent, acknowledge the papers in the literature review in their work. However, this has also resulted in very few papers from 2020 and 2021. The only paper included from 2020 was the paper by Kwon *et al.* [2020], and none from 2021. This criterion might have been too strict to include more recent research. Two options for solving this issue are: Either the citation limit should have been smaller, e.g., one for all papers included, or it could have been smaller for a specific range. Furthermore, the literature review might lack papers on, e.g., the development of scales for robotics, evaluations of robots who should not have direct interactions with people, and robot abuse. However, it is unknown where papers with such topics were excluded.

2.5.2 Developing systems

Section 2.1 showed that the majority of the papers included in the literature review developed a system for a robot. However, six of the papers that developed a system only investigated their system in a single user evaluation, of which five compared their system to another system. The literature review showed that simulations had three different purposes. Kwon *et al.* [2020] used the simulation to investigate the aspect of online learning when a robot interacts with a person, which arguably, enhanced their system in terms of developing good interactions with people. However, the systems developed by authors only investigating their system in a single user evaluation rely more on the participants' perception when developing systems than Kwon *et al.* [2020] did. An example of this is St-Onge *et al.* [2019] who investigated peoples' ability to recognise exaggerated motions. The argument here is that it is hard to simulate peoples' ability to recognise. Therefore, when developing systems to investigate technical aspects of robot design, a simulation can be helpful; however, user evaluations are more appropriate when the

development depends on peoples' perception. Another way to create systems is to conduct several user evaluations, where the first user evaluation creates a baseline, which was done by Oudah *et al.* [2015]. They created a system to generate cheap talk and investigated its effects, both between different kinds of cheap talk, cheap talk between two human partners, and no cheap talk.

2.5.3 The design of the user evaluations

The literature review showed that the authors used different subject allocations in their user evaluations; some used within-subject design, some used between-subject design, and some used a mix of the two. St. Clair & Mataric [2015] used a within-subject design in their investigation of communication in a pseudo-herding task with a robot. It can be discussed whether using a within-subject design was the right choice. Even though the interactions with the two robots were counter-balanced between participants, it can be argued that there might have been a carry-over effect. This carry-over effect can be problematic for the participants who first interacted with the communicative robot as they could use the information they gathered from the communicative robot when they interacted with the robot that was not communicative. Consequently, a between-subject design would possibly have been a more appropriate choice. Fitter *et al.* [2018] investigated the effects of personalisation in telepresent robots. The user evaluation compared a personalised and a non-personalised robot to each other. Due to recent times, where many students have endured a lot of online lessons, it could be interesting to investigate the effects of personalisation compared to another baseline: online lessons and exams. Using this type of baseline could give information on the effects of personalisation and how to solve some struggles with online lessons that students might have.

2.5.4 Multiple user evaluations

As explained in Section 2.1, Gielniak & Thomaz [2012] did two user evaluations: one where they investigated the effect of exaggerated motion on memorising a story, and one where they investigated participants' eye gaze during the story-telling. Furthermore, it was explained that the results of the first user evaluation were not used in the second user evaluation. An alternative could be to investigate the correlation between attention and memorisation from eye gaze and memorisation. Another paper that did not use the results from previous user evaluations in later user evaluations is Jacq *et al.* [2016], which investigated teacher-apprentice relationships to teach handwriting to children. Their first and third user evaluation both evolve around letters. Unfortunately, the data and results they present from their first user evaluation only revolve around the child's development. Suppose they had gathered information about the relationship between the increased difficulty and the examples presented from the child to the robot. It could have been

used to enhance the system to ensure that the difficulty progressed optimally. Another example of potentially useful data is about the setting they chose in the first user evaluation. They chose a setting where a robot was to write another robot who was on a mission. Such information about the setting could be compared to the child's engagement, which could have been used in the third user evaluation to enhance the setting.

2.5.5 Credibility of the analysis

Section 2.2 showed that some of the papers did not report important information about the analyses they used for their results. This limits the analysis presented, as potential readers have less information to deduce whether the differences are actually significant. Furthermore, Section 2.2 showed that of the papers using scales to measure qualitative aspects of their user evaluations, only three of them reported information regarding the consistency and reliability of the scales. This also limits the user evaluations as it cannot be deduced whether the scales did measure the intended.

2.6 Findings and Further works

The literature review showed that many aspects differ between the 22 papers. First of all, the use of simulations to test developed systems and the number of user evaluations in the papers are quite varied. Furthermore, the literature review showed that the researchers used different settings to evaluate robots. Some used online surveys with videos, some evaluated the robots in labs, and some in more ecological settings. The literature review also showed that both within- and between-subject designs are used in user evaluations of robots and robot systems. However, it also showed that maybe some researchers should have chosen another subject allocation than they did. Another aspect that varied throughout the literature review papers was the type of data they collected. The literature review showed that the researchers used strictly subjective methods, such as observations and interviews, methods of quantifying perception, such as questionnaires, and strictly quantitative measures, such as time and mistakes. It was also shown that the researchers also combined different methods. On the contrary, the literature review showed that some researchers did not provide sufficient information about the analyses they used for their data.

These findings from the literature review resulted in the desire to compare different data-collection methods to find out how the methods differ in concerns of the information they provide and the resources used for the methods. This comparison aims to provide guidelines on how to use different methods and at what aspects of robot design the different methods would provide information

that can be used in further development and validation of the design.

Part II

The user evaluation

Chapter 1

A user evaluation of data-collection methods in HRI

The literature review showed that researchers use different data-collection methods when investigating Human-Robot interaction. However, when researchers investigate aspects of robot design, few to no known guidelines are in place to help researchers choose what type of data they should collect from their user evaluations. Therefore, in this part of the study, a user evaluation investigating human-robot interaction will be designed to investigate different data-collection methods in the effort to answer the following question:

What are the advantages and disadvantages of using different types of data-collection methods in HRI?

Dividing this question gives the following specific questions which are more specific:

1. How do data-collection methods differ regarding the use of resources during preparation, the experiments, and the data analysis?
2. What can different data-collection methods deduce from the same interaction between a human and a robot?
3. How do the different data-collection methods supplement each other?

In this situation, resources is a measure of time spent on different parts of the user evaluation, such as:

- Time spent on preparing the method
- Time spent on data collection
- Time spent on analysing the data

Besides answering these questions, the study aims to develop guidelines for when, why, and how to use the different methods when investigating Human-Robot Interaction. When comparing data-collection methods have to be compared, the scenario should ensure that the different data-collection methods measure the same aspect of the interaction.

1.1 User evaluation scenario

To determine which scenario should be used for the user evaluation several options were considered, see Table 1.1.

	Description
Robot State	The robot indicates different states while solving a task
Proxemics	The robot passes the participant with different speeds and emotions
Robot Reacts	The robot reacts positively or negatively toward a participant solving a task
Story Telling	The robot tells a story, with different emotions and movements
Beer-pong	The participant and the robot plays a game of beer-pong as teammates

Table 1.1: Table showing the considered interactions.

The possible interactions in Table 1.1 can be categorised into three different groups:

1. Observing the robot; *Robot state* and *Story telling*.
2. Robot influencing participant; *Proxemics* and *Robot reacts*.
3. Robot and human collaborating; *Beer-pong*.

The interactions *Robot state* and *Story telling* have in common that the participant's task is to observe the robot, as it either solves a task or tells a story. The measurements for these scenarios could, for example, evolve around mistakes made in interpreting the robot's state or recollection of story details. The interactions *Proxemics* and *Robot Reacts* have in common that they investigate how the perceived emotions of the robot influence how the participants solve the task at hand. For the *Proxemics* this would be how far from the robot they decide to pass for different speeds and emotions of the robot. For the *Robot Reacts* this would be how many, e.g., mistakes they make in their task depending on how the robot gives the participant feedback. The interaction *Beer-pong* is categorised as the robot and human collaborating, as the robot and human have a goal of winning the game

together. Measurements for this interaction could, for example, be team performance. Another option is to measure the collaboration between the participants and the robot.

Arguably, the beer-pong scenario is the most versatile. The human-robot team has to collaborate to reach the desired goal of winning the game, and this operation depends on a set of decisions and processes. The decisions and processes are interdependent, and a diagram illustrates them.

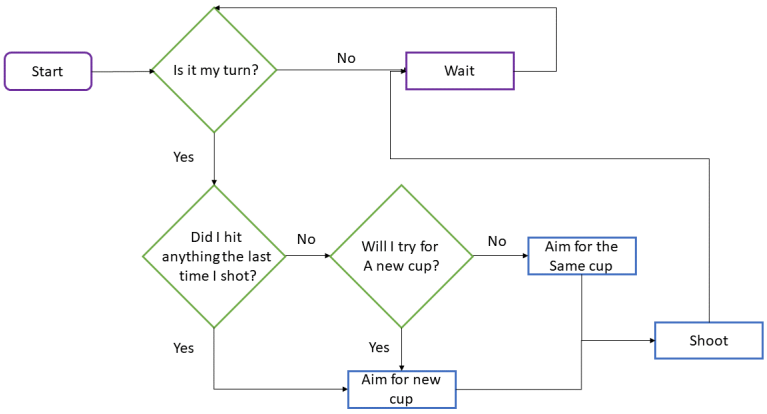

Figure 1.1: Diagram of the interdependent decisions and processes in a game of Beer-pong.

As can be seen from Figure 1.1, there are three decisions in each round throughout the game: Is it my turn, Did I hit anything the last time I shot, and Will I try for a new cup. Furthermore, four processes occur throughout the game: Aim for a new cup, Aim for the same cup, Shoot and Wait. From observing the game, the participants can make the decisions on the diagram. Typically the processes are internal. Another decision is to determine whether the game was won or lost at the end of the game. This decision typically does not lead to an internal process; contrary, it might lead to an external process of celebrating.

Designing the dynamic could be the human teammate making all the decisions from Figure 1.1 and telling the robot about these decisions so that the robot can solve the processes illustrated in Figure 1.2. Besides making all the decisions for the robot, the participant has to go through all the same decisions and processes of Figure 1.1 for themselves.

To be able to design the user evaluation, a robot needs to be chosen, designed, and programmed.

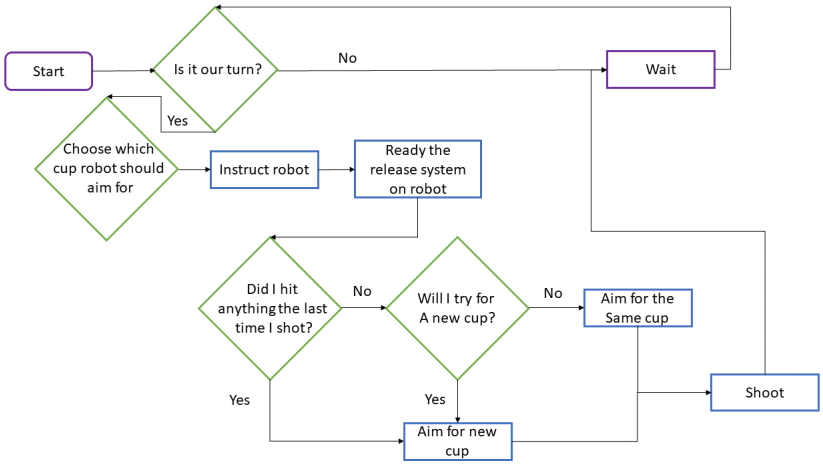

Figure 1.2: Illustration of the participants decisions and task during the interaction with the robot

1.2 Chosen robot

As the purpose of this study is to investigate data-collection methods in a beer-pong scenario, the criteria for the robotic platform are as follows:

- Short developing time.
- Has to be able to receive some input from a person.
- Has to be able to shoot a ball.
- Preferably cordless.

These criteria made the Fable robot from Shape Robotics[1] an appropriate robot for the user evaluation. It is a robot designed in Denmark that aims to teach pupils of state schools robotics and make the pupils interested in STEM education possibly. It is a module-based robot with different active components. Besides these active components, the robot also consists of passive components used to build different robot designs. This project will use the Fable Joint Module and the Fable Spin Module; see Figure 1.3 and 1.4 [2].

1.2.1 Fable Joint module

[1]https://www.shaperobotics.com/

[2]From the Fable User Guide from February 10th, 2022: v.1.4.3 https://www.shaperobotics.com/wp-content/uploads/2019/07/Fable-User-Manual-1.4.3.pdf

Figure 1.3: The Fable Joint module.

The Fable Joint module, Figure 1.3, is an active component consisting of two Servo Motors. These two motors make it possible to move the Joint module from -90° to 90° on both an X- and a Y-axis. Furthermore, the design of the Fable Joint Module's top, and bottom, makes it possible to connect it to other active and passive components.

1.2.2 Fable Spin module

Figure 1.4: The Fable Spin module.

The Fable Spin module, Figure 1.4, is also an active component. This component also consists of two motors connected to two wheels. Furthermore, it has a stabilising wheel, making it possible for the Spin module to drive forward and backward and spin up to 360°. Besides, the Spin module also consists of different sensors making it possible to detect, e.g., light, colours, and distance to an object.

1.3 Design and programming of the Fable Robot for the beer-pong scenario

Shape Robotics has developed a Python-based block program. When connecting a "Hub" to a computer, all active components of the robot can connect wirelessly to the Hub.

The Fable Robot also has different accessories. One of these is a glass fibre rod with a hole for a table tennis ball. The first part of designing the beer-pong robot was to investigate how far the robot could shoot using this rod. Connecting the rod to the Fable Joint Module and programming it to move the y-axis to -90°, wait a couple of seconds, and then move to -5°, made the robot shoot around 70 cm. The length of an ordinary table used for beer pong is 240 cm. Therefore, a solution was needed to make the robot shoot further. Connecting a hook to the Fable Joint Module and a release component to the glass fibre rod with a stabiliser to make it less flexible, the setup mimics a catapult. This setup made it possible to shoot around 170 cm. However, an inconvenience of this system is that the robot

cannot drag the catapult system into place; this has to be done by the participants. The hook and the release components were designed and printed on a 3D printer. Using the Fable Spin module makes the robot able to aim for different cups. Figure 1.5 shows the physical design of the robot when it is loaded.

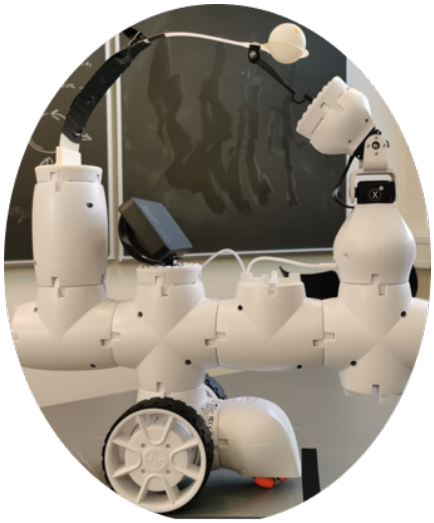

Figure 1.5: The physical design of the robot, in loading position.

As mentioned earlier, the Spin module has colour detecting sensors. These sensors can distinguish between 8 different colours: Red, blue, green, yellow, purple, turquoise, white, and black. However, as light is on when the sensors have to detect colours, it has some problems detecting white and black when the sensors are facing downward. When the light is on, they reflect on the spin module's surface, making it detect white even when nothing white is beneath it. The lights also make it hard for the sensors to detect black, as they make the black object reflect on the sensors making them not detect the black colour; this leaves six colours for the robot to recognise. Therefore, the beer-pong game will only consist of six cups per team. For the human-robot team, the six cups are each given a colour. This colour distribution is seen in Figure 1.6.

Locating the robot on a table beside the beer-pong table gave the best results. However, this also results in the wheels of the robot not having enough grip on the surface of the table—this influence the robot's ability to turn the correct amount of degrees. However, whether or not it turns the right amount or how much it varies from the right amount seems arbitrary. This shortcoming resulted in a design of the physical setup and programming of the robot, which limited the times the robot had to turn. Therefore, the program is designed based on the setup illustrated in Figure 1.6.

Figure 1.6 illustrates the robot's starting point related to the cups it is to hit. The slight angle of the starting point toward the cups the robot should hit means that the robot can hit the red, blue, and turquoise cups without turning. This angle means that the arbitrary differences from the wheels mean less than they would if the robot's starting-point would be parallel to the cups. For the robot to hit the remaining three cups, it has to turn 60°. The coloured stars in Figure 1.6 illustrate the robot's position when aiming for the six coloured cups. The robot's positions to aim for the green, purple and yellow cups mean that the robot has to turn left (from the robot's point of view). An explanation of the program for this is in the following section.

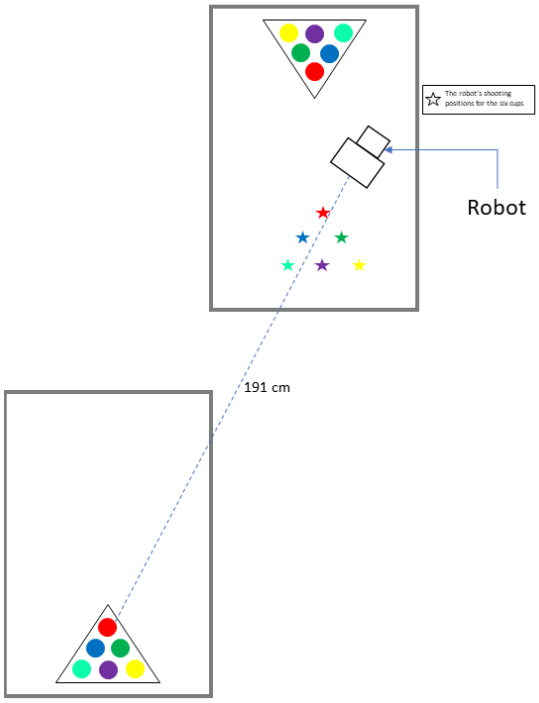

Figure 1.6: The setup from which the robot is programmed. The stars illustrate the six positions the robot has to drive toward, to aim for the cups in the same colours as the stars. From the robot to the edge of the red cup, the distance is 191 cm, the robot is programmed in relation to this distance.

1.3.1 Programming the robot

The program has nine different functions: one for each cup, two for turning the robot, and one for throwing the ball. The robot uses the throwing function for all of the six functions for the different cups, and the turning functions for the cups marked with green, purple, and yellow, in Figure 1.6. Consequently, the explanation divides the functions of the six cups into two different groups: The robot does not turn, and the robot has to turn.

Throw ball function

Figure 1.7: The code for throwing the ball.

The function of throwing the ball is only for the joint module of the robot. As can be seen from Figure 1.7, this function only consist of 6 lines. The y-axis on the joint module moves to $-45°$, waits five seconds, and moves to $-90°$, waits 2 seconds, and moves to $0°$. The reason for moving the y-axis to $0°$ is that if the motor were in the angle $-45°$ for an extended amount of time, the robot would shoot unexpectedly. Consequently, this left another degree of freedom that seemed somewhat challenging to take into account, in any other way than making sure that the motor is only in the $-45°$ angle for five seconds, for it to be the same for each round. This decision also had an advantage. As can be seen from Figure 1.2, the participants had to load the robot. When the Joint-module went into the $-45°$ angle, it served as an indicator that the robot was ready to be loaded.

Robot turning

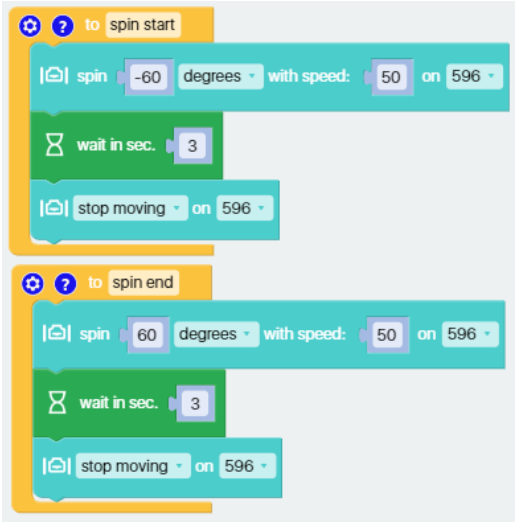

Figure 1.8: The code for turning the robot.

As can be seen from figure 1.8, two functions were made for when the robot was to turn. Figure 1.8 shows that each of the functions consists of three lines. The first is making the robot spin $\pm 60^\circ$, then waiting 3 seconds, and then stopping. When two commands were used right after each other, they would overlap in the execution. In this case, that means that the robot would stop before it had turned the $\pm 60^\circ$. Using the delay ensures that the robot would turn the right amount of degrees before stopping. This bug in the program, is also balanced for in other functions for the robot.

The functions for the cups where the robot drives straight

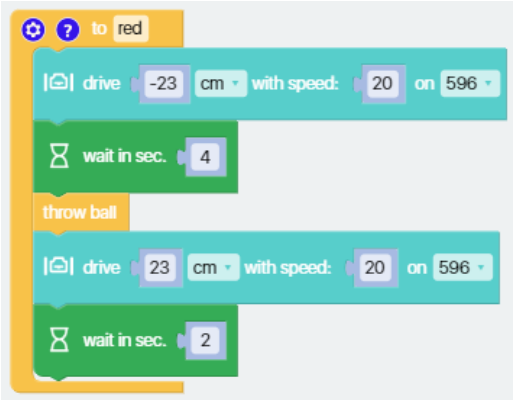

Figure 1.9: Code example from the function for hitting the red cup.

Figure 1.9 shows the code for when the robot has to hit the red cup from Figure 1.6. The functions of the blue and the turquoise cups are similar, except for the distances the robot has to drive. As can be seen from Figure 1.9, the robot drives -20 cm, throws the ball, and drives 20 cm back to its starting point. First of all, this means that the robot's starting point for the game is outside of the range from which it would hit the cup. Furthermore, it drives back to that starting point to ensure that the order in which the robot shoots for the cups is entirely up to the participants. Therefore, this is the case for all functions of the six cups. For the blue and the turquoise cups, see Figure 1.6, the only difference from Figure 1.9, is that the distances are different. These distances can be found in Table 1.2.

	Blue Cup	Turquoise Cup
Distance	±27 cm	±40 cm

Table 1.2: Distances for the Blue and Turquoise cup.

Manually finding the placements for the robot ensured a 30% success rate for the red, blue, and turquoise cups. Putting tape on the table ensured the robot programming had the proper distances for the three cups. Hereafter, the starting point was moved 20 cm from the placement for hitting the red cup.

The functions for the cups where the robot has to turn

The right side of Figure 1.10 shows the code for when the robot has to hit the green cup from Figure 1.6. The functions of the purple and the yellow cups are similar, except for the distances the robot has to drive. The Figure shows that all four functions are used twice in the code as the robot has to drive back to its starting point. Furthermore, the function *Drive 1* changes from when the robot drives forward to when it has to drive back. When it drives forward, the distance is 21 cm. When it drives back, the distance is only 20 cm. The distance also changes for *Drive 2*. Testing showed that the robot needed this change to be more precise at the starting point. Arguably, this change is necessary because the robot has to turn, as the change is not necessary for the cups where the robot does not turn.

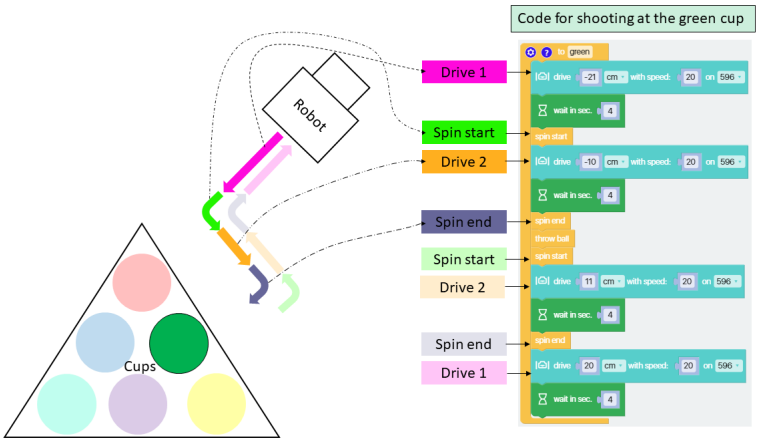

Figure 1.10: Code example from the function for hitting the green cup, and a visual representation of the functionality of the code.

Why the turns make this necessary is unknown; however, it could be because of the lack of friction mentioned earlier. As will be seen in Table 1.3 this change is also present in *Drive 1* for the Yellow cup, and in *Drive 2* for the Purple cup.

The left side of Figure 1.10 shows a visual representation of the functionality of the code example. As can be seen from Figure 1.10 the different parts of the function are visualised with arrows in different colours. The pink and orange arrows illustrate the robot driving. The light green and dark purple arrows illustrate the turns. Table 1.3 shows the distances for the purple and yellow cups.

Function	Purple Cup	Yellow Cup
Drive 1	± 27 cm	-21 cm/20 cm
Drive 2	-10 cm/9 cm	± 16 cm

Table 1.3: Distances for the Purple and Yellow cup.

The trial and error method also found the distances for the green, purple, and yellow cups. In Table, the functions Drive 1 and Drive 2, are shown. These functions relates to Figure 1.10.

1.4 The data-collection methods

The user evaluation consists of Subjective measures, Psychophysical measures, and Quantitative measures to compare data-collection methods of the interaction between the participants and the robot. The following sections will present the methods.

1.4.1 Subjective measures

The use of subjective data should depend on the scenario. For the scenario of this user evaluation, the decision stood between observations, think aloud, and interviews. When the purpose is to collect data about the interaction, the think-aloud method could influence other measures of the interaction, as the participants need to reflect on the interaction during the user evaluation. The interview could also influence other measures collected throughout the user evaluation. However, if the interview is the last part of the user evaluation, this problem is minimised. The advantage of observations is that they can reflect the participants' spontaneous reactions. Therefore, the subjective data collection will assess the interaction in two ways. The first is observation: how the participants react when different things happen in the game. Possible scenarios of observation are: *What is their body language/reaction when*

- They hit a cup
- Their teammate hits a cup
- They do not hit a cup, but their teammate does.
- They hit a cup, but their teammate does not
- Neither them nor their teammate hits a cup
- They win the game
- They lose the game

However, observation alone might not contain enough data about the participants' subjective perceptions of their own and teammate's performance in the beer-pong game. Therefore, an interview might be necessary to collect the desired data. Such an interview should debrief the participants after the user evaluation. It should consist of questions that all participants have to answer, but it should also encompass that some of the observations need explanation. A preliminary user evaluation should investigate the need for an interview and the questions within it.

1.4.2 Psycho-physical measures

To collect data in a psycho-physical manner, scales will be used. The literature review found that researchers often develop new scales for measuring when they use scales, sometimes leading to a long-developing process and many preliminary investigations of the scales. The basis of the scales in this user evaluation is from other researchers' scales.

Previously in this Chapter, it was determined that the interaction of the user evaluation in this project was considered a collaborative task. Therefore, the scales investigated for inspiration should measure aspects of collaborative tasks between people and robots.

Evaluating Fluency in Human-Robot Collaboration [Hoffman, 2019]

Hoffman [2019] measures *fluency* of a human-robot collaboration. Hoffman [2019] defines fluency as: The quality of the interaction in a shared activity. Hoffman [2019] mentions that *Fluency* is a separate construct than efficiency, as an interaction that is not efficient can still have good quality.

Table 1.4 shows questions about fluency in Human-Robot collaboration from Hoffman [2019].

Human-Robot Fluency	**Working Alliance for H-R Teams**
"The human-robot team worked fluently together"	**Bond sub scale**
"The human-robot team's fluency improved over time"	"I feel uncomfortable with the robot"
"The robot contributed to fluency of the interaction"	"The robot and I understand each other"
Robot Relative Contribution	"I believe the robot likes me"
"I had to carry the weight to make the human-robot team better"	"The robot and I respect each other"
"The robot contributed equally to the team performance"	"I am confident in the robot's ability to help me"
"I was the most important member on the team"	" I feel that the robot appreciates me"
"The robot was the most important team member on the team"	"The robot and I trust each other"
Trust in Robot	**Goal sub scale**
"I trusted the robot to do the right thing at the right time"	"The robot perceives accurately what my goals are"
"The robot was trustworthy"	"The robot does not understand what I am trying to accomplish"
Positive Teammate Traits	"The robot and I are working towards mutually agreed upon goals"
"The robot was intelligent"	**Additional**
"The robot was trustworthy"	""I find what I am doing with the robot confusing"
"The robot was committed to the task"	**Individual Measures**
Improvement	"The robot had an important contri-butionn to the success of the team"
"The human-robot team improved over time"	"The robot was committed to the success of the team"
"The human-robot team's fluency improved over time"	"I was committed to the success of the team"
"The robot's performance improved over time"	"The robot was cooperative"

Table 1.4: An overview of the questions about fluency in Human-robot collaboration [Hoffman, 2019].

The questions in Table 1.4 are all measured on a 7-point Likert scale. The boxes marked with blue and with text in **bold** are the headlines of the subscales in the fluency scale. The boxes beneath the blue boxes show the questions within the subscales. Furthermore, the statements in *italic* are statements asked in two different subscales. In the paper by Hoffman [2019] they investigate the perception of fluency in a human-robot collaborative task compared to objective metrics of a human-robot collaborative task. They investigate four different objective measures: 1) Human idle time, 2) Robot idle time, 3) Concurrent activity, and 4) Functional delay. The human idle time is the percentage of total task time the human partner in a human-robot collaborative task is inactive. Similarly, the robot idle time is the percentage of total task time the robot *appears* inactive. However, the robot idle time can be due to different things: Either the robot is inactive, or it only appears inactive but is, for example, processing data (internally active). The concurrent activity is the overlapping time where both the human and the robot agents are active. The functional delay is the accumulated ratio of the total task time and the time between one agent finishing an activity and the other beginning one [Hoffman, 2019].

Their investigation presents the participants with 5 of 50 different videos, where the objective metrics were changed. The participants were not familiarised with the clips before they were to evaluate them. After each video, the participants answered eight questions on a 7-point Likert scale. These eight were chosen based on the questions in Table 1.4. They chose only eight as the participants were to answer them after each video. They collected their data using an online survey.

Their investigation found a significant correlation between the perception of fluency and the objective metrics found that Human idle time. Furthermore, the functional delay was significantly reverse-correlated with the perception of fluency [Hoffman, 2019]. Even though they do not find significant correlations between the robot idle metric and the perception of fluency, they do find them to be consistently reverse-correlated [Hoffman, 2019].

Regarding the human-robot team performance in a beer-pong scenario, it could be interesting to investigate this reverse correlation between the robot idle time and the perception of fluency by expanding the visual processing time of the robot for different participants. Changing the visual time processing would also influence the functional delay, as the participants' action is to give the robot instructions, and the robot's task is to follow these instructions. Another way the robot idle time can be manipulated without manipulating the functional delay is to put in more delays between the different movements the robot makes after being given the instructions from the teammate.

Whose Job is it Anyway? - A Study of Human-Robot Interaction in a Collaborative Task [Hinds *et al.*, 2004]

Hinds *et al.* [2004] investigate what effects robot appearance would have on how much people rely on a robot they collaborate with and how much people are ceding responsibility to the robot. They investigate this by setting up a collaborative task between two agents. They manipulate one of the agent's appearance, using a human, a human-like robot, and a machine-like robot. The other agent is the participant.

They measure reliability and responsibility using scales, videotaping the interaction, and coding different interaction elements. Furthermore, they measure the perception of human likeness, which is outside this project's scope.

The scales used by Hinds *et al.* [2004] are as follows:

Human likeness	**Responsibility**
To what extent does the robot:	*To what extent did you feel:*
Have human-like attributes?	It was your job to perform well on the task?
Look like a machine or mechanical device?	Ownership for the task
Have characteristics that you feel you would expect of a human?	That your performance on this task was out of your hands?
Look like a person?	That good performance relied largely on you?
Have machine-like attributes?	Obligated to perform well on this task?
Act like a person?	**Attribution of credit**
Act like a machine?	Our success on the task was largely due to the things I said or did
Attribution of blame	I am responsible for most of the things that we did well on this task
I hold my partner responsible for any errors that we made on this task	Our success on this task was largely due to the things my partner said or did
My partner is to blame for most of the problems we encountered in accomplishing this task	My partner should get credit for most of what we accomplished on this task.

Table 1.5: The scales used by Hinds *et al.* [2004] to measure responsibility and attribution of blame and credit.

Besides using these scales, they also coded videotapes of the experiments. For example, they coded the videotapes for *shared social identity* by investigating the language the participants used in the experiment. They coded for *I, my, you* and *your* which is individualistic language, and for collective language such as *us, we* and *our* [Hinds *et al.*, 2004]. Besides this, they also investigated the mood of the participants to see whether this affected the results they got [Hinds *et al.*, 2004].

It could be interesting to use the scales concerning responsibility, attribution of credit, and attribution of blame concerning this project. In the scales concerning attribution of credit, they also investigate what the participants or the robots said during the user evaluation. As the robot in this user evaluation will be non-verbal

and cannot use verbal cues from the participants as instructions, the scales will only account for what the participant or the robot did during the user evaluation.

Scales to measure the perception of the interaction

In terms of the fluency scale by Hoffman [2019] using subscales 1 through 5 could measure the team performance. However, it can be seen from Table 1.4 that two statements are in more than one of the five subscales. The statement *"The human-robot team's fluency improved over time"* is used both in sub-scale 1 and 5. It can be hard to decide which of these two subscales is the better fit for this statement, as subscale 1 measures the human-robot fluency and subscale 5 measures the improvement, see Table 1.4. The statement *"The robot was trustworthy"* is in both subscales 3 and 4.

From the scale presented in Hinds *et al.* [2004] the subscale human likeness is removed. However, the subscales about responsibility, attribution of credit, and blame could be interesting for this study.

Common for the two scales is that they used 7-point Likert scales. However, Hoffman [2019] measures from *"Strongly disagree"* to *"Strongly agree"* and Hinds *et al.* [2004] measures from *"less"* to *"more"*. Arguably the subscales attribution of credit and attribution of blame from Hinds *et al.* [2004] can be measured from *"Strongly disagree"* to *"Strongly agree"* without any changes to the statements. However, in terms of the responsibility subscales, it would be necessary to rewrite the statements to a more first-person view. For example: from *"to what extent did you feel it was your job to perform well on the task?"* to *"I feel it was my job to perform well on the task"*.

Table 1.6 shows that these three subscales, as well as the five subscales of the fluency scale, result in the participants having to answer 24 questions. Arguably that is too many questions. Considering that the participants only answer them once in the user evaluation, the number of questions might not be problematic - a preliminary user evaluation should investigate this. As shown in Table 1.6, the wording of the questions was changed a bit to make the questionnaire consistent. "The robot" is changed to *"My robot teammate"*, and "the human-robot team" is changed to *"My robot teammate and I"*.

An Empathic Robotic Tutor for School Classrooms: Considering Expectation and Satisfaction of Children as End-Users [Alves-Oliveira *et al.*, 2015]

Another measure of interest in the user evaluation is a measure of expectation and satisfaction of the participants. Alves-Oliveira *et al.* [2015] has developed Technology-Specific Expectation Scale (TSES) and Technology-Specific Satisfaction Scale (TSSS) measures for their user evaluation. They investigated children's expectations toward a robotic tutor before they met the tutor and the children's related

From Hoffman [2019]	From Hinds *et al.* [2004]
Human-Robot Fluency	**Responsibility**
1: My robot teammate and I worked fluently together	*To what extent did you feel:*
2: The fluency of my robot teammate and I improved over time	**14:** It was your job to perform well on this task?
3: My robot teammate contributed to the fluency of the interaction	**15:** Ownership of this task
Robot Relative Contribution	**16:** That your performance on this task was out of your hands?
4: I had to carry the weight to make our team better	**17:** That good performance relied largely on you?
5: My robot teammate contributed equally to our team performance	**18:** Obligated to perform well on this task?
6: I was the most important member on our team	**Attribution of credit**
7: My robot teammate was the most important team member on our team	**19:** Our success on this task was largely due to the things I said or did
Trust in Robot	**20:** I am responsible for most of the things that we did well on this task
8: I trusted my robot teammate to do the right thing at the right time	**21:** Our success on this task was largely due to the things my robot teammate did
9: My robot teammate was trustworthy	**22:** My robot teammate should get credit for most of what we accomplished on this task
Positive Teammate Traits	**Attribution of blame**
10: My robot teammate was intelligent	**23:** I hold my robot teammate responsible for any errors, we made on this task
11: My robot teammate was committed to this task	**24:** My robot teammate is to blame for most of the problems we encountered in accomplishing this task
Improvement	
12: My robot teammate and I improved over time	
13: My robot teammate's performance improved over time	

Table 1.6: The sub scales from Hoffman [2019] and Hinds *et al.* [2004], which will be investigated in the preliminary user evaluation.

satisfaction toward the robotic tutor after they met the tutor. They developed the two scales with inspiration from the expectation-confirmation theory. The confirmation comes when the satisfaction of a product lives up to the expectations toward that product [Alves-Oliveira *et al.*, 2015]. In this case, the product is the interaction with a robot.

The scales Alves-Oliveira *et al.* [2015] developed to measure expectation and satisfaction, each consist of 10 questions. These questions are answered on 5-point Likert scales, where *1 = very low expectation/satisfaction* and *5 = very high expectation/satisfaction*. The questions used

Item	Technology-Specific Expectation Scale	Technology-Specific Satisfaction Scale
1	I think the robot will have superhuman capabilities	I think the robot had superhuman capabilities
2	I think the robot will be more than a machine	I think the robot is more than machine
3	I think the robot will be able to perceive what I am going to do before I do it	I think the robot was able to perceive what I was going to do before I did it
4	I think I will be able to interact with the robot	I think I was able to interact with the robot
5	I think the robot will be similar to the robots I see in movies	I think the robot was similar to the robots I see in movies
6	I think the robot will understand my emotions	I think the robot was able to understand my emotions
7	I think the robot will be able to recognize when I look at it or when I shift my gaze to something else	I think the robot was able to recognize when I looked at it or when I shifted my gaze to something else
8	I think the robot will have sense of humour	I think the robot had sense of humour
9	I think the robot will be able to understand me	I think the robot was able to understand me
10	I think the robot will be able to read my thoughts	I think the robot was able to read my thoughts

Table 1.7: The questions used in the Technology-Specific Expectation and Satisfaction Scales [Alves-Oliveira *et al.*, 2015].

Table 1.7 shows the questions from the TSES and the TSSS only differ in terms of the tense they are asked in: The TSES questions are future tense, and the TSSS questions are past tense, which makes them comparable.

Alves-Oliveira *et al.* [2015] investigates interaction in a tutor-student scenario, where a robot tutored children. Arguably the scales are specifically designed to evaluate the interaction of this scenario. The interaction in the present study has quite different characteristics, as explained in Section 1.1. Therefore, the specific questions used in the TSES and TSSS do not fit this user evaluation's interaction very well. However, the format of questionnaires can be used as inspiration to design a similar set of expectation vs. satisfaction scales.

Expectation and Satisfaction scales

One of the questions in the TSES and the TSSS fits this study. This is question 4, see Tabel1.7. However, more information regarding the expectation of the interaction with the robot is needed.

Dividing the statement *"I think I will be able to interact with the robot"* into different statements could make a more thorough investigation of expectation and satisfaction. Presenting the expectation questions to the participants after the initial briefing could also measure how well they understood the task. Based on the different sub-tasks of the participants explained in Section 1.1 the questions in Table 1.8 could measure the expectation and satisfaction of the participants.

	Expectation	**Satisfaction**
1	I will be able to communicate well with my robot teammate	I was able to communicate well with my robot teammate
2	I think my robot teammate and I will succeed at the task	I think my robot teammate and I did succeed at the task
3	I think I will be skilled at the task	I think I was skilled at the task
4	I think my robot teammate will be skilled at the task	I think my robot teammate was skilled at the task
5	I think I will be able to interact with my robot teammate	I think I was able to interact with my robot teammate

Table 1.8: Overview of the expectation and satisfaction questionnaires.

As can be seen in Table 1.8, statement 3 is not directly related to the interaction with the robot. Including the question could measure how skilled the participant believes they are at beer-pong beforehand, as it is a pretty common game in Denmark. Individual differences in skill level can influence the outcome of the game played with the robot. Furthermore, statement four measures the participants' initial impression of a robot teammate.

Rating these statements should also be on 7-point Likert scales, from *"Strongly disagree"* to *"Strongly agree"*.

The order of the questionnaires

It is essential to consider which order these questionnaires during the user evaluation. The expectation questionnaire should be before the game begins. However, an important question is whether the participants should answer the satisfaction or the questionnaire intended to measure the perception of the interaction should be presented as the first one after the game has ended. A disadvantage of answering the satisfaction questionnaire last is that the questionnaire about the interaction could make the participant reflect too much about the interaction they went through. Therefore, the satisfaction questionnaire will be presented as the first one

after the game is over to give more spontaneous ratings.

1.4.3 Quantitative measures

In the beer-pong scenario, the following quantitative measures will measure the performance of the human-robot team:

- How many times do they throw a ball
- How many times do they hit a cup
  - How many times do only one player on the team hit a cup
  - How many times do none of them hit a cup
  - How many times do both players hit a cup and is it the same cup.
- How long do they play (in minutes)
- Which of the teams won

These measures will make it possible to measure the performance of each of the participants and the robot, and the team performance. The count of how many times they threw a ball could indicate good interactions; the more throws, the better the interaction is. Moreover, as the interaction between the participant and the robot is the primary interest, additional objective measures should be used:

- How many interactions were there between the participant and the robot
  - How many of these were successful
  - How many of these were unsuccessful

In this case, *successful interaction* is the participant flawlessly indicating to the robot which cup it should aim for, as well as flawlessly preparing the ball for the robot. If any of these steps go wrong, the interaction is unsuccessful. This measure is arguably a more direct indication of the interaction between the participant and the robot than how many times they threw a ball.

In this study, the user evaluation will be a within-subject design, as all participants interact with the same robot and answer the same questions.

1.5 The rules of the beer-pong game

In a game of beer-pong, there are several rules. These rules concern: how to start the game, shoot, and the result of hitting a cup.

Concerning how the game starts, there are different rules. Some just choose a starting team, and if this team also wins, the opposing team gets one more shot

before the game is over. Some choose to do a mini-game to decide who starts - by one from each team having eye contact and shooting simultaneously. When one person hits a cup while having eye contact with one from the opposing team, that person "wins", and their team starts the game. Different opportunities exist for an opposing team in the beer pong game. One was to have one person play against the human-robot team; another recruited participant or the same person for all user evaluations. However, these two methods could result in variable user evaluation conditions for the human-robot team. First of all, ensuring the same skill level for recruited participants would be hard. Suppose it were to be the same person for all participants - this could result in that person getting better throughout the user evaluations, consequently making it harder for the last participant to win than it was for the first one. Therefore, the user evaluation will have a simulated opponent - meaning that the conductor of the user evaluations will remove cups for the "opposing team" during the user evaluation. A simulated opponent ensures that the skill level of the opposing team is consistent. As can be seen in Table 1.8 one of the questions for these two questionnaires is about succeeding at the task. Therefore, the simulated opponent will not "hit" more cups than the human-robot team - ensuring that even if the team does not win, they will still be better than the opposing team.

Concerning how to shoot, there are several sub-rules. First of all, one rule is that both teammates have to shoot one ball each time it is the team's turn. The order in which they do this is not essential in an ordinary beer-pong game. However, in the user evaluation, who shoots first can only be decided once to make the data collection easier. Therefore, the participant has to decide before the first shot who should shoot first, and then that order should be the same throughout the user evaluation. Another rule included in the user evaluation is that the participants have to shoot from a specific distance to the cups. As can be seen in Figure 1.6, the tables are not connected. Therefore, the participants have to shoot from a distance behind the cups on their table end. In an ordinary beer-pong game, there is a rule that if people's elbows cross the edge of the table at their end while shooting, the shot does not count, and the opposing team gets to shoot. However, the user evaluation will not use this rule - as there is no opposing team.

There are six different scenarios concerning the rules about what happens when the ball hits a cup. Table 1.9 shows the six scenarios.

A *Bounce* is the ball hitting anything (besides a cup) before hitting the cup, e.g., the table, a wall, or the floor. These will also be the rules of the user evaluation.

Different scenarios for hitting cups in the game	Rule
One cup hit	One cup is removed (after both opponents have shot)
Two cups hit	Two cups are removed.
Bounce in one cup	Two cups are removed (after both opponents have shot)
Bounce in two cups	Four cups are removed
Two hits in one cup	Two cups are removed, and the balls are given back to the team who just shot
Two *bounces* in one cup	Four cups are removed, and the balls are given back to the team who just shot

Table 1.9: Table showing the possible scenarios for hitting the cups in the beer-pong game and what happens with each of the scenarios.

Chapter 2

Preliminary User Evaluation

This Chapter aims to describe the preliminary user evaluation, including; the purpose, Its execution, and the results. Chapter II.1 describes aspects of the chosen data-collection methods that need investigation in a preliminary user evaluation. This Chapter describes the gathering of information about these aspects. Hereafter, it describes the structure of the preliminary user evaluation, with introductions and debriefings. Finally, the preliminary user evaluation results will be analysed, and a description of the usage of the findings from the preliminary user evaluation in the final user evaluation.

The order of the preliminary user evaluation will be as follows:

1. The participants will receive an introduction to the user evaluation.
2. The participants will answer the expectation questionnaire.
3. The participants will be introduced to how to interact with the robot.
4. The participants will play a beer-pong game with their robot teammate.
5. The participants will answer the satisfaction questionnaire.
6. The participants will answer the interaction questionnaire.
7. The participants will be debriefed.

2.1 1) The participants will receive an introduction to the user evaluation

First of all, the participants sign an informed consent form, See Appendix 10.2. To make sure that the participants understand the task at hand, they will receive an introduction to the task. However, the question is how thorough this initial introduction has to be. As the participants are to answer the expectation questionnaire

after the introduction, the introduction has to be vague enough for the participants to generate mental models of the robot without being biased by the information in the introduction. On the other hand, it has to be thorough enough for the participants to answer the expectation questionnaire. Another question is whether to introduce the participants to the robot before answering the expectation questionnaire. Therefore, the first thing this preliminary user evaluation will investigate is:

Should a picture of the robot be shown to the participants in the introduction?

Half of the participants see a picture of the robot before they answer the expectation questionnaire, while the other half will only receive the introduction. The preliminary user evaluation debriefing will investigate the participant's experience with the expectation questionnaire.

2.2 2) The participants will answer the expectation questionnaire

After the introduction, the participants answer the expectation questionnaire. While they answer the expectation questionnaire, the researcher will not intervene unless the participants ask questions about the questionnaire's wording. Suppose the participants have questions about the questionnaire - an analysis of these comments and the questionnaire can reveal whether it needs changes. The expectation questionnaire can be seen in Table 1.8.

2.3 3) The participants will be introduced to how to interact with the robot

After the participants have answered the expectation questionnaire, they will receive the interaction tools they need to interact with the robot. These tools are six cards in the different colors the robot reacts to and a map of which card corresponds to which cup. Furthermore, the participants receive an introduction to the placement of the sensors used to detect these colors on the robot. After these instructions, the participants can ask questions. When the participants are ready to begin, the game will begin.

2.4 4) The participants will play a game of beer-pong with their robot teammate

The investigation of the observation schema and the objective measures happens while the participants play. The purpose of investigating these is to determine whether unexpected interactions or reactions to the interactions happen during the task. However, after the user evaluation, this investigation happens by analysing recorded videos of the interaction and the task. Notes will be made during the task to ensure that the debriefing after the user evaluation can take foundation in the task. The notes taken during the investigation will revolve around the result of the game, as well as unsuccessful interactions. Consequently, the second thing investigated in the preliminary user evaluation is:

Are the observations of interest and the quantitative measures of interest sufficient to interpret the interaction?

An overview of observations of interest is in Section 1.4.1 and the objective measures of interest is in Section 1.4.3. Investigating this question relates to what happens in an ordinary game of beer-pong. The quantitative data collection happens during the user evaluation in the preliminary user evaluation, and the qualitative data collection happens through the videos. The user evaluation's primary purpose is to compare data-collection methods that measure the interaction between robots and people, meaning that the measures explained in Chapter II.1 are not sufficient. Furthermore, the preliminary user evaluation will also help determine *when* different data collection should happen.

2.5 5) The participants will answer the satisfaction questionnaire

After completing the task, the participants answer the satisfaction questionnaire. This questionnaire can be found in Table 1.8.

2.6 6) The participants will answer the interaction questionnaire

After the satisfaction questionnaire, the participants will answer a questionnaire about their interaction with the robot and the task they just completed. This questionnaire will henceforth be called the *interaction questionnaire*. As mentioned in Section 1.4.2, the questionnaire consists of 24 questions, and it should be investigated in the preliminary user evaluation whether this is too many questions. Therefore, the third thing investigated in the preliminary user evaluation is:

Does the interaction questionnaire consist of too many questions?

The debriefing and the participant's comments about the questionnaire will determine possible changes to the questionnaire.

Furthermore, the debriefing will also indicate the participant's perspective on the relevance of the questions used in the questionnaire.

2.7 7) The participants will be debriefed

For the preliminary user evaluation, the debriefing consists of three parts: Subjective information about the interaction with the robot, their experience with the different questionnaires, and a rounding of the user evaluation.

As explained in Chapter II.1, the measurements in the user evaluation all revolve around the interaction the participants had with the robot. The first part of the debriefing will revolve around the participants' subjective perspectives on their interaction with the robot. This preliminary user evaluation will determine what questions would be useful to asked the participants. Therefore, this part of the debriefing will be dialog-based. In the preliminary user evaluation, the debriefing consists of two premade questions:

- What do you think of the task you were given
- What do you think about the interaction you had with the robot

As shown in this Chapter, the preliminary user evaluation investigates several aspects. Therefore, in the second part of the debriefing, the participants will be asked questions about:

1. Their experience concerning the expectation questionnaire related to whether they saw the robot beforehand or not.
2. Their experience with the satisfaction questionnaire.
3. Their experience with the interaction questionnaire
    - The number of questions
    - The relevance of the questions
    - What their interpretation of the questions was

These questions are semi-structured - meaning that if participants have comments about these questionnaires, they need to elaborate on the comments to get a more thorough analysis of the questionnaires. Furthermore, the answers to the questionnaire will structure the debriefing.

Chapter 3

The Findings from the preliminary user evaluation

This chapter discusses the findings from the preliminary user evaluation. This discussion will consequently lead to the changes to the main user evaluation. The chapter will conclude with the structure of the main user evaluation and an overview of the data collection and analysis.

3.1 The participants

The participants in the preliminary user evaluation were all students from Aalborg University. Two study Engineering Psychology, one studies philosophy, one studies sociology, one studies Signal Processing, and the last studies Control and Automation - meaning that four of the six participants, all to some extent, know robotics.

Unfortunately, the camera did not record the experiment for the first participant - resulting in the lack of observations and not knowing which cup the participant aimed for in three of the shots.

The sixth participant saw the robot from the beginning of the user evaluation to investigate how this affected the questionnaires.

3.2 Findings

3.2.1 The robots ability to hit

Through these preliminary studies, the robot hit between 1 and 4 cups. An overview, of the cups the robot hit can be seen in Table 3.1

Amount of cups the robot hit	Number of participants
1	1
2	2
3	1
4	2

Table 3.1: A count of how many times the robot hit a cup during the 6 preliminary studies.

As explained in Chapter II.1 the robot has some different degrees of freedom, which are hard to control, resulting in these differences in how many times the robot hits a cup.

3.2.2 Qualitative measures

Observation

As mentioned in Section 1.4.1, the observational data consisted of looking at how the participants reacted when; they hit a cup, the robot hit, none of them hit, and when they won/lost. Looking through how the participants reacted during the preliminary user evaluation indicated that observations of hits might not be relevant to the interaction the participants had with the robot. However, it is relevant to the robot's and the participant's performance and comparing the number of cups both hit during the experiment. The observations of the preliminary user evaluation found that some of the participants looked surprised or confused when the robot turned for the green, purple and yellow cups. Furthermore, the participants were to move the robot if it did not return to the correct starting point marked with black tape. This sub-task of the interaction with the robot could also be interesting to investigate through observation. The preliminary user evaluation showed that some participants were more aware of where the robot stopped on its way back to make sure that they could move it to the correct point. Throughout the calibration of the robot before the preliminary user evaluation and the preliminary user evaluation, the robot would sometimes not follow the programming. For example, it would sometimes start driving without being shown a card. Furthermore, on some occasions, the robot would not drive after it had turned the first time when shown the green, purple or yellow card. These are unexpected situations for the robot, and the observations can investigate the participants' reactions to these situations.

Interview

As mentioned in Chapter II.1 and Chapter II.2 the participants were asked different questions after the experiment. The first question was about what they thought about the task given and their interaction with the robot. Hereafter, it was about the introduction given before the experiment and their thoughts on the expectation-satisfaction questionnaire and the interaction questionnaire.

Five of the six participants were mainly positive concerning the questions about the task and the interaction. They thought that the task was fun and exciting and that the interaction with the robot was intuitive. The last participant said that the task quickly became *"a "we" task"*, but was also optimistic about the task. From the question about the interaction, this participant mentioned that they needed feedback from the robot. When asked what type of feedback, they mentioned verbal feedback mainly concerning the robot's performance on the shot it had just shot.

In terms of the questions about the introduction and the questionnaires, all participants were positive. Before the preliminary user evaluation, a concern was that the interaction questionnaire was too long. However, the participants did not think that this was a problem. However, some participants got confused about the wording of some statements in the expectation-satisfaction questionnaire. One participant mentioned that the *"I think I will be skilled at the task"* was confusing, as it was not clear whether it was the task of playing beer-pong or the task of giving the robot the necessary information for it to play. As mentioned in Section 1.4.2 this statement was added to get an indication of the participant's perception of their skills when playing beer-pong. It might need rephrasing to ensure that the participants understand that this question concerns their skills in a beer-pong game. Three participants mentioned that it was difficult to judge whether they would be able to communicate with the robot concerning the expectation questionnaire, as they did not know how to yet. One of them mentioned that they expected verbal communication. Two of these participants did not see a picture of the robot beforehand. However, one of them had prejudiced expectations of the robot's abilities. One participant mentioned that they needed statements about the unison of the task. Another participant mentioned that it was hard for them to judge whether the robot was trustworthy, as they think trustworthiness relates to people and not machines. However, they chose to answer it concerning whether the robot understood the command they gave and gave its best shot.

The sixth participant, who had the present robot throughout the entire experiment, did not comment on this during the interview. This participant should have originally answered the expectation questionnaire without seeing the robot beforehand. Considering that the two remaining participants who did not see a picture of the robot said that it was hard to answer the question about communication, it could indicate that it is an advantage for the participants if they have a visual im-

pression of their robot teammate. Therefore, the participants should see the robot from the beginning of the main user evaluation.

3.2.3 Quantitative measures

As mentioned in Section 1.4.3, the quantitative measures were counts of the amount of successful and unsuccessful interactions between the participant and the robot.

During the preliminary user evaluation, participants primarily had successful interactions. However, a count of whether the participants showed the robot a card for a cup not present could also indicate unsuccessful interactions. An example could be that the red cup, see Figure 1.6, had already been hit and removed, and the participants still interacted with the robot using the red card.

The robot

After watching the videos of the interactions in the preliminary user evaluation, a question occurred: What did the participants think about the robot's speed? The question occurred as the videos gave the impression that the participants waited for the robot to drive for a long time, and some seemed somewhat impatient. After contacting the participants with the following question:

What do you think about the speed of the robot

Four of the six participants said that the robot drove slow. Two of these said that the robot was especially slow when they aimed for the green, purple and yellow cup, see Figure 1.6, as the robot had to turn as well. One of these elaborated by saying they changed their tactic to make the robot aim for the red, blue, and turquoise cups, as the robot did not have to turn. They said that this gave them the possibility of shooting more times during the experiment, and the robot was not very precise, so they did not think that it would mean anything as long as the robot aimed for the cups. Further analysis of which cups the participants aimed for, compared to the path of the robot, can be seen in Table 3.2.

Participant	Red, blue, turquoise (robot drives straight)	Green, purple, yellow (robot has to turn)	Unsure	Sum
1	12	9	3	24
2	17	6	0	23
3	7	13	0	20
4	13	9	0	22
5	15	6	0	21
6	5	6	0	11
Sum	69	49	3	121
Percentages	57,02 %	40,5 %	2,48 %	

Table 3.2: Table showing how many times the participants aimed for the two different groups of cups, the sum across the participants, and the percentages.

As can be seen from Table 3.2, the participants aimed for the red, blue, and turquoise cups more frequently than the other three. Even though this could indicate that the robot's speed and the delays should be changed, it was not. This decision was made based on the data collection. A slower robot could give some interesting observational data on the participants' body language. Furthermore, it could also investigate the tactic of getting the robot to primarily aim for the red, blue, and turquoise cups.

3.3 Structure of the main user evaluation

The first change to the main user evaluation is guiding the participants into the room where the study will take place before they get the introduction, sign the consent form, and answer the expectation questionnaire. The visualisation of the robot gives the participants an idea of the possible ways they could interact with the robot.

The second change from the preliminary user evaluation to the main user evaluation is that the guide for the colors of the cups is changed. Several participants from the preliminary user evaluation mentioned, that they had a hard time distinguishing between the green and turquoise cups on the guide. Therefore, the colors have been slightly changed. The changes shown in Figure 3.1 should minimize the

(a) Cup color guide from preliminary user evaluation. **(b)** Changed cup color guide for the main user evaluation.

Figure 3.1: The old and the new cup color guide. The pictures shows the differences between the color of the green and turquoise cups.

misinterpretation of the cup colors compared to the cards.

There was no specific order of the cards the participants used to interact with the robot in the preliminary user evaluation. They were mixed randomly by hand - meaning that the presentation of the cards was probably not truly different for the participants. For the main user evaluation, ensuring that all participants get different presentations of the cards is done with a randomisation tool[1]. The tool requires a number of participants. This number was set to 40 to ensure enough randomisations; however, it might not be possible to recruit 40 participants. The

[1] http://www.jerrydallal.com/random/permute.htm

goal is to recruit at least 20 participants. Table 10.4, shows the order of the cards for the main user evaluation, can be seen in Appendix 10.4.

The quantitative data collection happened during the user evaluations in the preliminary user evaluation, and the observational data collection happened through the recorded videos. The observational data collection will happen during the user evaluations for the main user evaluation. Moreover, the quantitative data collection will happen through the videos. Interpretation of the situation at hand could get lost when collecting the observational data mainly through video recordings; therefore, this change. The only quantitative data collection during the user evaluation is the order of hitting the cups and a timestamp for hitting the cups - as the recordings cannot cover both the participants, the robot, and the cups they shoot for in the same frame.

3.4 Data-collection

The different data-collection methods are described in Chapter II.1.

3.4.1 Collecting qualitative data

Concerning the collection of subjective data some changes has been made. For the observational data the following will be looked at during the user evaluations:

- Body language
  - Expression when shooting
  - Impatience due to speed of robot
  - Reactions to unexpected events
  - Reactions to the path of the robot
- Indication of their tactics
  - Are they aiming for specific cups?
  - Are they aiming for the cups where the robot has to drive less?

These are expectedly important to look out for during the experiment. However, other scenarios could appear during the user evaluations, which would also be important for analysing the data.

In concerns of the debriefing of the participants after the interaction questionnaire, the questions will be as follows:

- Questions also used in the preliminary user evaluation
  - *What do you think of the task you were given?*

  - *What do you think about the interaction you had with the robot?*

- Questions that will be added for the main user evaluation
  - *Did you have a specific tactic for interacting with the robot, for the goal of winning the game?*
  - Questions concerning observations, if anything stands out during the user evaluation

3.4.2 Collecting data with questionnaires

The only change to the questionnaire is that the participants answer the questionnaires online, limiting the time spent compiling the data. The participants will be answering the questionnaires on the experimenter's phone.

3.4.3 Collecting quantitative data

Besides the number of successful and unsuccessful interactions, explained in Chapter II.1, the following quantitative data will be collected:

- Number of cups hit
  - By the participant
  - By the robot
- The order of cups hit, as well as the timestamp
- Which cups are the robot told to aim for and timestamps for these.
- The number of shots from both the participant and the robot
- The time between shots.
- How many times did the participants correct the robot's starting point.

These additional quantitative measures will result in a more extensive evaluation of successful interactions with the robot. Furthermore, it makes it possible to make more comparisons between the data-collection methods. How many cups the participants and the robot hit during the user evaluation will also be used to divide the participants into three different groups for the analysis:

1. The participant hit more cups than the robot
2. The robot hit more cups than the participant
3. The robot and the participant hit an equal amount of cups.

3.5 Outline for the data analysis

This section will outline which methods used to analyse the data that was just described.

3.5.1 Analysing the subjective data

To analyse the observations and the interviews a thematic analysis will be used. It was considered to use a content analysis for the qualitative data. However, as this method quantifies the qualitative data, it was chosen to use a more qualitative method as other data will be collected and analysed quantitatively. Furthermore, from Tables 10.1 to 10.3, it can be seen that none of the papers that collect observational and interview data use content analysis. Instead, they extract essential observations and statements from their participants to support the other data. Therefore, the analysis used in this study should support the methods already used by researchers. The thematic analysis divides the data into different themes concerning the interactions the participants had with the robot. For this purpose, the observations and the interview will be transcribed during the user evaluations, and the program NVIVO will be used to code the data and find themes.

3.5.2 Analysing the expectation-satisfaction questionnaires

The analysis of the expectation and satisfaction questionnaires will be twofold. First, a t-test will investigate overall differences in the answers to the two questionnaires. Second, a Multivariate Analysis of Variance will investigate the differences between the three groups mentioned earlier and the differences between the two questionnaires.

3.5.3 Analysing the interaction questionnaire

The analysis of the interaction questionnaire will also be twofold. First, an Exploratory Factor Analysis will investigate possible factors of the questionnaire - as the interaction questionnaire is a compilation of two different questionnaires. Second, a One-Way Analysis of Variance will investigate whether the three groups affect the answers to the interaction questionnaire.

3.5.4 Analysing the quantitative data

The analysis of the quantitative data will use different methods: Correlation analysis of how many times the participants corrected the robot's starting point and the number of cups the robot hit. One-Way Analysis of Variance between the three groups explained earlier to analyse successful interactions. Analysis of Variance

between the three groups to analyse the number of shots. Furthermore, an analysis of the timestamps for aiming and hitting could give a quantitative indication of the participants' tactics. However, this is of low priority.

3.6 Comparing data-collection methods

The last part of the analysis of the user evaluations will consist of comparing the different data-collection methods subjectively to answer the questions presented at the beginning of Chapter II.1.

Chapter 4

Analysis of the subjective data

As mentioned in Section 3.5.1 the analysis of the subjective data consists of thematic analysis based on the observations from the user evaluations and the participant's answers to the debriefing interview, see section 3.4.1.

section The thematic analysis of the qualitative data Thematic analysis is a qualitative data analysis method. Braun & Clarke [2006] reports a step-by-step guide for using thematic analysis, consisting of the following six steps:

1. Familiarisation with the data
2. Generating initial codes
3. Searching for themes
4. Reviewing the themes
5. Defining and naming themes
6. Producing the report

There are different approaches when doing a thematic analysis. The analysis can be inductive or deductive. An inductive approach means basing the themes on the data. The deductive approach means basing the analysis on preconceived ideas of themes, which can both be personal experiences or theories. Furthermore, thematic analysis can also use either a semantic or latent approach. A semantic approach means using the data quite literally. The latent approach means looking into subtext and assumptions of the data [Braun & Clarke, 2006].

This thematic analysis uses a latent inductive approach to finding codes and themes within the data.

For the thematic analysis, the data were read through to get an overview and to brainstorm ideas for codes. After that, the data were read through several times to find elements fitting specific codes. After defining all the codes and arranging data

into the codes and sub-codes, they are organised into different themes. The themes were named and looked at again to determine whether they fit. Table 4.1 shows the final themes, codes, and sub-codes from the observations and interviews. The entire data set can be found in the supplementary work for this project.

Theme	Code	*Sub-code*
Interaction	It did what I asked	
	Understandable	
	Is it verbal?	
	Fun	
	Missing control	
	Missing feedback	
Time	Time management	*Impatience*
	Optimizing	
Negative feelings	Disappointment	*In self*
		In robot
	Nervousness	
	Trust	*Not trusting*
Invested in robot	Focused on robot shot	
	Emotional investment	
Comments on the game	Positive comment	*Liking game*
		Positive self assessment
		Impressed by robot
	Negative comment	*Self ability*
		Robot ability
	Talking about robot	
	Talking to robot	
	Simulated opponent	
	Whose turn?	
	Not winning	
	Thinking aloud	
	Questions about user evaluation	
	Comments for conductor of the user evaluation	
Aiming	Cup aim	*Specific cup*
		Aiming for straight line
		Same cup multiple times
		Importance
		Trying all colors
		Random colors
		Account for unpredictability
	Analysing aim	*Contemplating aim*
	Correction	*Contemplating importance*
		Compensating for last shot
Robot path	Unexpected event	*Robot not reacting*
		Robot slow
		Confused over loading time-frame
	Path recognition	*Confused over robot turning*
		When should it be loaded?
		Loading ready quicker than expected

Table 4.1: Themes, codes and sub-codes found in the observations and interviews, through a latent inductive thematic analysis.

4.1 The themes of the thematic analysis

As can be seen from Table 4.1, the thematic analysis revealed seven themes from the data. The theme *Interaction* consists of codes about the comments the participants had about the interaction with the robot during the debriefing interview. The theme *time* consists of codes about how the participants interacted with the robot with the time frame of the user evaluation in mind. The codes for this theme are from the observations of the user evaluation and the debriefing interview. The theme *Negative feelings* consists of codes concerning the negative feelings the participants had during the user evaluation. These codes primarily came from the observations, where participants both showed their feelings through body language and comments during the user evaluation. The theme *Invested in robot* consists of observations and comments from participants concerning. Throughout the user evaluations and the debriefing, the participants had several other comments; therefore, the next theme is *Comments on the game*. Some of the comments were positive, and some were negative. In the preliminary user evaluation, see Chapter II.3, the participants had different tactics for the robot - this became apparent in the main user evaluation as well. Therefore, the next theme is called *Cup aim* and consists of how the participants made the robot aim. Furthermore, the theme consists of questions the participants had, comments for the conductor, and how they talked to the robot during the user evaluations. The last theme found is *Robot path* and consists of the different reactions the participants had to the robot's trajectory and the loading mechanism throughout the user evaluations.

4.1.1 The theme: Interaction

As can be seen from Table 4.1, the theme Interaction consists of six codes:

- *It did what I asked*
- *Understandable*
- *Fun*
- *Missing control*
- *Missing feedback*
- *Is it verbal?*

As mentioned in Section 3.4.1, the participants had to answer what they thought about the given task, as well as the interaction they had with the robot during the task. The theme *Interaction* is primarily from the participant's answers to these questions. Of the codes for this theme, the code *Understandable* has the most coverage in the data (4.36 % of the interview data). The comments from the participants

are about the colors for the interaction and that it was easy. For example, one participant says:

(1) *"It was simple giving it tasks. It was easy. Intuitive. (Danish: Det var simpelt at give den opgaver. Det var nemt. Intuitivt.)"*

Another participant said that the robot was very clear in its interaction with the participant:

(2) *"It was relatively straightforward in its communication of what it wanted to do. If it could not recognise the colour, it would not do anything, and it was evident in its communication of when it wanted the ball. (Danish: Den var rimelig tydelig til at komunikere hvad den ville. Hvis den ikke kunne kende en farve så gjorde den ikke noget, og den var tydelig i hvornår den gerne ville have bolden.) "*

Another code in this theme revolves around the feedback from the robot. Some participants mentioned that they wanted more feedback from the robot. For example, one of the participants had the following comment about the interaction they had with the robot:

(3) *"There weren't a lot of interaction. It can't exactly talk to me. I just had to put it back into place, show it a color and put a ball in. It can't make it's own decisions, so the interaction was pretty bad. When you play with people they can cheer for each other, the robot didn't really care about that. (Danish: Der var ikke meget interaktion. Den kan jo ikke snakke med mig. Jeg skulle bare sætte den på plads, og vise farve og sætte bolden i. Den kan ikke selv tage en beslutning, så det var en ringe interaktion. Når man spiller med mennesker kan man heppe på i hinanden, det var robotten lidt ligeglad med.)"*

The codes *Fun*, *It did what I asked* and *Missing control* consist of only one reference each, and the name of the codes explains the references.

The walk-through of the data showed that some participants were unsure of how to interact with the robot:

(4) *""How am I supposed to tell it what to aim for?" At first I thought I had to talk to it, and then I thought "Oh no", because of Siri and such. (Danish: "Hvordan i alverden skal jeg fortælle den hvad den skal sigte efter?" først tænkte jeg at jeg skulle snakke og så tænkte jeg åh nej, pga. Siri og sådan noget.)"*

This comment relates to the expectation questionnaire the participants had to answer before the user evaluation began. As can be seen from Table 1.8 the first statement is about the participants being able to communicate with the robot.

4.1.2 The theme: Time

As can be seen from Table 4.1 the theme *Time* consists of two codes and one subcode:

- *Time management*
  - *Impatience*
- *Optimizating*

These codes were found by how the participants interacted with the robot and what they answered during the debriefing. The code *Time management* consists of the thoughts the participants had during their interaction. Several participants used the time when the robot returned to its starting point to shoot themselves. Some even commented on this during their interaction:

(5) *"Can I shoot while it resets? that is more time-efficient. (Danish: Må jeg skyde mens den resetter? Det er mere time efficient.)"*

Another participant said the following when asked whether they had specific tactics during the debriefing:

(6) *"Yes, I had to be done shootinng fast, so that I could utilise the time the best. So when the robot was back I should be done throwing. (Danish: Ja, jeg skulle være så hurtig færdig med at kaste min bold, så vi kunne udnytte tiden bedst muligt. Så når robotten var tilbage skulle jeg være færdig med at kaste.)"*

In terms of the code *Impatience*, it consists of references from both the observations and the debriefing. Some participants mentioned that they thought the robot was slow during the debriefing. One specified this the following way:

(7) *"For the ones to the left [green, purple, yellow] it was slow as it had to turn(...) (Danish: Dem til venstre var den langsom til fordi den skulle dreje(...))"*

Through the observations, several of the participants showed signs of impatience through their body language, for example, by tapping their foot or playing with the ball while the robot drove. Some participants even expressed their impatience verbally; for example, one participant expressed during the interaction:

(8) *"I am just standing here, waiting. (Danish: Jeg står bare og afventer.)"*

The code *Optimizing* consist of both statements about how the participants optimized the number of times the robot shot, as well as how they could have optimized. The following statement is an example of a participant explaining how they should have optimized:

(9) *"I should have aimed more for the cups to the left [green, purple, yellow]. I should also have optimized how much I tried the different colors to see when the robot hit well. Furthermore, I figured out late how long it would take for the robot to be ready to shoot. The more shots, the more possibilities for hitting, so I figured that I would get more throws if the robot were quickly ready to throw again. (Danish: Jeg skulle have sigtet mere til venstre, for at den kunne ramme flere kopper. Jeg skulle også have optimeret hvor meget jeg prøvede de forskellige farver, for at se hvornår den ramte godt. Det var også sent jeg fandt ud af hvor lang tid der gik for at robotten var klar til at skyde. Jo flere skud, jo flere muligheder for at ramme, så jo hurtigere den er klar, jo oftere kan jeg skyde.)"*

The statement indicates that this participant reflected on how changing tactics could optimize the number of cups hit during the user evaluation.

4.1.3 The theme: Negative Feelings

As can be seen from Table 4.1, the theme concerning negative feelings observed throughout the user evaluations have three codes and and three sub-codes:

- *Disappointment*
  - In self
  - In robot
- Nervousness
- Trust
  - Not trusting

The references found for the code and sub-codes concerning disappointment are from the observations' data. The code *Disappointment* relates to the general disappointment observed from the participants during the user evaluation. These signs of disappointment ranged from the participants' body language when something happened to verbal outbursts. Examples of body-language indicating disappointment could be looking down or shrugging their shoulders. Verbal outbursts indicating disappointment could be like the following examples from the code:

(10) *"Shiiiiiiit" when they almost hit (Danish: "Piiiiiiis"* or *"So close" (Danish: "Så tæt på"*

The code of disappointment has two different sub-codes as well. These are named *In self* and *In robot*. The references for these sub-codes have the same characteristics as those for the general code of disappointment; however, the participants'

outbursts or reactions were straight after they had shot themselves or straight after the robot had just shot. The reactions indicating disappointment relate to the shot and the one who shot. The sub-code *In robot* consists of 36 references from the data, and the sub-code *In self* consists of 35 references from the data.

The code *Nervousness* primarily concerns the path of the robot combined with the robot being on a table. Some participants indicated nervousness about the robot driving off the table both throughout the user evaluation and in the debriefing interview. An example of a reference from this code is:

(11) *"Please do not drive off the table. (Danish: Please ikk' kør ned af bordet.)"*

Furthermore, the code of nervousness is also about being nervous when the robot starts to drive or whether the robot would slip in the indication card.

The last code of the theme is *Trust* and its sub-code of *not trusting*. These two are, for the most part, inter-related as participants often used both of them in the same sentence, for example, as follows:

(12) *"I think that in the beginning I did not have a lot of trust in the robots ability to hit anything, but it turns out that it could. (Danish: Jeg tænker at i starten havde jeg ikke så meget tro til at den kunne ramme noget, men det kunne den så.)"*

The beginning of the sentence indicates not trusting the robot's abilities at first glance, but after interacting with it, changing their perception to trusting the robot's abilities. However, some participants also only indicated not trusting the robot; for example, one participant said the following during the debriefing:

(13) *"If it had been a dangerous tool, I would have been afraid to use it, as it doesn't tell when it is going to let go. However, it isn't dangerous, so I was mostly stressed in the beginning. Had it been knives I would have been afraid. (Danish: Hvis det nu havde været et farligt værktøj ville jeg være bange for at bruge den fordi den ikke fortæller hvornår den slipper. Men det er jo ikke så farligt, så jeg var mest bare stresset i begyndelsen. Hvis nu den kastede med knive ville jeg være bange.)"*

This nervousness is also related to the sub-code previously explained, *Missing feedback*.

4.1.4 The theme: Invested in Robot

Throughout the user evaluation, some participants seemed very invested in making the robot shoot and its ability to hit cups. Therefore, this theme consist of the codes *Focused on robot shot* and *Emotional investment*. Although these codes do not consist of many references, they are important as they indicate the relationship

between the robot and the participants. For the code *Focused on robot shot*, one of the participants stated the following in the debriefing:

(14) *"I might have been more focused on it [robot] to shoot than shooting myself. (Danish: Jeg fokuserede måske mere på at få den til at skyde end at skyde selv.)"*

Another participant said the following during the user evaluation:

(15) *"I want it [robot] to hit so much that I forget to shoot. (Danish: Jeg vil så gerne have at den rammer, at jeg glemmer selv at skyde.)"*

The other code in this theme; *Emotional investment*, consists of answers for the debriefing. One participant cheered during the user evaluation and was asked about it later - resulting in the following statement:

(16) *"It became a game, where one starts cheering for the robot when it starts hitting. Furthermore, I am a person who thinks aloud a lot. But yes, one starts cheering. (Danish: Det bliver lidt en leg til sidst, hvor man begynder at heppe på robotten, når man finder ud af at den rammer. Og så som person tror jeg også jeg er en der tænker meget højt. Men ja, man kommer til at heppe.)"*

Another participant also cheered a lot during the user evaluation. Furthermore, this participant also named the robot. Therefore, the participant was asked about the personality of the robot - resulting in the following answer:

(17) *"It is a little shy, it's like "Ohh, Ohh, I am not that good" but when it really counts I knew it would come through. And it did. [...] (Danish: Den er sådan lidt sky, den er sådan lidt "aaaargh, aaaargh, jeg er ikke særlig god" men når det virkelig gælder så vidste jeg at den ville come through. Og det gjorde den. [...])"*

4.1.5 The theme: Comments on the game

Throughout the user evaluations and the debriefing, the participants had other comments about the game. These comments resulted in the codes and sub-codes shown for this theme in Table 4.1. As there are many codes for this theme, this section will only explain a few. These are positive comments, negative comments, Talking about robots, and Talking to robots. As can be seen from Table 4.1 the codes of Positive comments consist of three sub-codes. For the code *Positive comment*, most of the references are from the observational data. These primarily consist of outbursts when the robot, e.g., hits a cup. One of the participants also had this comment during the debriefing:

(18) *"But I was happy to have it on the team, as it hit. (Danish: Men jeg var glad for at have den på holdet, fordi den ramte jo.)"*

Several of the participants were impressed by the robot for different reasons. Some were hopeful when the robot hit; some were impressed by the consistency, and some were impressed by the robot's ability to recognise colors. In concerns of the sub-code *Liking game*, the majority of the participants said that the given task was fun. For example, one participant said:

(19) *"It is fun that it is a known game. (Danish: Det er sjovt at det er et spil man kender.)"*

Only two of the participants commented positively about their own abilities. One of these said the following:

(20) *"It was great, as it incorporated something I knew I was good at. (Danish: Det var fedt fordi de inkorporerede noget jeg vidste jeg var god til.)"*

The primary reason the participants had *Negative comments* was due to the ability of the robot. These comments are primarily about the functionality of the robot. For example, it needed much help, the aim was not that good, it had some struggles with recognising some of the colors, and it did not hit.

As mentioned, another part of this theme is the codes *Talking about robot* and *Talking to robot*. These codes show a difference in how participants talked during the user evaluation. The *Talking about robot* code consists of language using, e.g., "it". For example, one participant said:

(21) *"It is going to miss so bad now. (Danish: Den kommer til at skyde så meget ved siden af nu.)"*

On the other hand, the sub-code *Talking to robot* consists of language using, e.g., "you" and "we". The following show examples of both types:

(22) *"I have faith in you, robot, I can feel that you are going to hit again. (Danish: Jeg har tiltro til dig robot, jeg kan mærke at du rammer igen.)"*

(23) *"Okay, one left, are we ready for a victory? (Danish: Okay, en tilbage, er vi klar til en sejr?)"*

4.1.6 The theme: Aiming

During the user evaluation, different strategies for making the robot aim for cups became apparent through observation. Some of these strategies were also investigated further during the debriefing after the user evaluation. The sub-codes for the code *Cup aim* all about how the participants used the cards. Some participants also switched between the different methods. In concerns of the sub-code *Account for unpredictability* one of the participants said the following:

(24) *"I the beginning, I aimed for the center, ad it gave a bigger possibility of hitting a cup. (Danish: I starten sigtede jeg lidt i midten, fordi der så er større chance for at ramme en kop.)"*

Furthermore, some participants analysed the aim of the robot. For example, some would go behind the robot when they had loaded it to see if the general direction of the aim would result in a successful shot. The code also has a sub-code named *Contemplating aim*. This consist of references where participants were investigating the guide and the cups and when they seemed unsure of which cup they should get the robot to aim for. The code *Correction* have a similar sub-code: *Contemplating importance*. For example one participant stated:

(25) *"I figured that it [robot] would hit better if I did it [correct]. (Danish: Jeg havde en idé om at den ville ramme bedre hvis jeg gjorde det.)"*

The other sub-code is *Compensating for last shot* which consists of observations and answers to the debriefing, where participants indicated that they would use the same card multiple times and correct the starting point of the robot dependent on how it shot the last time.

4.1.7 The theme: Robot path

As mentioned in Chapter II.1, the robot had different paths for the different cups it was programmed to hit. Sometimes during the user evaluations, different events would happen. This theme is an overview of the participants' reactions to these events. Sometimes, the robot would drive without being shown a card, recognise a card wrong, or shoot badly. The reactions for these types of events are in the code *Unexpected event*. For example, the robot recognised the green card when shown the turquoise card, and the participant stated:

(26) *"It read the color wrong. (Danish: Den læste farven forkert.)"*

When the robot shot weird, the reactions were either confusion or laughing at the robot's abilities. Another unexpected event was when the robot would not drive back to its starting point correctly. When this happened, it led to reactions such as:

(27) *"Does it think it is home now? (Danish: Synes den selv den er hjemme nu?)"*

This code also consists of sub-codes. For these sub-codes, the participants showed that these events were no longer unexpected after some familiarisation with the robot. However, for some participants, it would result in some frustration toward the robot. For example, when the robot would not react to a card, a participant said the following:

(28) *"That is annoying. (Danish: Det var da irriterende.)"*

The other code of this theme is *Path recognition*, which furthermore consists of three sub-codes. The sub-code path recognition is also a temporary event, minimised when the participants are familiar with the robot. When they seemed familiarised, more of the observations would go into the code *Path recognition*. From the observations, it became apparent that most of the participants did not expect the robot to turn when it had to aim for three cups, but after a couple of tries, they learned that it would take longer for the robot to be ready to be loaded. However, the opposite also occurred, where participants aimed for, e.g., the red cup and did not expect the robot to be ready for loading so quickly - usually after they had aimed for the "turn" cups.

Chapter 5

Analysis of the data from the questionnaire

This chapter will analyze the data from the expectation, satisfaction, and interaction questionnaire.

5.1 Analysing the expectation and satisfaction questionnaires

As mentioned in Section 1.4.2 the expectation and satisfaction questionnaires were answered on 7-point Likert scales from *Strongly disagree* to *Strongly agree*. To illustrate the raw data collected throughout the user evaluations see Figure 5.1.

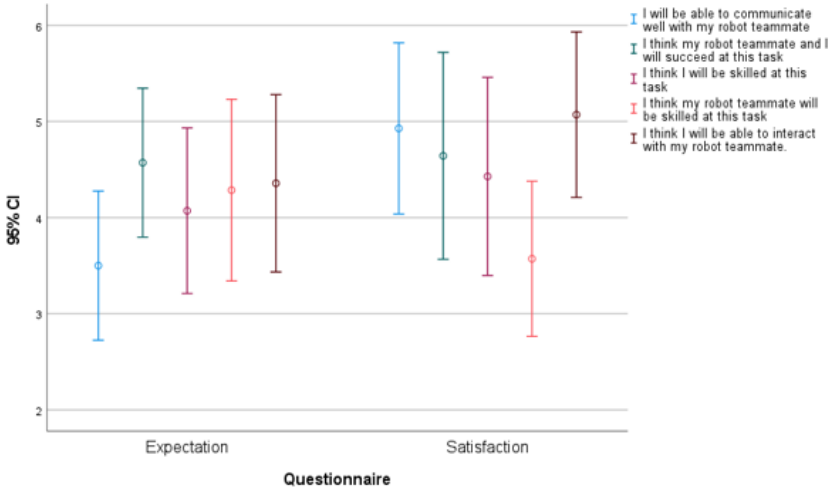

Figure 5.1: Error-bar plot of the means from the two questionnaires, across participants.

Figure 5.1 is an error-bar plot of based on the means across participants. The error bars are 95 % confidence intervals. Another aspect considered before the analysis is the internal reliability of the questionnaires. However, as there are only eight

sub-scales and only 14 participants, it was decided to look at the mean inter-item correlations of the sub-scales instead of a Cronbach's alpha. For the expectation questionnaire, the mean inter-item correlation (referenced to as MIIC) is .422, and for the satisfaction questionnaire, it is .275. According to Pallant [2011] the MIIC should be between .2 and .4, however, according to Glen, S. [n.d.] the MIIC should be between .15 and .50. As the MIIC of the expectation questionnaire is only .022 above the first guideline, and within the second, arguably both questionnaires give consistent and appropriate results [Glen, S., n.d.].

As mentioned in Section 3.5 it was planned to analyse these questionnaires by doing an overall paired sample t-test to compare the answers to the two questionnaires. Secondly, the participants were to be divided into groups depending on how many cups they and their robot teammate hit during the user evaluation and used to analyse the questionnaires with a multivariate analysis of variance.

5.1.1 Analysing with the paired sample t-test

To be able to analyse the data from the expectation and satisfaction questionnaire with a parametric test, several general assumptions need to be met [Pallant, 2011][p. 203-204]:

- The data type needs to be either interval or ratio
- The sample needs to be random
- The sample needs to be independent of each other
- The difference between the pairs of data should be normally distributed

Even though the data is collected using a 7-point Likert scale, the consequence of the limited number of participants is that interval data cannot be assumed. The data can be ranged, but the distance between data points is unknown. The more participants, the easier it is to assume interval data from Likert scales. Assuming that the data is ordinal, the non-parametric test Wilcoxon Signed Rank test will be used instead of a Paired Sample t-test.

5.1.2 Analysing with the wilcoxon signed rank test

For the non-parametric test wilcoxon signed rank, only two assumptions needs to be met:

- The sample needs to be random
- The sample needs to be independent of each other

The data from the expectation and satisfaction questionnaires all meet these assumptions.

Contrary to the t-test, the Wilcoxon signed-rank test does not compare means. The Wilcoxon signed-rank method organises the data into ranks. In this case that means a count of how many participants rated the questions from the satisfaction questionnaire higher than the questions from the expectation questionnaire *(positive rank; satisfaction>expectation)*, vice versa *(negative rank; satisfaction<expectation)*, and how many times they rated the questions equally *(ties; satisfaction=expectation)*. Table 5.1 shows the ranks of the test.

		N	Mean Rank	Sum of Ranks
SQ1-EQ1	Negative Ranks	1	2	2
	Positive Ranks	8	5.38	43
	Ties	5		
	Total	14		
SQ2-EQ2	Negative Ranks	6	7.5	45
	Positive Ranks	7	6.57	46
	Ties	1		
	Total	14		
SQ3-EQ3	Negative Ranks	4	6.25	25
	Positive Ranks	7	5.86	41
	Ties	3		
	Total	14		
SQ4-EQ4	Negative Ranks	9	6.56	59
	Positive Ranks	4	8	32
	Ties	1		
	Total	14		
SQ5-EQ5	Negative Ranks	4	4.88	19.5
	Positive Ranks	7	6.64	46.5
	Ties	3		
	Total	14		

Table 5.1: The Table shows the ranks when the data from the Expectation questionnaire is compared to the Satisfaction questionnaire. In the Table S = satisfaction, E = expectation, Q=question, Number = number of question (see Table 1.8 for reference).

In Table 5.1 *N* is the number of cases in the rank, *Sum of Ranks* is the sum of the data-points that match the rank, and *Mean Rank* is the sum of ranks divided with the number cases. These ranks investigate whether there is a significant difference between the questions of the two questionnaires.

Table 5.2 shows that only the ratings of *"I will be able to communicate well with my robot teammate"* and *"I was able to communicate well with my robot teammate"* was significantly different. This difference means that the participants rated significantly higher on the communication question in the satisfaction questionnaire($median = 5$), than they did in the expectation questionnaire($median = 3.5$), $Z = -2.459^b, p = 0.014$ with a medium effect size of 0.46.

	SQ1-EQ1	SQ2-EQ2	SQ3-EQ3	SQ4-EQ4	SQ5-EQ5
Z	-2.459[b]	-0.035[b]	-0.725[b]	-0.957[c]	-1.224[b]
Asymp. Sig. (2-tailed)	**0.014**	0.972	0.468	0.339	0.221

Table 5.2: Results of the Wilcoxon signed rank test. b = based on negative ranks, c = based on positive ranks.

5.1.3 Dividing the data-set

During the data collection, it became apparent that the participants hit more cups than the robot in some of the user evaluations. Therefore, the data-set for the satisfaction questionnaire could be divided into two groups, depending on whether the participant hit more cups than the robot or not, to investigate whether this significantly affects the satisfaction score. Section 3.5.2 shows that the method that should have been used was a MANOVA. However, the limited number of participants changed this to a Mann Whitney U only investigating the satisfaction questionnaire. The group sizes for the comparison of the satisfaction questionnaire can be seen in Table 5.3

Group	Group size
Participant >robot	9
Participant $\leq$ robot	5

Table 5.3: The table shows the two groups of participants, dependent on whether the participant hit more cups than the robot.

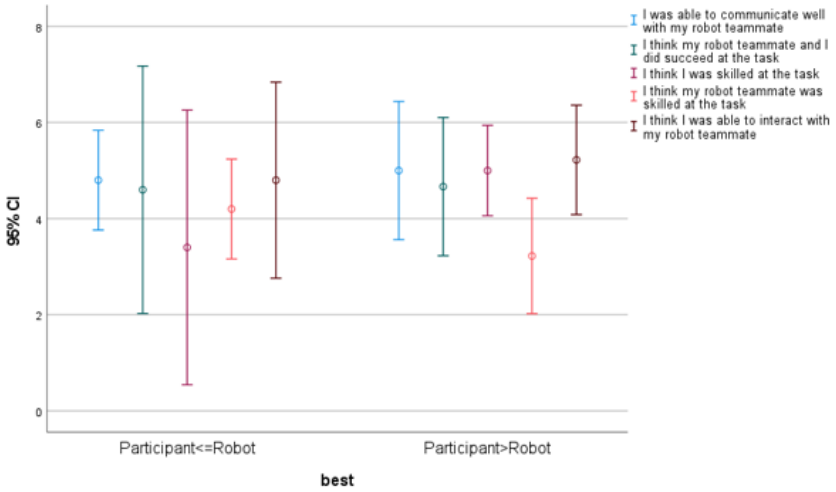

Figure 5.2: Error-bar plot based on the two groups mean values to the satisfaction questionnaire. The error-bars are made from 95 % confidence.

Furthermore, Figure 5.2 shows the data from the satisfaction questionnaire based on means between the two groups.

Table 5.4 shows the ranks of the Mann-Whitney U.

		N	Mean Rank	Sum of Ranks
SQ1	P≤R	5	6.2	31
	P>R	9	8.22	74
SQ2	P≤R	5	7.5	37.5
	P>R	9	7.5	67.5
SQ3	P≤R	5	5.2	26
	P>R	9	8.78	79
SQ4	P≤R	5	9.10	45.5
	P>R	9	6.61	59.50
SQ5	P≤R	5	6.7	33.5
	P>R	9	7.94	71.5

Table 5.4: Table shows the ranks for the satisfaction questionnaire, using the Mann-Whitney U test.

In Table 5.4 the column *Sum of Ranks* shows the sum of all the ratings for the two groups, and the *Mean Rank* is the sum divided by the number of data points in the groups. The Ranks investigate whether the satisfaction scores are significantly different between the two groups. Table 5.5 shows the results of the significance test. As can be seen from Table 5.5 the two groups did not rate the questions significantly different based on the current data. However, it is interesting that the p-value for the second question in the satisfaction questionnaire is $p = 1$. Furthermore, the median for this question is $M = 5$ for both groups indicating that

this question is not dependent on whether the participant hit more cups than the robot. Therefore, an additional significance test is done for this question. The data showed that 5 of the participant won the game within the 15-minute time frame. The ratings for question 2 are compared between the participants who won and those who did not. The Mann-Whitney U revealed a significant difference between the participants who won ($Md = 6, N = 5$) and those who didn't ($Md = 3, N = 9$); $U = 8, z = -1.997, p = 0.048, r = .5$.

	SQ1	SQ2	SQ3	SQ4	SQ5
Mann-Whitney U	16.5	22.5	11	14.5	18.5
Z	-.898	.000	-1.552	-1.102	-.563
Asymp. Sig. (2-tailed)	.369	1	.121	.270	.574
Effect Size	0.2	0	0.4	0.3	0.2

Table 5.5: Table showing the Mann-Whitney U significance test, for the satisfaction questionnaire.

5.2 Analysing the interaction questionnaire

The initial thought on how to analyse the data from the interaction questionnaire was to perform an exploratory factor analysis, see Section 3.5.3. However, different elements of the data made this not possible. First of all, some of the questions had negative eigenvalues. Second, a loading plot showed that the data was spread widely across the factors - indicating that factor analysis is not appropriate for the data.

Another idea for analysing the data from the interaction questionnaire was to divide the participants into the same groups, as shown in Table 5.3, according to whether the participants hit more cups than the robot or not. Furthermore, the analysis should have been an independent sample t-test. Unfortunately, according to a Shapiro Wilks test of normality, the data was not normally distributed. Therefore, the non-parametric alternative Mann-Whitney U should be used. Before analysing the data from the interaction questionnaire, it is essential to consider whether all of the questions load in the same direction. As for the data from the interaction questionnaire, 11 questions were inverted before the analysis:

- *I had to carry the weight to make our team better*
- *I was the most important member on our team*
- *I felt it was my job to perform well on this task*
- *I felt ownership if this task*
- *I felt that my performance on this task was out of my hands*
- *That good performance relied largely on me*

- *I felt obligated to perform well on this task*
- *Our performance on this task was largely due to the things I said or did*
- *I am responsible for most of the things that we did well on this task*
- *I hold my robot teammate responsible for any errors, we made on this task*
- *My robot teammate is to blame for most of the problems we encountered in accomplishing this task*

Inverting the data from these 11 questions makes it possible to group the data from the interaction questionnaire into the eight sub-scales presented in Table 1.6. As for the expectation and satisfaction questionnaires, the MIIC is also calculated for the interaction questionnaire. The analysis showed a MIIC of .325, indicating that the subscales give consistent and appropriate results.

The data from these 8 sub-scales are illustrated in Figure 5.3.

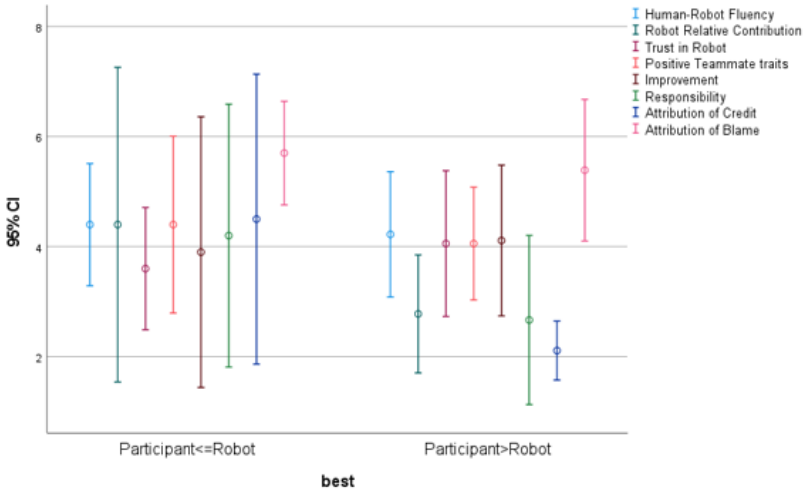

Figure 5.3: Error-bar plot based on each participants median for each sub-scale, divided into the groups from Table 5.3. The points are the means across all participants and the error-bars are 95 % confidence.

These eight sub-scales are used to find significant differences between the two groups. For this, each participant's median of the data within a sub-scale was found. These medians are used to investigate whether significant differences exist between the two groups shown in Table 5.3 for the eight sub-scales. The ranks of the Mann-Whitney U can be seen in Table 5.6. In Table 5.6, the Sum of Ranks are the sum of the data for the two groups for each of the sub-scales. The Mean Rank is the Sum of Ranks divided by the number of participants in the group. The comparison of the Ranks is shown in Table 5.7.

	Participant ≤ Robot			Participant >Robot		
Sub-scale	N	Mean Rank	Sum of Ranks	N	Mean Rank	Sum of Ranks
Human-Robot Fluency	5	7.7	38.5	9	7.39	66.5
Robot Relative Contribution	5	9.8	49	9	6.22	56
Trust in Robot	5	7.1	35.5	9	7.72	69.5
Positive Team-mate Traits	5	8.4	42	9	7	63
Improvement	5	7.2	36	9	7.67	69
Responsibility	5	9.5	47.5	9	6.39	57.5
Attribution of Credit	5	10.6	53	9	5.78	52
Attribution of Blame	5	7.5	37.5	9	75	67.5

Table 5.6: The ranks of the Mann-Whitney U test, for the two groups, for each of the sub-scales.

Table 5.7 shows that for one of the sub-scales (Attribution of credit), the difference between the two groups is significant. For this sub-scale the significant difference indicates that the group where the participants hit more cups than the robot ($Md = 2$) gave the robot significantly less credit than the group where the participants *did not* hit more cups than the robot ($Md = 4$), $U = 7, z = -2.085, p = .037$. This significant difference has a large effect size ($r = -.56$). Furthermore, it can be seen from Table 5.7 that the significance level for the sub-scale Attribution of blame is 1 - which means that mean ranks for the two groups for this sub-scale are equal. Therefore, the sub-scale Attribution of blame is analysed again, depending on whether the participants won the beer-pong game. The Mann-Whitney U between those who won ($Md = 7, N = 5$) and who did not ($Md = 5.5, N = 9$) revealed no significant difference between the two groups, $U = 9.5, z = -1.763, p = .078, r = -.47$.

Sub-scale	Mann-Whitney U	Z	Asymp. Sig. (2.tailed)	Effect Size
Human-Robot Fluency	21.5	-.138	.890	-.04
Robot Relative Contribution	11	-1.563	.118	-.4
Trust in Robot	20.5	-.269	.788	-.07
Positive Team-mate Traits	18	-.627	.531	-.17
Improvement	21	-.202	.844	-.05
Responsibility	12.5	-1.362	.173	-.36
Attribution of Credit	7	-2.085	**.037**	-.56
Attribution of Blame	22.5	.000	1.0	.000

Table 5.7: The results of the Mann-Whitney U significance test, for each of the sub-scales.

Chapter 6

Analysing the quantitative data

This Chapter will go through the analyses of the quantitative data collected in the user evaluations.

6.1 Analysing the difference between aiming for the cups straight in front of the robot and cups where the robot had to turn

In the preliminary user evaluation, it seemed that the participants more often aimed for the red, blue, and turquoise cups, resulting in the robot not having to turn, making it quicker and limiting the number of times it needed correcting at the starting point. Therefore, an analysis to investigate whether there is a significant difference in the number of aims toward the red, blue, and turquoise (straight) cups and the green, purple and yellow (turn) cups is made. As all participants both aimed for the *straight* and the *turn* cups, the data analysis should be made with a paired sample t-test. The assumptions for the paired sample t-test is explained in Section 5.1.1. None of the assumptions are violated by the data. Figure 6.1 for the two variables.

Based on the paired sample t-test there was no significant difference between the amount of times the participants chose to shoot for the *straight* (M=9.86, SD=4.72) and *turn* (M=8, SD=3.51) cups, $t(13) = 1.02$, $p = .326$. The eta squared statistic (r=-.08) indicated a moderate effect size.

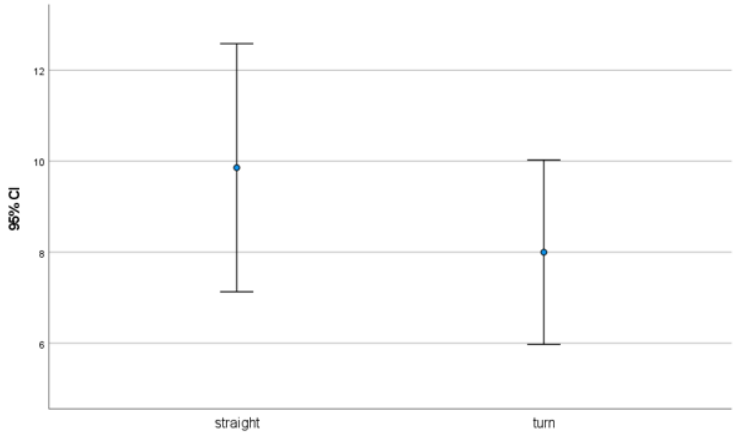

Figure 6.1: The means for how many times the participants aimed for the straight cups, and turn cups, with 95 % confidence intervals.

6.1.1 Does the robot hit what it aims for?

It is interesting to investigate the robot's performance as the design has several degrees of freedom. Table 6.1 shows an overview of what the robot was aiming for when it hit cups. Participant five is not represented in the table, as the robot did not hit any of the cups in this user evaluation. The letters in the Table indicate the color of the cups, e.g., r = red cup, b = blue, and so forth.

Participant	First		Second		Third	
	Robot Aim	Robot Hit	Robot Aim	Robot Hit	Robot Aim	Robot Hit
1	t	t	r	r		
2	b	g	y	y	t	p
3	b	r				
4	p	g				
6	g	p	t	b	g	y
7	r	r				
8	g	y	r	r		
9	r	g	r	r		
10	y	g	r	r		
11	t	b	p	g		
12	b	g	b	r	t	b
13	t	g	t	b	r	r
14	y	r	t	b		

Table 6.1: Table showing what the robot aimed for the times it hit a cup during the user evaluations. Participant 5 is not represented as the robot did not hit any cups.

As Table 6.1 showed that the robot did not hit its target very often, it was found interesting to investigate the frequencies of when the robot hit. These frequencies can be found in Table 6.2. As can be seen from the Table, the robot performed between a 0 % (participant 5) and an 18.75 % (participant 2) hit rate through the user evaluations. Furthermore, the table shows that the robot did not hit the cups aimed at for half of the participants. Throughout the user evaluations, the robot

shot a total of 237 times, of which it hit a cup 27 times, which gives an 11.39 % hit rate. From Table 6.2 the total shots, the total hit, and the percentages of hit cups are also represented for the participants. The Table shows that the participants had a hit rate of between 0 % (Participant 1) and 36.36 % (Participant 14). Furthermore, the Table shows that the total hit rate of the participants is 16.03 %. Furthermore, of the 27 times, the robot hit a cup, only 29.63 % of the time was the cup the participant had aimed for.

Participant	Total Shots Robot	Total Hit Robot	Robot hit Percent	Robot Hit Aimed Percent	Total shots Participant	Total hit Participant	Participant Hit Percent
1	16	2	12.50	50	16	0	0.00
2	16	3	18.75	33	16	1	6.25
3	22	1	4.55	0	21	2	9.52
4	16	1	6.25	0	16	2	12.50
5	16	0	0.00	0	16	4	25.00
6	20	3	15.00	0	20	3	15.00
7	16	1	6.25	100	18	4	22.22
8	14	2	14.29	50	14	4	28.57
9	14	2	14.29	50	14	4	28.57
10	17	2	11.76	50	17	4	23.53
11	17	2	11.76	0	17	3	17.65
12	19	3	15.79	0	19	2	10.53
13	22	3	13.64	33	22	1	4.55
14	12	2	16.67	0	11	4	36.36
Total	237	27	11.39	29.63	237	38	16.03

Table 6.2: Table showing the amount times the robot and the participant shot and hit a cup, and the percentages for hit rate based on these numbers. Furthermore, the table also shows the percentages based on how many times the robot hit what was aimed for.

6.2 Analysing the differences in successful and unsuccessful interactions

A part of the quantitative data was a count of successful and unsuccessful interactions; see Section 1.4.3 for a definition of a successful interaction. A Paired Sample t-test is used to investigate whether there is a significant difference between successful and unsuccessful interactions. The test revealed significantly more successful (M=16.21, SD=2.445) interactions than unsuccessful (M=2, SD=1.664) interactions, $t(13) = 16.3, p = .000$. The mean difference between the successful and unsuccessful interactions was 14.214, with a 95 % confidence interval ranging from 12.33 to 16.1. The eta squared statistic (1.05) indicates a large effect size. As there is an overall significant difference between the number of successful and unsuccessful interactions, it is interesting to investigate whether there is a difference between the two groups described in Table 5.3 as well. Figure 6.2 shows the data concerning successful and unsuccessful interactions for these two groups.

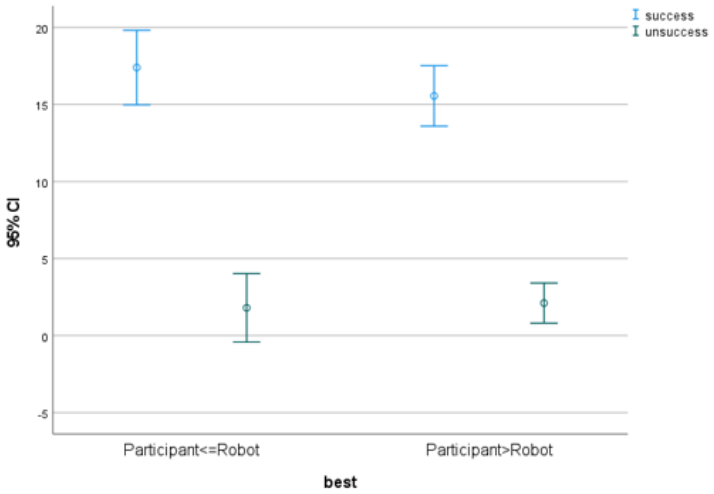

Figure 6.2: Means of the successful and unsuccessful interactions for the two groups, with 95 % confidence intervals.

This is done by analysing one dependent variable at a time and comparing the two groups with an independent sample t-test. The t-test revealed that the two groups did not differ, either in terms of successful or unsuccessful interactions. Table 6.3 shows the results of the two independent Samples t-tests.

Variable	Group	Mean	Std. Deviation	t	df	Sig. (2.tailed)	Cohen's d
Successful	Participant ≤ Robot	17.4	1.949	1.395	12	.188	.188
	Participant >Robot	15.56	2.555				
Unsuccessful	Participant ≤ Robot	1.8	1.789	-.323	12	.752	1.725
	Participant >Robot	2.11	1.691				

Table 6.3: Results of the independent Sample t-test for successful and unsuccessful interactions between the two groups.

6.3 Analysing the time between shots

The user evaluations found that half of the participants chose to start themselves, and the other half chose that their robot teammate should start. Therefore, it was interesting to investigate whether these choices affected the time differences between shots. The time differences between the robot shot, the time differences between the participant shot, and the time differences between the robot and the participant shot were all calculated. Figure 6.3 shows the means of these measures.

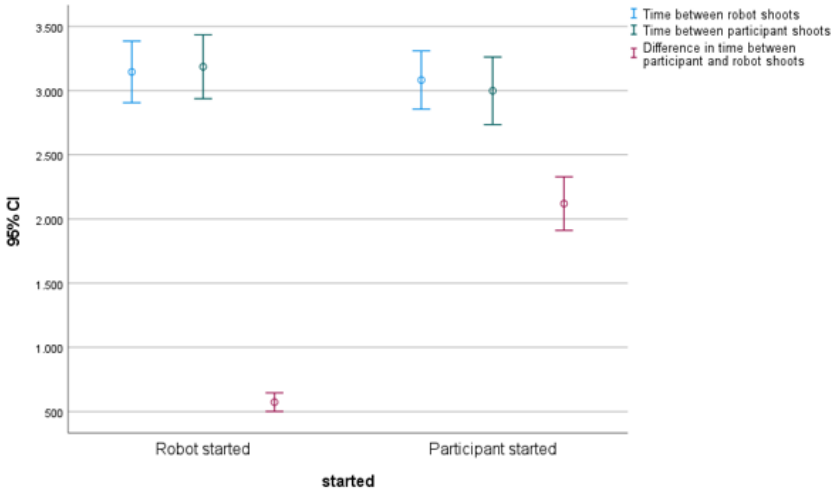

Figure 6.3: Means of the time differences, in the user evaluations, with 95 % confidence intervals.

These measures were used to find significant differences based on who started (the participant or the robot). A Mann-Whitney U test was used to investigate this, as the two groups of participants were independent of each other. The significance test showed only a significant difference between the two groups regarding the time difference between the robot shot and the participant shot. For the group where the robot shot first the median was $Md = 51$ and for the group where the participant shot first was $Md = 48$, $U = 737, z = -12.06, p = .000, r = -3.22$. The effect size indicates that the group where the robot started had the smallest difference between when the robot shot and the participant shot. Furthermore, the

difference between the two groups is larger than three standard deviations.

The time differences between the first and last shots of the robot, the participant, and the differences between them, were also investigated as it could indicate whether the participants optimised the shots.

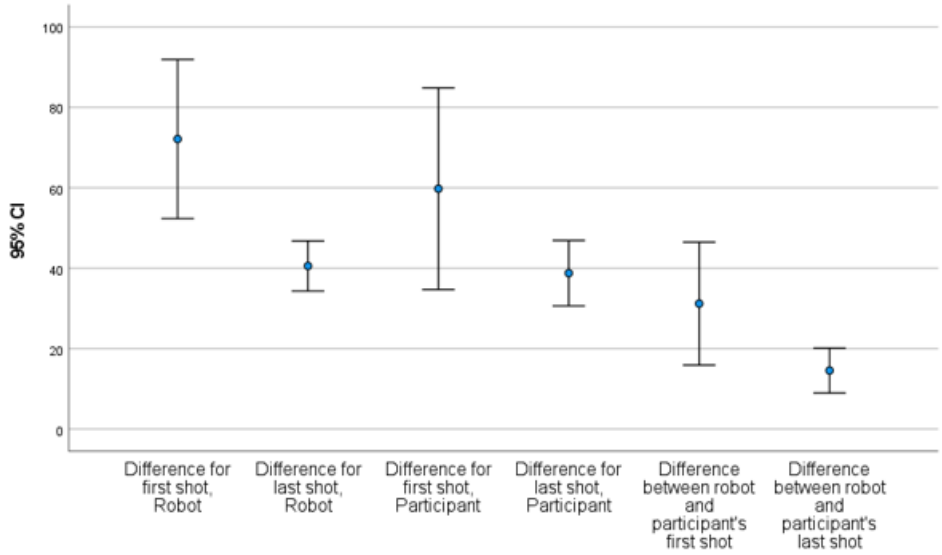

Figure 6.4: Mean of the differences between the first shot of the robot and participant, the last shot of the robot and the participant, and the difference between the first shots of the robot and the participant, and the last shots of the robot and participant. Error-bars indicate an 95 % confidence interval.

Wilcoxon Signed-Rank test was used to investigate this. The Wilcoxon Signed-rank test revealed a significant reduction of the time difference between the first and the last shot, for both the robot $z = -2.64, p = .008, r = -.5$ and the difference between the robot shot and the participant shot $z = -2.48, p = .013, r = -0.47$. For the difference between the first and the last shots for the robot, the medians were; first $Md = 67$ and last $Md = 37.5$. For the difference between the first and last shot between the robot shot and the participant shot, the medians were; first $Md = 27$ and last $Md = 11$. For both of the significant differences, the effect sizes are large.

Chapter 7

Comparing the data-collection methods

As mentioned in Chapter II.1, the scope of the project is to try to answer the following:

> How do data-collection methods differ regarding resources during preparation, the experiments, and the data analysis? What can different data-collection methods deduce from the same interaction between a human and a robot? How do the different data-collection methods supplement each other?

Therefore, this chapter will compare the different data-collection methods to answer the three questions. First, the time resources are presented and compared qualitatively. Hereafter, the outcome of the different data collections will be discussed in terms of what they have in common and how they differ.

7.1 Time resources spend on the different data-collection methods

As mentioned in II.1, one of the aspects of investigating the data-collection methods is to compare the resources used on each of them. For this purpose, it has been estimated how long it has taken to prepare the data-collection method for the user evaluation, how long it took to collect the data for the method, and how long it took to analyse the data.

7.1.1 Observation and Interview

	∼Hours
Preparation	4
Collection	5
Analysis	20

Table 7.1: The hour estimate of the different elements of collecting the subjective data

For the qualitative data-collection method, the preparation consisted of investigating which aspects of the interaction were interesting for the scenario. First of all, an outline was made and investigated in the preliminary user evaluation and then reviewed. The data collection for the observations and interviews were done during the user evaluation. The data analysis consisted of the thematic analysis documented in Chapter II.4. Table 7.1 shows the estimated resources in terms of time spent on the subjective data-collection method.

Questionnaires

	∼Hours
Preparation	5
Collection	2
Analysis	10

Table 7.2: The hour estimate of the different elements of collecting the expectation and satisfaction data

In terms of the expectation and satisfaction questionnaires, the preparation consisted of investigating how similar questionnaires have been designed previously, designing the questionnaires to fit this user evaluation, and investigating the questions in the preliminary user evaluation. The data collection was via Google Forms, where the participants answered the questions. The answers in the forms were converted into excel sheets. The data analysis consisted of the Wilcoxon Signed Rank and Mann Whitney U explained in Section 5.1. Furthermore, the data-analysis estimation also consists of figuring out which methods to use to analyse the data and converting the data to fit these methods. All estimates for the expectation and satisfaction questionnaires can be seen in Table 7.2.

	∼Hours
Preparation	10
Collection	4
Analysis	8

Table 7.3: The hour estimate of the different elements of collecting the interaction data

For the interaction questionnaire, the preparation time consisted of investigating what has been done previously to investigate people's perception of a collaborative task with a robot, deciding which to use, and rewriting them to have similar wording. The data collection consists of the same elements as it did for the expectation and satisfaction questionnaires. The data analysis consisted of the Mann-Whitney U explained in Section 5.2, finding appropriate data-analysis methods, and converting the data. The estimates for the interaction questionnaire can be seen in Table 7.3

Quantitative data

	∼Hours
Preparation	3
Collection	5
Analysis	10

Table 7.4: The hour estimate of the different elements of collecting the quantitative data

For the quantitative data-collection method, the preparation consisted of investigating which aspects of the interaction were interesting for the scenario. First of all, an outline was made and investigated in the preliminary user evaluation and then reviewed. The data collection for the quantitative was done through videos of the user evaluation. The data analysis consisted of parametric and non-parametric methods, all described in Chapter II.6. Furthermore, the data sheets were also converted to fit the purpose of the analyses best. Table 7.4 shows the estimated resources in terms of time spent on the quantitative data-collection method.

7.1.2 Comparing the time-resources between the data-collection methods.

The first thing to be compared is the time spent on preparation. As can be seen from the Tables 7.1 through 7.4, the data-collection method that took the longest to prepare was the questionnaires. These methods took longer to prepare than the others, as they rely on previous research on how to investigate interactions between Humans and Robots in a collaborative task. Furthermore, it can be seen from Section 1.4.2 that the questionnaires from previous research used in this project needed to be reviewed to fit into the interaction investigated in the user evaluations of this project. In terms of preparation, all three data-collection methods have in common that the time estimates do not include the time spent on investigating different scenarios and deciding which to choose for this project. For both the qualitative and quantitative data-collection methods, the preparation estimate is smaller than for the questionnaires. The reason for this is that the chosen scenario also provides an idea of which qualitative and quantitative data could be interesting or important to collect from the user evaluation. The next part of the time estimate of the data-collection methods is the time spent collecting the data from the methods. As mentioned in Chapter II.3 the qualitative data were collected during the user evaluations, the data from the questionnaires were collected via an online platform, and the quantitative data were collected from the recordings of the user evaluations. These choices of how and when to collect the different types of data also resulted in optimising the data collection; therefore, these estimates are similar for the different data-collection methods. The last estimate of the data-collection methods is the time spent analysing the data. Tables 7.1 through 7.4 show that the data-analysis for the qualitative took the longest. As mentioned in Chapter II.4, the qualitative data were analysed with thematic analysis, using the inductive and latent approach. The inductive approach dictated that the themes should

evolve from the data instead of predetermined ideas of themes from, e.g., previous research. Even though it is not investigated, it could be argued that the analysis time would be shortened if the themes and codes were made from a deductive approach. Arguably it minimises the time spent on figuring out what should be a code or a theme and what should not, as it is determined beforehand. The time spent on analysing the questionnaires sums up to 18 hours, see Tables 7.2 and 7.3 which is slightly less than the time spent on the qualitative data. The least amount of time was spent on analysing the quantitative data, see Table 7.4. The time spent analysing the quantitative data consists of the same element as it did for the questionnaires.

7.2 What can the different data-collection methods deduce from the same interaction

This section will be divided into two parts. First, it will be discussed what the methods have in common in terms of the data collected. Lastly, it will be discussed what the methods do not have in common in terms of the collected data.

7.2.1 What the data-collections have in common

Throughout the data analyses, it became apparent that the different events could be analysed with more than one of the methods.

Improvement

Table 1.6 shows that one of the sub-scales of the interaction questionnaire measures the participants' perception of improvement during the user evaluation. Another measure of improvement is the time between the first and last shots. This is argued as being a measure of improvement, as it shows how much the participants (and the robot) have improved during the interaction. With this in mind, it can also be argued that the theme *Time*, from Table 4.1, could also be a measure of this. The theme shows that some participants improved their interaction with the robot during the user evaluation, using different techniques to optimise the number of shots to get a more substantial possibility of hitting a cup.

Tactics

Another thing the data-collection methods have in common is arguably the measure of aiming. Table 4.1 shows a theme called *Aiming* which shows different tactics of the participants during the user evaluation. Another measure of tactics is arguably the quantitative measure of which colors the participants show the robot

during the user evaluations. The quantitative measure can also be used to see the structure of when participants aim for, e.g., specific cups, random cups, or when they try all the colors.

Disappointment

Throughout collecting the data via observation, it became apparent that some participants expressed their disappointment, both toward themselves and the robot, through body language and verbally. Another way which could indicate disappointment, both toward the robot and self, is the sub-scale *Attribution of credit* from the interaction questionnaire. This sub-scale consists of questions that attribute credit to the robot and questions that attribute credit to the participant. Arguably, the more disappointed the participant is toward the robot, the less credit they will attribute to the robot.

Fluent interaction

Another measure that different data-collection methods have in common is whether the interaction between participant and robot was fluent. In the interaction questionnaire, one of the sub-scales is called *Human-Robot Fluency* it measures the participants' perception of fluency in the interaction. Within the quantitative measures, one was a count of successful and unsuccessful interactions between participants and robots. This measure could indicate fluency, with the hypothesis that the more successful interactions, the more fluent the interaction is.

Measures of satisfaction

Other measures which could be measured with other data are the ratings of the satisfaction questionnaire. These questions can be seen in Table 1.8. Arguably the participants' perception of the measured elements could also be derived from the debriefing interview and the interaction questionnaire. The interview asked what the participants thought about their interactions with the robot, which could account for both the first and last questions of the satisfaction questionnaire. The interaction questionnaire could account for the questions concerning succeeding, own skills, and robot skills.

7.2.2 Differences in the data from the three data-collection methods

Throughout the analyses, it was also apparent that the methods differed in the data collected from the same event.

Succeeding at the task

One of the questions of the satisfaction questionnaire was *I think my robot teammate and I did succeed at this task*. From the analysis of this question, it became apparent that the related factor for the perception of succeeding was whether the teams managed to win within the 15 minutes of the user evaluation or not. However, the quantitative measure of success was the number of participants' successful interactions with the robot. Therefore, it can be argued that a measure of success in a human-robot interaction scenario with a game differs from psychophysical data to objective data.

Expectations

Based on the collected data in this project, the data derived from the expectation questionnaires differ from the other data collected, as these are collected before any of the other collections begin. Furthermore, the analysis of the participants' expectations was directly paired with the participants' satisfaction.

7.3 How does the different data-collection methods supplement each other

How data-collection methods can supplement each other is arguably related to whether the different methods can measure the same aspect of an interaction and what that aspect is. Furthermore, through the analyses of the data collection methods and whether these can measure the same thing, it also hypothesised that whether they can supplement each other depends on how many personal differences affect the aspect. Improvement is an aspect that arguably does not depend very much on personal differences. During the user evaluation, the observations and debriefing interviews suggested that some participants thought a lot about time management throughout the interaction and how to improve their interaction. However, the time differences between shots across all participants indicate that all participants improved during the user evaluation and time-optimised. Therefore, in this case, the subjective data alone cannot determine whether participants improved their interaction in terms of time management alone. On the other hand, the quantitative data does not reflect how the participants were able to optimise their time-management throughout the user evaluations, which is possible to derive from the subjective data.

Another aspect, which depends less on personal differences, is the measure of fluency in the interaction. Fluency was measured somewhat directly through the interaction questionnaire, and the quantitative measure of successful interaction could also relate to this aspect. These two measures supplement each other in

investigating whether the perception of fluency matches the quantitative measure of fluency. However, neither of the measures supplement each other regarding why the interaction is deemed fluent. However, the debriefing interview could, to some extent, reveal why the participants found the interaction fluent. As can be seen from Table 4.1, one of the sub-codes of the *Interaction*-theme is *Understandable*. Quote (1) which could indicate why the interaction was fluent.

During the user evaluations, an aspect measured in more than one way was the tactics. This measure arguably depends more on personal differences than that of, e.g., improvement. The observations and debriefing interviews made it possible to collect data concerning how and why the participants used their tactics. However, the quantitative data only made it possible to determine what was aimed for, in what order, and how many times the robot's starting point was corrected. Therefore, it is argued that in the case of tactics, these two data-collection methods supplement each other well. Furthermore, the quantitative data can indicate whether the tactics worked if this is of interest.

Another measure that relates to personal differences is the measure of disappointment. As mentioned, this measure was both done through observations and the interaction questionnaire. However, it can also be argued that these measures also relate to different kinds of disappointment: Disappointment toward specific events and disappointment toward the experience as a whole. Therefore, these two measures supplement each other quite well, as all participants, to some extent, showed disappointment toward themselves and/or the robot during the user evaluation. However, the difference in the attribution of credit, which arguably measures the overall disappointment, revealed a difference dependent on the groups based on whether the participants hit more cups than the robot. Furthermore, the measure of disappointment could also be supplemented by the difference between the expectation and satisfaction questionnaires, dependent on the same groups. However, this is not investigated as it should be done with a MANOVA.

Previously, it was mentioned that the participants' satisfaction measures could also have been derived from the debriefing interview and the interaction questionnaire. However, this could make the analysis of the expectations more difficult, as they cannot be directly paired with the data from the interviews or the questions from the interaction questionnaire. Therefore, the interaction questionnaire and subjective measures should not replace the satisfaction questionnaire but could be a supplement to get more insight into why the satisfaction questionnaire reveals what it does.

Another aspect of the data collection that arguably improves with supplementary data is that of unsuccessful interactions. The measurement of unsuccessful interactions is a quantitative measure. The measurement only reveals how many times an interaction was unsuccessful for each of the participants. Furthermore, this project has only analysed it to find whether the participants had more suc-

cessful than unsuccessful. However, it does not reveal why the participants had unsuccessful interactions with the robot - which could be explained by the code *Path recognition* from the subjective data, see Chapter II.4. The code has three related sub-codes, which to some extent does reveal why some interactions were unsuccessful in the user evaluations presented in this project.

Chapter 8

Discussion

This chapter will discuss different aspects of the user evaluation.

8.1 The scenario of the user evaluation

In the literature review, See Part I, one of the aspects of the discussion was whether the scenario chosen for the user evaluations fitted the possible context the system/robot could be used in. Therefore, this will be discussed for this user evaluation as well. When discussing the scenario of the user evaluation of this project, see Section 1.1, it is essential also to consider the purpose of the user evaluation and how it differs from the papers of the literature review. For most of the papers in the literature review, the purpose was to either validate a robot or a "proof of concept" investigation. In this user evaluation, the purpose was not to design a robot; on the contrary, it was to investigate how different data-collection methods can be used to investigate an interaction. It should, however, be considered that the setting of this user evaluation may be quite limiting. A collaborative robot should help people with tasks to make them more efficient and pleasurable for the person. As can be seen from Table 5.3, the robot arguably only made the task more efficient in half of the user evaluations. However, the analysis of the observations and interview did suggest that the robot made the task pleasurable, though it was not compared to the pleasure of an ordinary game. As the setting is limiting, it is unsure whether the findings from this user evaluation could be used in other user evaluations with collaborative robots.

8.2 The chosen robot

As mentioned in Section 1.2, the Fable robot from Shape Robotics was used for the user evaluation. The robot has several advantages, one of them being the platform

from which it is programmed. However, the robot also has some disadvantages. One of them is its lack of precision in the Fable Spin module, which was mentioned earlier. Another disadvantage of the Fable Spin module is its ability to recognise colors. First of all, it is only able to recognise eight colors. Although, when the sensors were placed under the robot, which they were for the design used in the user evaluation, the robot sensed a white color a lot due to reflection from lamps. Furthermore, during the programming of the robot, it was found that the robot was not able to recognise black, which was also an option; it seemed that this problem was also due to reflection from the lamps. Another problematic colour was turquoise. It was tough to find a color match on paper to get the robot to detect the right color. As the observational data have shown, this was also a struggle during the user evaluations.

8.3 The design of the robot

The design of the robot can be seen in Figure 1.5. The design of the robot is non-humanoid, as the robot does not have any anthropomorphic characteristics. Some participants from the user evaluation indicated through the observational data and the debriefing interview that they would have preferred a robot with more anthropomorphic characteristics. Based on the data, these anthropomorphic characteristics mainly concerned cheering, for example, when the robot hit a cup or when the participants did. Moreover, the robot should also be able to recognise itself when cups are hit. Another participant also had some concerns based on the lack of feedback from the robot, see Quote (13). This quote indicates that the lack of feedback from the robot regarding when it is going to shoot is troublesome. Fortunately, in the case of this scenario, the thrown object is not harmful.

Another design choice of the robot that should be discussed is the catapult setup. The setup was chosen so that the robot could shoot longer. However, it did also result in inconsistencies in the shots. The glass-fiber rod used in the setup was enforced using a welding thread. As can be seen in Figure 1.5 the rod is bent each time the robot has to shoot, which could be an explanation as to why the robot shot inconsistently throughout the user evaluations. The force, fed to the rod results in the rod becoming warmer. This change might not be detectable, but it can expand the rod, resulting in the robot shooting longer, shorter, or skewed. This uncontrollable factor of the design of the robot is a limitation of the design used in the user evaluation.

8.4 The data-collection methods

The user evaluation was chosen to collect data using three different categories of data collection: Subjective, psycho-physically, and quantitative.

8.4.1 The subjective data-collection methods

In Section 1.4.1, it was described that the subjective measures collected during the user evaluation would be observations. Furthermore, it was discussed that the observations alone might not consist of enough data to analyse the interactions. Therefore, a debriefing interview was designed as well. Whether this was necessary depends on different things. One of these is the recruited participants. People show their emotions differently, and the subjective data analysis showed that some participants were more verbally in their expression of others. A debriefing interview might not be necessary for these participants, as they already expressed their feelings about the given task during the user evaluation. However, some of the participants were not very verbal during the user evaluation, and for these participants, it was harder to analyse their thoughts about the task merely through observations. When the purpose is to investigate interactions between people and robots, it could be an advantage to also ask participants directly about their thoughts after completing the user evaluation.

8.4.2 Psycho-physical measures

When using psychophysical measures, a consideration is what changes the measure is meant to detect. In Section 1.4.2, it was decided to use two questionnaires to collect psychophysical data from the participants. In concerns of the questionnaire, it can be argued that the expectation-satisfaction could, to some extent, reflect psychophysical measures, as a change happens between the two questionnaires. However, for the interaction questionnaire, it might be a questionable definition of the questionnaire. The user evaluation was done using a within-subject design, meaning that all participants went through the same user evaluation, collaborating with the same robot. Looking objectively at the user evaluation, there was no change in the interaction with the robot between participants. On the other hand, the number of cups hit by the robot, and the participants, during the user evaluation might suggest that subjectively the user evaluations did provide a change between participants. This is to some extent also reflected by the analysis in Section 5.2, which showed that when the participant hit more cups than the robot, they rated the sub-scale *Attribution of Credit* significantly lower. The change between participants was not intended but happened due to the degrees of freedom in the design of the robot. Therefore, in this user evaluation, the interaction questionnaire cannot be defined as a psychophysical measure, but it happened to reflect one.

8.4.3 Quantitative measures

In Section 1.4.3 the outline for the collection of the quantitative measures can be seen. From the outline, it can be discussed what the intended measure's purpose is. The section shows that most of the measurements concern the performance instead of how the interaction. Arguably, the interaction is only measured using the count of successful and unsuccessful interactions.

8.4.4 Preliminary user evaluation

In the literature review in Chapter I.2, it was shown that some of the papers used preliminary user evaluations to investigate aspects of their robot design before their main user evaluations. Furthermore, it was discussed that preliminary studies could have been an advantage for two of the papers. One of these was the paper by St-Onge *et al.* [2019] which could have used this to investigate the effect of different constraints they had in their user evaluation. To minimise possible errors in the user evaluation of this study it was chosen to do a preliminary user evaluation. The purpose was to investigate the use of the data-collection methods. The preliminary user evaluation findings first made it clear that when to collect the observational and quantitative data should be changed so that the observational data was collected during the user evaluation instead of afterward through the recordings. Making the observations reflect the situation at hand and the environment of the user evaluation. Furthermore, the preliminary user evaluation findings indicated that the measures from the observations and quantitative methods should be changed. Instead of focusing on reactions and counts concerning hitting cups and not hitting cups, they were changed to reflect the interaction better. Therefore, the preliminary user evaluation was an advantage for this user evaluation, as it ensured that the measures would measure what was actually intended.

The language of the user evaluations

The participants in the user evaluation were all Danish. Therefore, most of the user evaluation was conducted in Danish Involving; the introduction, the consent form, and the interview. However, as can be seen from Section 1.4.2, the questionnaires are in English. For the user evaluation, it was decided that this should not be changed, meaning that the participants answered the questionnaire in English. This choice can have both advantages and disadvantages. An advantage is argued to be that the meaning of the questions is not distorted, which could happen if they were translated. On the other hand, a disadvantage could be that the participants did not fully understand the meaning of the questions, as they were not in their preferred language. However, during the user evaluations, this did not seem to be a problem, as the participants did not ask questions about the meaning of the

questions. Even though it did not seem to be a problem, it should be considered to conduct the entire user evaluation in English if it is reproduced, as it could ensure consistency and possibly more participants as other than Danish people can be recruited for the user evaluations.

8.5 Comparing the data-collection methods

As can be seen in Section 7.1, one way the three different data-collection methods were compared was in terms of time resources. The assessment of the time resources used first of all showed that the time spent on the preparation of the methods was longer for the questionnaires used than for the subjective and quantitative methods. As mentioned, the questionnaire needed more preparation, as they were made from questionnaires from previous research on Human-Robot Collaboration. Even though this resulted in a longer preparation time, it should be considered that if one were to develop an entirely new questionnaire, this would possibly take longer, as a more thorough investigation of the questions should be done before the user evaluation. In terms of the time used to collect the data, it can be seen that the time is similar for all three methods. For the subjective and quantitative methods, it was changed when the data should be collected from the user evaluations; see Chapter II.3. This change reduced the collection time for the subjective data.

Section 7.2 and 7.3 showed what the different methods had in common, how they differed and how they could supplement each other in the analyses of the data. In relation to how the methods can supplement each other, it was argued that this depended on personal differences in the measurement. An example of a measure, depending on personal differences, is tactics. It is argued that this depends on personal differences, as not all participants had a specific tactic, and some thought a lot about how to optimise their interaction and performance.

When choosing data-collection methods, it is important to consider the purpose of the user evaluation. The purpose has much impact on which method to use. Arguably the more subjective data concerns the *"why"* of an interaction. The *why* data can be beneficial in prototyping, as it indicates why, e.g., interactions were not successful. On the other hand, quantitative concerns the *"what"* of an interaction. *What* went well, *what* went wrong, *what* did the participants do. This type of data is of interest when validating a design. For example, it gives evidence of good interactions when more went well than went wrong. The psychophysical measure is arguably an in-between method that measures the *how much*. This is argued as it, in this case, is a measure of "how much did x subjective's perception of the interaction affect the rating of the interaction". It is a used way of quantifying a subjective measure and is arguably a good supplement for both subjective and quantitative measures in robot development. For the subjective data, it can supplement by indicating how much an element of an interaction affected participants.

On the contrary, it can be used to indicate why a quantitative measure shows what it does. However, this is only possible if both the psychophysical and the subjective or quantitative measure, measure the same aspect of a user evaluation.

8.6 The analyses used for the data

There are many ways of analysing subjective data. From the literature review, see Chapter I.2 it seemed that this type of data was mainly used to extract a limited amount of statements, to supplement their other data. For the observations and interviews of the user evaluation in this project, an inductive thematic analysis using a latent approach was used. As different types of data collection methods were used in the user evaluation, and as one of these was the interaction questionnaire, it could be argued that the deductive method should have been used instead. The deductive approach could have indicated how well the observations and interview fitted the sub-scales from the interaction questionnaire, which would have been very useful if, for example, the purpose was to investigate one of the initial prototypes for an actual beer-pong playing robot. However, as the purpose was to, among others, compare the outcome of the data and analyses, it was deemed necessary not to let the different data analyses influence each other. Another argument could have been that a content analysis should have been used. *Content analysis* is a method used to categorise subjective data and quantify the observations after that. The content analysis was also decided against, as it was deemed more important to analyse how participants reacted to, e.g., unexpected events than how often they reacted to them.

In Section 7.2, it was mentioned that the measure of success in the satisfaction questionnaire and through quantitative data did not reflect the same thing. The analysis of the satisfaction questionnaire showed a significant difference in the responses to the question *"I think my robot teammate and I did succeed at this task"* between the groups of who managed to win within the 15-minute time frame and who did not. This difference can be due to the phrasing of the introduction before the expectation questionnaire. In the introduction, the participants were told: *"The goal of the game is that you and the robot wins the game, by hitting all six cups. If this is not done within 15 minutes, the experiment will stop"*. This phrasing could have biased the participants. Objectively speaking, a game that was restricted to 15 minutes would arguably also be successful if the team had hit more cups than the simulated opponent, which all of them did. However, the questionnaire results suggest that the participants only view success based on winning. Therefore, introductions in the future should not identify specific goals if the aim is to analyse interactions rather than performance. The analysis of the interaction questionnaire showed a significance level of 1, see Table 5.7. This significance level was for the analysis of the sub-scale *Attribution of blame*, analysed between the groups depending on

whether the participant hit more cups than the robot, which is interesting as it indicates that the participants do not blame the robot. The analysis investigating this sub-scale concerning whether they won within the 15 minutes did not show a significant difference for this sub-scale either. It could be due to the perceived intelligence of the robot; however, this is not investigated and is only speculation. However, it is an interesting result that could be investigated further.

Chapter 9

Conclusion and Further works

This project investigated the following research question:

What aspects of robot design could benefit from systematic user evaluation methods, to improve human-robot interaction?

To answer this question the first part of the project presents a literature review. Through the literature review it was found that researcher use different data-collection methods for their user evaluations. Furthermore, the literature review also showed that some researchers use preliminary user evaluations, and simulations before their main user evaluations, and that other researcher should have used them. The literature review also investigated the settings of the user evaluations, and discussed how these could be changed, to enhance the results. The findings from the literature review indicated that the fist step toward a systematic user evaluation should be comparing different data-collection methods. Therefore, the second part of the project investigate what subjective, psychophysical, and quantitative measures have in common, how they differ, how they can supplement each other, and the time resources spent on the methods. The investigation was done through a collaborative task of beer pong. From this setting the following data was collected and analysed: observations, interview, ratings on three different questionnaires, and a variety of quantitative measures.

The comparison of the data from these measures indicated that subjective data explains *why*, the psychophysical explains *how much*, and the quantitative explains *what*. Therefore, the three different data-collection methods supplement each other well. However, the setting of a user evaluation is an important consideration when choosing data-collection methods.

Considering, the results of this project it is suggested that the three different data-collection methods are useful in different stages of robot development. Subjective data-collection methods that explains why something happens, or why the participants do what they do, are useful for initial investigation of a robot prototype. Psychophysical data-collection methods, can be used in different stages, to

supplement other methods. Quantitative data-collection methods are useful in the final stages of robot development, as it gives an indication of what goes wrong, unrelated to the perception of the participants. However, it is also argued that the quantitative measures should not stand alone, even in the final stage.

The next step toward a systematic user evaluation method in human-robot interaction can consist of different things. Either the assumptions of when to use different data-collection methods should be investigated thoroughly, in different developmental stages of robots. Another option could be to investigate when it is necessary to compare robots, or robot systems, to others in a user evaluation. The hypothesis for such an investigation could be that comparisons are not necessary until the final stage of robot development.

Chapter 10

Appendix

10.1 Appendix for the literature review

10.2 Consent form used in the preliminary user evaluation

The consent form used in the preliminary user evaluation will be shown in English. However, it is anticipated that the majority of the participants of the preliminary user evaluation will be Danish, therefore, it will be translated to Danish. Whether the English or the Danish version of the consent form is used, depends on which language the participants prefer.

10.2.1 Consent form for participants in this user evaluation (English)

With this consent form I am asking you to participate in a study titled "Human-Robot Interaction in a collaborative task", for my master thesis in Engineering Psychology at Aalborg University. In this consent form you will receive written information about the study. Besides this written information, the information will also be given to you verbally. Please note that your participation in this study, is completely voluntary, and you can at any point of the study, for any reason, choose the stop. If this becomes the case, please let me know.

About the Study

In this study I am investigated interactions between humans and robots, when they have to collaborate on solving a task. The task you will have to collaborate with a robot about is a game of beer-pong, where the goal is for your team to win the game, by being the first ones to hit all six cups. Your teammate in this task will be a robot, design to play beer-pong.

What I will ask you to do

To be able to complete the task, you will have to interact with the robot in a certain way. How you will interact with the robot, will be described to you at a later point. During this study I will collect different types of data. The data will be collected through observation, interview and questionnaires. All of the collected data will be anonymous, and stored responsively as of current GDPR regulations. This means that all data will only be saved locally, until the end of the project in the summer of 2022. All video material will only be used by I, the undersigned. I will use the data in my master thesis, which will be published at Aalborg University's project Library. Be advised that this consent can at all times be withdrawn. If you choose to withdraw your consent, I will delete all the data I have collected from you for this study immediately after your withdrawal.

To be able to collect all the desired data for my master thesis, you interaction with the robot will be video recorded. By signing this consent form, you automatically consent to the video recordings as well.

If you have any questions about this consent form please feel free to ask me.

Contact information

If you at any point, after this study, need to get in contact with me please contact me on the email address below. cpje15@student.aau.dk

10.2.2 Samtykke erklæring for forsøgspersoner i dette studie (Danish)

Med denne samtykke erklæring beder jeg om din deltagelse i et studie ved navn "Human-Robot Interaction in a collaborative task", som er til mit speciale i Produkt- og Designpsykologi på Aalborg Universitet. I denne samtykke erklæring vil du modtage skriftlig information omkring studiet. Udover denne skriftlige information vil du også få information omkring studiet mundtligt. Venligst vær opmærksom på at din deltagelse er helt frivillig, og du må til enhver tid, af hvilken som helst grund, vælge at stoppe. Hvis dette bliver aktuelt, så venligst fortæl mig det.

Omkring studiet

I dette studie undersøger jeg interaktioner mellem mennesker og robotter, når de skal samarbejde omkring en opgave. Opgaven du skal samarbejde med en robot om, er et spil beer-pong, hvor målet er at i skal vinde ved at være de første til at ramme alle seks kopper. Din holdkammerat i opgaven vil være en robot der er designet til at spille beer-pong.

Hvad jeg vil bede dig om at gøre

For at løse opgaven, skal du interagere med robotten på en bestemt måde. Måden du skal interagere med robotten vil jeg fortælle dig om senere. I løbet af studiet vil jeg indsamle forskellige slags data. Dataet bliver indsamlet via observation, interview og spørgeskemaer. Alt det indsamlede data vil være anonymt og i den forbindelse vil det blive lagret forsvarligt i forhold til gældende GDPR regler. Dette betyder at alt data udelukkende gemmes lokalt, indtil projektets afslutning i sommeren 2022. Alt video materiale vil kun blive brugt af undertegnede. Jeg vil bruge dataet i mit speciale som publiceres på Aalborg Universitets projektbibliotek. Du bør være opmærksom på at dit samtykke til enhver tid kan trækkes tilbage. Hvis du vælger at tilbagetrække dit samtykke så vil alt data jeg har omkring dit forsøg blive slettet med det samme. For at jeg kan indsamle det data jeg lige har beskrevet, vil jeg optage din interaktion med robotten. Ved at underskrive denne samtykkeerklæring samtykker du automatisk til at jeg må optage interaktionen.

Hvis du har nogle spørgsmål omkring denne samtykke erklæring så sig endelig til.

Kontakt information

Hvis du på noget tidspunkt har brug for at kontakte mig efter studiet er overstået, så venligst kontakt mig på nedenstående email adresse.

cpje15@student.aau.dk

10.3 Introduction to the preliminary user evaluation

As well as for the consent form, some participants might prefer an introduction to the user evaluation in Danish. Furthermore, the introduction to the preliminary user evaluation will be given to the participants before entering the room where the user evaluation are to take place.

Introduction (English)

"Hello, and thank you for participating in my study. In this study I want to investigate interactions between humans and robots when these are to collaborate on a task. For this purpose what you have to do today is to play a game of beer-pong with a robot as your teammate. However, for the robot to play the game it has to get information from you.

Before explaining any further, I would like for you to read and sign this consent form. Please note that i will video record the interactions you have with the robot, and by signing the consent form, you also consent to the recordings. The recordings will be deleted after the project is done in the summer of 2022.

Participants read and sign the consent form
As mentioned, you are going to play a game of beer-pong with a robot. In relation to that there is a set of rules I want you to follow. You and the robot are to both shoot in each round. Who of you are going to shoot first is up to you, however the order you choose in the beginning will have to be the order throughout the game. In the experiment the opponent of the game will be simulated, meaning that I will remove one of more cups for the opponent once in a while. The goal of the game is for you and the robot to win, by hitting all 6 cups. The experiment will stop after all 6 cups have been hit, or after 15 minutes. The rules for removing cups are the same as they would be in an ordinary game of beer-pong. If you hit one cup, on is removed, if you hit two, two is removes, if you hit the same cup, two will be removed and you'll get the balls back. The next thing I will ask you to do, is to answer a questionnaire. The questionnaire is about your expectations. The questions are in English.
After the participants have answered the expectation questionnaire
Thank you. Now we are ready to start the study. Please follow me into this room.*(Only in the preliminary study)*

This is the robot you will interact with during this study. The interactions you'll have with the robot is twofold. The first thing you'll have to do is to show the robot which of the 6 cups you want the robot to hit. This is done by sliding one of these six cards underneath the robot **shows the motion, with a hand**. The meaning of the cards can be seen on this peace of paper. When you have shown a card to the robot, it will start driving. When it stops driving, this **points at joint module** will move to an angle of 45°, that means that the robot is ready to be loaded. Therefore, the second half of the interaction begins. You'll have to take this **takes the glass fiber rod**, and secure it in this hook, and then put a ball in it. The robot will then shoot by itself. The time from the robot is ready to be loaded until it shoots on its own is 5 seconds. It might seem like a short time-frame, however it should be sufficient. When the robot have shot, it will drive back to its starting-point, which is marked by this black square. However, sometimes the robot does not hit the right place in the square. If this happens you are welcome to move it around, until it is inside the square **demonstrates how it could be done**.

When you shoot, you have to behind this line on the floor.

Do you have any questions?

Okay, so I'm just gonna start the recording, and the robot, and then we'll begin the game."

10.4 Randomisation of cards

Reference	Subject design	Number of Participants	Measurement	Analysis method
Menne & Lugrin [2017]	Within-subjects design	33	Quantitative measures: Recordings of participants faces Qualitative measures: Self-reported answers to PANAS questionnaire	One-tailed t-test
Totsuka *et al.* [2017]	Within-subjects design	15	Qualitative measures: 7-point likert scales measuring 3 items of appropriateness of the robot and 1 item evaluating the robot as a walking partner Observation of the robot interview of the participants Quantitative measures: Number of utterances from robot and their comparison to the surroundings	ANOVA for the the scale measurements. Percentages and frequency results for system behaviour. Extraction of behaviors and statements for the observation and interview.
Rakita *et al.* [2018]	Between-subjects design	32	Quantitative measures: metrics of improvement in task performance. Qualitative measures: 7-point likert scales measuring 2 items on fluency, 3 items on robot intelligence, 2 items on trust in the robot, and 3 items on the goal understanding of the robot	ANOVA
Gielniak & Thomaz [2011]	Within-subjects design All participants watched videos of all three categories, but only wathed 4/20 of each category	41	Quantitative measures: Motion capture of; mimicking their own motion. And a count of the number of views before mimicking and choosing the best and the most natural motion. Whether they recognised a motion, what name they would give the motion Qualitative measures: Which motions were easiest and hardest to mimic, whether they perceived a difference in the motions in part 2, which they found better and most natural, and their resoning of these choices	Acumulated recognition: how many participants recognised the specific motion of the three categories. ANOVA to test whether their system makes more human-like motion based on the results of participants mimicking
Jayaraman *et al.* [2018]	Within-subjects design	30	Quantitative measures: distance to collision, jaywalking time, crossing time, crossing speed, and waiting time. Qualitative measures: survey measuring trust and propensity to trust	Mixed linear model to find relationship between trust and the quantitative measures.
Rouanet *et al.* [2011]	Between-subject design Participants used one of four interfaces for interaction	107 33 used a iPhone interface 27 used a wiimote 33 used a wiimote-laser 15 used geatures (WOZ)	Quantitative measures: The participants pictures during the experiment, with the names the participants gave the pictures, and the time it took to complete the task Qualitative measures: Pre-questionnaire about technological profile and attitude towards robots Post-questionnaire consisted of 6 statements about the interface and 4 about the game, measured on 5-point likert scales	Pictures were divided into three groups, and an one-way ANOVA was used to compare the groups One-way ANOVA was used on the post-questionnaire, between the four interfaces.
Kruse *et al.* [2014]	Within-subject design	17	Qualitative measures: Path behaviours of the participants, when they "interfered" with the robot two semantic-differential scales about the robots: "clear-confusing", "uncomfortable-comfortable"	A qualitative deduction of the path behaviours ANOVA on the semantic-differential scale answers
Fitter *et al.* [2018]	Between-subject design 2 conditions: one where participants could personalise the appearance of the robot, and one where they couldn't	24	Quantitative measures: completion-time of the obstacle-course, and the answers to the questions at the stations on the course Qualitative measures: 11 questions based on other surveys about telepresent robots. 8 measured on7-point bipolar likert scales, and 3 on 5-point unipolar likert scales	*Not mentioned*
Javed *et al.* [2019]	Within-subject design	18 13 traditionally developing (TD), 5 diagnosed with Autism Spectrum Disorder (ASD)	Quantitative measures: Engagement of the children, measured by looking at 6 different target behaviours. Qualitative measures: Baseline-, pre-session, post-session questionnaires, filled by parents.	T-test to test for differences in engagement between TD and ASD children. T-test to test for differences between the two robots. Engagement contribution for each station in the experiment.
St-Onge *et al.* [2019]	Within-subject design	27 all with experience and knowledge within dancing	Quantitative measures: The performance of the classifier system. Qualitative measures: Questionnaire based on the QUEAD survey, measured on 7-point likert scales	Confusion matrices to find differences in classifiers. Friedmans ANOVA were used to analyse the questionnaire data.

Table 10.1: This table shows the subject design, the number of participants, which measurements were used, and which methods were used to analyse the data in the 22 papers.

Reference	Subject design	Number of Participants	Measurement	Analysis method
Oudah *et al.* [2015]	Between-subject design divided into three different groups	48	Quantitative measures: Avarge payoff: how many points did the participants gather in the two games. how well the system has learned to interact with people Solution quality: how many achieved differens solutions for the games. how well the system cooperates with people Qualitative measures: Questionnaires administered to participants after each of the two games	ANOVA to compare the different conditions. Also compared to the results of their initial study (which did not make use a robot)
Hanheide *et al.* [2017]	Within-subject design	13 (only usability study)	Quantitative measures: Whether the participants completed the tasks within the time limits, how many mistakes they made in the tasks, how many hints they needed to complete the tasks. Qualitative measures: semi-structured interview and questions measured on 5-point likert scales	Primarialy percentages, also extractions from the interview, and observations derived from the task-performance
Dragan & Srinivasa [2014]	Study 1: within-subject design Study 2: within-subject design Study 3: mixed design	Study 1: 25 Study 2: 24 Study 3: 16	Study 1 and 2: Quantitative measures: did correctly anticipate trajectory, before and after familiarization. Qualitative measures: self-reported measures of utility, improvement, and confidence Study 3: Quantitative measures: the chosen distance to the robot Qualitative measures: 7-point likert scale answer to their willingness to work side-by-side with the robot ("if it moved the way i saw" after familiarization)	In all 3 studies they analysed using logistic regression and ANOVA.
Kwon *et al.* [2020]	Within-subject design	Study 1: 50 Study 2: 10	Study 1: Quantitative measures: how often does the robot anticipate the participants choices correctly Study 2: Quantitative measures: Efficiency in terms of time it took to build the tower. Safety in terms of trajectory length of the robots motions when it had to interfeer with the participant. Qualitative measures: Four-items measured on 7-point likert scales, and which they prefered and which anticipated their behaviors best.	t-test
Mavrogiannis *et al.* [2018]	Within-subject design	180 (online survey)	Quantitative measures: Participants were to answer how to agents in the video would pass each other and were given points for fast and correct answers, and points were removed for slow and wrong answers	Linear models, t-test, Pearson's correlation coefficient
St. Clair & Mataric [2015]	Within-subject design	15	Quantitative measures: Recordings of locations of people, robot, sheep, and virtual objects. Audio recording of participants, and recordings of the interaction (overview and sideview) Qualitative measures: demographics, and 27 questions answered on 7-point likert scales	They used a post hoc test called linear mixed effects regression, however they do not mentioned what method they used beforehand. Could be a t-test, given that use a within-subject design and only have two conditions
Murakami *et al.* [2014]	Within-subject design	20	Qualitative measures: Ratings of naturalness, perceived safety and an overall evaluation, after each session, as well as an interview after each session.	ANOVA for 7-point likert scale answers extractions from the interviews
Nikolaidis *et al.* [2018]	All three were between-subject design	User evaluation: 151 Follow-ups: 52	User evaluation (three conditions): Quantitative measures: How many participants adapted to the robot, in each of the conditions Qualitative measures: 5-point likers scale rating of trusting the robot. Open-ended questions Follow ups: The same as the user evaluation, as these were compared	User evaluations and follow ups: Chi-squared, equivalence test, and TOST equivalence, Analysis of open-ended questions

Table 10.2: This table shows the subject design, the number of participants, which measurements were used, and which methods were used to analyse the data in the 22 papers.

Reference	Subject design	Number of Participants	Measurement	Analysis method
Jacq *et al.* [2016]	within-subject design	First case study: 1 Second case study: 1 User study: 8	Case 1: Quantitative measures: Commitment, as the number of demonstrations by the child, time used on the demonstrations and the duration of the sessions Qualitative measures: post-sessions interview of parents Case 2: Quantitative measures: The demonstrations of the child, the time spend on the demonstrations and the duration of the sessions. User evaluation: Quantitative measures: The same measures as Case 1. The childrens evaluation of the robots performance (thumbs up or down for each time the robot wrote something) Correlation between the evaluations, and the distances between the robots letters and reference templates, as a measure of the childrens awareness of the robot's progress.	They don't explain any significance tests.
Gielniak & Thomaz [2012]	Between-subject design for both studies	Study 1: 54 Study 2: 68	Study 1: Quantitative measures: Fill-in-the-blank questions from the story Qualitative measures: 16 questions measured on 7-point likert scales. Short-answer questions about the content of the story. Their favorite part of the story, favorite motion, and the reasoning for these two choices. Study 2: Quantitative measures: recordings of trajectory of the participants eye gaze. The measure of the exagerrated motions the participants preference for 5 different motions. Qualitative measures: Same as study 1.	ANOVA percentages of the fill-in-the-blank questions, compared between conditions. also percentage of right answers for the two types of motions they investigated
Doering *et al.* [2019]	Within-subject design	16	Quantitative measures: The number of questions asked by the robots. Qualitative measures: four questions measured on 7-point likert scales. observed behaviors of the robot	It seems that they used t-test for the 7-point likert scale questions No methods were explained for the other measures.
Kwon *et al.* [2018]	Preliminary studies: Within-subject design Main study: Between-subject design when participants had to elicit goal and cause of the robot Within-subject design For perceptions of the different robots	Preliminary studies: 60 in both Main study: 120	Qualitative measures for all three: Questions answered on 5-point likert scales Main study: open-ended questions about goal and cause of inbcapability of the robot	ANOVA

Table 10.3: This table shows the subject design, the number of participants, which measurements were used, and which methods were used to analyse the data in the 22 papers.

Participant	1.	2.	3.	4.	5.	6.		Participant	1.	2.	3.	4.	5.	6.
1								21						
2								22						
3								23						
4								24						
5								25						
6								26						
7								27						
8								28						
9								29						
10								30						
11								31						
12								32						
13								33						
14								34						
15								35						
16								36						
17								37						
18								38						
19								39						
20								40						

Table 10.4

Bibliography

Agresti, Alan. 2018. *Statistical methods for the social sciences*. 5 edn. Boston: Pearson.

Alves-Oliveira, Patrícia, Ribeiro, Tiago, Petisca, Sofia, di Tullio, Eugenio, S. Melo, Francisco, & Paiva, Ana. 2015. An Empathic Robotic Tutor for School Classrooms: Considering Expectation and Satisfaction of Children as End-Users. *Pages 21–30 of: Social Robotics*. Lecture Notes in Computer Science, vol. 9388. Cham: Springer International Publishing.

Bartneck, Christoph, Kulić, Dana, Croft, Elizabeth, & Zoghbi, Susana. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International journal of social robotics*, **1**(1), 71–81.

Bartneck, Christoph, Belpaeme, Tony, Eyssel, Friederike, Kanda, Takayuki, Keijsers, Merel, & Sabanović, Selma. 2020. *Human-Robot Interaction: An Introduction*. Cambridge: Cambridge University Press.

Basu, Chandrayee, Singhal, Mukesh, & Dragan, Anca D. 2018. Learning from Richer Human Guidance: Augmenting Comparison-Based Learning with Feature Queries. *Page 132–140 of: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '18. New York, NY, USA: Association for Computing Machinery.

Braun, Virginia, & Clarke, Victoria. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, **3**(2), 77–101.

Doering, Malcolm, Liu, Phoebe, Glas, Dylan F., Kanda, Takayuki, Kulić, Dana, & Ishiguro, Hiroshi. 2019. Curiosity Did Not Kill the Robot: A Curiosity-Based Learning System for a Shopkeeper Robot. *J. Hum.-Robot Interact.*, **8**(3).

Dragan, Anca, & Srinivasa, Siddhartha. 2014. Familiarization to Robot Motion. *Page 366–373 of: Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '14. New York, NY, USA: Association for Computing Machinery.

Fitter, Naomi T., Chowdhury, Yasmin, Cha, Elizabeth, Takayama, Leila, & Matarić, Maja J. 2018. Evaluating the Effects of Personalized Appearance on Telepresence Robots for Education. *Page 109–110 of: Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '18. New York, NY, USA: Association for Computing Machinery.

Gielniak, Michael J., & Thomaz, Andrea L. 2011. Spatiotemporal Correspondence as a Metric for Human-like Robot Motion. *Page 77–84 of: Proceedings of the 6th International Conference on Human-Robot Interaction*. HRI '11. New York, NY, USA: Association for Computing Machinery.

Gielniak, Michael J., & Thomaz, Andrea L. 2012. Enhancing Interaction through Exaggerated Motion Synthesis. *Page 375–382 of: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*. HRI '12. New York, NY, USA: Association for Computing Machinery.

Glen, S. *Average Inter-Item Correlation: Definition, Example*. shorturl.at/oFUWX.

Hanheide, Marc, Hebesberger, Denise, & Krajník, Tomáš. 2017. The When, Where, and How: An Adaptive Robotic Info-Terminal for Care Home Residents. *Page 341–349 of: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '17. New York, NY, USA: Association for Computing Machinery.

Hinds, Pamela J, Roberts, Teresa L, & Jones, Hank. 2004. Whose job is it anyway? A study of human-robot interaction in a collaborative task. *Human-computer interaction*, **19**(1-2), 151–181.

Hoffman, Guy. 2019. Evaluating Fluency in Human–Robot Collaboration. *IEEE Transactions on Human-Machine Systems*, **49**(3), 209–218.

Jacq, Alexis, Lemaignan, Séverin, Garcia, Fernando, Dillenbourg, Pierre, & Paiva, Ana. 2016. Building successful long child-robot interactions in a learning context. *Pages 239–246 of: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

Javed, Hifza, Burns, Rachael, Jeon, Myounghoon, Howard, Ayanna M., & Park, Chung Hyuk. 2019. A Robotic Framework to Facilitate Sensory Experiences for Children with Autism Spectrum Disorder: A Preliminary Study. *J. Hum.-Robot Interact.*, **9**(1).

Jayaraman, Suresh Kumaar, Creech, Chandler, Robert Jr., Lionel P., Tilbury, Dawn M., Yang, X. Jessie, Pradhan, Anuj K., & Tsui, Katherine M. 2018. Trust in AV: An Uncertainty Reduction Model of AV-Pedestrian Interactions. *Page*

133–134 of: Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction. HRI '18. New York, NY, USA: Association for Computing Machinery.

Kruse, Thibault, Kirsch, Alexandra, Khambhaita, Harmish, & Alami, Rachid. 2014. Evaluating Directional Cost Models in Navigation. *Page 350–357 of: Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '14. New York, NY, USA: Association for Computing Machinery.

Kwon, Minae, Huang, Sandy H., & Dragan, Anca D. 2018. Expressing Robot Incapability. *Page 87–95 of: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '18. New York, NY, USA: Association for Computing Machinery.

Kwon, Minae, Biyik, Erdem, Talati, Aditi, Bhasin, Karan, Losey, Dylan P., & Sadigh, Dorsa. 2020. When Humans Aren't Optimal: Robots that Collaborate with Risk-Aware Humans. *Pages 43–52 of: 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

Mavrogiannis, Christoforos I., Thomason, Wil B., & Knepper, Ross A. 2018. Social Momentum: A Framework for Legible Navigation in Dynamic Multi-Agent Environments. *Page 361–369 of: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '18. New York, NY, USA: Association for Computing Machinery.

Menne, Isabelle M., & Lugrin, Birgit. 2017. In the Face of Emotion: A Behavioral Study on Emotions Towards a Robot Using the Facial Action Coding System. *Page 205–206 of: Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '17. New York, NY, USA: Association for Computing Machinery.

Murakami, Ryo, Morales Saiki, Luis Yoichi, Satake, Satoru, Kanda, Takayuki, & Ishiguro, Hiroshi. 2014. Destination Unknown: Walking Side-by-Side without Knowing the Goal. *Page 471–478 of: Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '14. New York, NY, USA: Association for Computing Machinery.

Nikolaidis, Stefanos, Kwon, Minae, Forlizzi, Jodi, & Srinivasa, Siddhartha. 2018. Planning with Verbal Communication for Human-Robot Collaboration. *J. Hum.-Robot Interact.*, **7**(3).

Oudah, Mayada, Babushkin, Vahan, Chenlinangjia, Tennom, & Crandall, Jacob W. 2015. Learning to Interact with a Human Partner. *Page 311–318 of: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. HRI '15. New York, NY, USA: Association for Computing Machinery.

Pallant, Julie. 2011. *Survival manual*. Vol. 4.

Rakita, Daniel, Mutlu, Bilge, Gleicher, Michael, & Hiatt, Laura M. 2018. Shared Dynamic Curves: A Shared-Control Telemanipulation Method for Motor Task Training. *Page 23–31 of: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '18. New York, NY, USA: Association for Computing Machinery.

Rouanet, Pierre, Danieau, Fabien, & Oudeyer, Pierre-Yves. 2011. A Robotic Game to Evaluate Interfaces Used to Show and Teach Visual Objects to a Robot in Real World Condition. *Page 313–320 of: Proceedings of the 6th International Conference on Human-Robot Interaction*. HRI '11. New York, NY, USA: Association for Computing Machinery.

Sanghvi, Jyotirmay, Castellano, Ginevra, Leite, Iolanda, Pereira, André, McOwan, Peter W., & Paiva, Ana. 2011. Automatic Analysis of Affective Postures and Body Motion to Detect Engagement with a Game Companion. *Page 305–312 of: Proceedings of the 6th International Conference on Human-Robot Interaction*. HRI '11. New York, NY, USA: Association for Computing Machinery.

Schulz, Trenton, Torresen, Jim, & Herstad, Jo. 2019. Animation Techniques in Human-Robot Interaction User Studies: A Systematic Literature Review. *J. Hum.-Robot Interact.*, **8**(2).

St. Clair, Aaron, & Mataric, Maja. 2015. How Robot Verbal Feedback Can Improve Team Performance in Human-Robot Task Collaborations. *Page 213–220 of: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. HRI '15. New York, NY, USA: Association for Computing Machinery.

St-Onge, David, Côté-Allard, Ulysse, Glette, Kyrre, Gosselin, Benoit, & Beltrame, Giovanni. 2019. Engaging with Robotic Swarms: Commands from Expressive Motion. *J. Hum.-Robot Interact.*, **8**(2).

Totsuka, Ryusuke, Satake, Satoru, Kanda, Takayuki, & Imai, Michita. 2017. Is a Robot a Better Walking Partner If It Associates Utterances with Visual Scenes? *Page 313–322 of: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '17. New York, NY, USA: Association for Computing Machinery.

Venture, Gentiane, & Kulić, Dana. 2019. Robot Expressive Motions: A Survey of Generation and Evaluation Methods. *J. Hum.-Robot Interact.*, **8**(4).