The effect of input modalities, different levels of task complexity and embodiment on users' overall performance and perception in human-robot collaboration tasks.

Master's Thesis

to confer the academic degree of Dipl.-Ing.
from the Faculty of Natural Sciences at the
Paris-Lodron-University of Salzburg

by
Gerald Stollnberger, BEng.

Supervisor of the thesis:
Tscheligi, Manfred, Univ.Prof., Dr.

Advisor of the thesis:
Weiss, Astrid, Dr.

Department of Computer Science

Salzburg, February 21, 2014

Acknowledgment

At this point I want to thank everyone who supported me in writing this thesis.
First of all I want to thank my professor and supervisor Univ.-Prof. Dr. Manfred Tscheligi. Without him, this thesis would not have been written.
Furthermore, I want to especially thank Dr. Astrid Weiss, my advisor for her encouragement, helpfulness, and motivation. She spent much time on this project, and also evoked me on this topic and robotics in general.
I want to thank my supervisors for giving me the chance to connect hobby (to build and program robots) with work, to have a job in the research, and to get to know the robotics community.
My thanks also go to my colleagues from the ICT&S Center, who always helped me with words and deeds for writing this thesis.
I am also much obliged to the people participating in my studies.
Gratitude also to my mother and father and my sister for their comprehensive support and patience.
Furthermore, to my wife and daughter for their love and understanding in this also sometimes exhausting and stressful process.
Last but not least, I want to thank my friends and my university colleagues, who accompanied me during my academic studies.
My heartfelt thanks go to all of you. I really appreciate your contributions.
Thanks a lot!

Declaration

I hereby declare and confirm that this thesis is entirely the result of my own original work. Where other sources of information have been used, they have been indicated as such and properly acknowledged.

I further declare that this or similar work has not been submitted for credit elsewhere.

Salzburg, February 21, 2014 __________________

Signature

Abstract

Human-Robot Interaction (HRI) is a growing multidisciplinary field of science. The focus of HRI is to study the interaction between robots and humans with a contribution of social science, computer science, robotics, artificial intelligence, and many more.

Nowadays one of the research goals of HRI is to enable robots and humans to work from shoulder to shoulder. In order to avoid misunderstandings or dangerous situations, choosing the right input modality, especially for different levels of task complexity, is a crucial aspect for successful cooperation. It can be assumed, that for specific levels of task complexity, there is always one complementing input modality which increases the corresponding user satisfaction and performance.

In order to identify the most appropriate input modality in relation to the level of task complexity, two user studies are presented in this thesis, as well as the complete prototyping process for the robot.

The first study was in a public space where participants could choose between two different input modalities (a PC-remote control and a gesture-based interface) to drive a race with a Lego Mindstorm robot against the other participant. The second study was conducted within a controlled laboratory environment where participants had to solve three assembly tasks, which differed in complexity. In order to get the required parts, they had to use three different input modalities (the PC-remote and a gesture-based interface from the first study plus a speech control interface). Besides investigating the correlation of task complexity and input modality, it was explored in this second study, whether the appearance of the robot also has an impact on how the collaboration is perceived by the user.

One of the main findings was that all of these factors (input modality, level of task complexity as well as the appearance of the robot) had a severe impact on the results. In addition to that, many interdependencies could be identified, especially between input modality and task complexity. Differences in user satisfaction measures and also in performance were often highly significant, especially for hard tasks. Furthermore, it was found, that the perceived task complexity was strongly dependent on users' cognitive workload, driven by the used input modality, which also emphasized the strong coherency of task complexity and input modality. Regarding the influence of the appearance of the robot, it could be shown, that the human-like shape increased users' self confidence, to be able to solve a task together with the robot without any help of someone else.

Contents

A. Appendix

1. Outline

The goal of this work is to investigate whether there exists an interdependency between input modalities and task complexities, and its effect on performance and user satisfaction in human robot interaction (especially in a cooperative scenario). Therefore, three different input modalities were implemented to control a robot prototype, which was realized by using the Lego Mindstorms NXT 2.0 toolkit.

The input modalities used in the user studies were:

- A traditional PC-remote control using the keyboard for commanding the robot,
- a mobile gesture-based interface on an Android mobile which depends on the accelerometer sensor of the mobile,
- and a speech control system using directly the voice of the user.

In order to evaluate the approach, two user studies were conducted involving behavioral and attitudinal measures.

This thesis is structured as follows:

- In the first chapter a short overview about the history of robotics is given, followed by the research objectives and research questions.
- The second chapter depicts the related work, especially regarding the state of the art of input modalities in the field of HRI, as well as the motivation for the used input modalities in this work.
- The third chapter explains the process and design of the robot prototype used for the two user studies and, moreover, describes the planning and the implementation of the three input modalities. Also some problems during the design process are described.

- The fourth chapter is about the two user studies (one in a public context, and one in a controlled laboratory setup) which were conducted in order to achieve the research goals and to answer the research questions. Firstly, the study setup is described, followed by the results of each study and a short summary of the most important findings.

- In the conclusion, the research questions are answered and discussed.

- The chapter about future work contains a short description about the prospective work, which could be performed in order to deepen the understanding of the impact of task complexity in HRI.

- In the appendix, three already accepted papers of this research, as well as all the questionnaires used in the studies, and the outputfiles of the data analysis are provided.

2. Introduction

The history of robots is a long and diversified one. (cf. [YVR11], [Bed64]) Already in the ancient world first experiments with machines were conducted. Around 420 B.C. Archytas of Tarentum [Arc05], a Greek mathematician, for example, created a wooden steam propelled bird, which was able to fly and was storied to be the first known robot (depicted in Figure 2.1).

Figure 2.1.: Archytas of Tarentum's Dove (mobile.gruppionline.org)

Similarly, Heron of Alexandria [Her99] described more than a hundred machines, for example automatic music machines or theaters in the first century A.D. and earlier.

With the downfall of the ancient cultures, scientific knowledge temporarily disappeared. About 1205 Al-Jazari, an Arabic engineer, wrote a book about mechanical equipment, for example humanoid automata or a programmable automaton band, which influenced Leonardo da Vinci in his research. Da Vinci [Dav06] designed a humanoid robot in the 15th century, but unfortunately, the technical knowledge was not sufficient enough to put such a machine into practice. However, some drafts of his mechanical knight (Figure 2.2) outlasted the centuries.

About 1740 Jacques de Vaucanson [Bed64] designed and implemented a machine, which could play a flute, a first programmable fully automatic loom, and an automatic duck, which can be seen in Figure 2.3 .

Figure 2.2.: Model of a Robot based on Drawings by Leonardo da Vinci (From Wikimedia Commons)

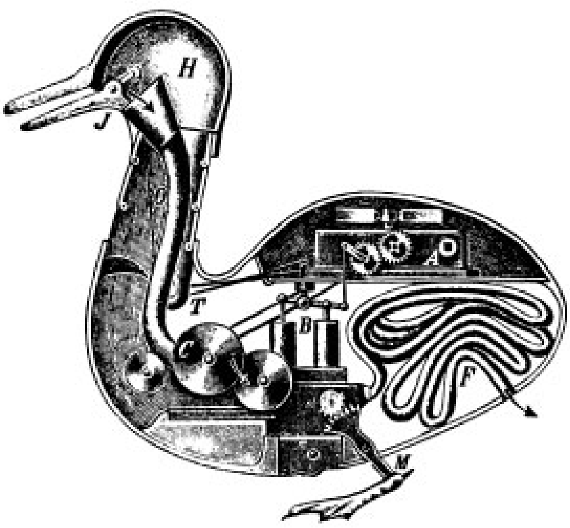

Figure 2.3.: Duck of Jacques de Vaucanson (From Wikimedia Commons)

1941 Isaac Asimov [Asi42], a science fiction author, firstly used the word “robotics” in one of his stories and assumed that robotics is referred to the science and technology of robots. He also became famous for his “three laws of robotics” which he introduced in his story “Runaround”.

1. “A robot may not injure a human being or, through inaction, allow a human being to come to harm.

2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.

3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.”

He also added a fourth one, the “zeroth law”

0. “A robot may not harm humanity, or, by inaction, allow humanity to come to harm.”

After the Second World War, most likely because of the invention of the transistor by Bell Laboratories, robotics made rapid advances from the technological point of view.

According to Matarić [Mat07] who defined a robot as follows:

“A *robot* is an autonomous sysytem which exists in the physical world, can sense its environment, and can act on it to achieve some goals.”

the first real robot was considered to be the “Tortoise” of Grey Walter, because it was the first machine which met the definition.

About 1970 the first autonomous mobile robot Shakey (Figure 2.4) was invented by the Stanford Reseach Institute.

In 1973 at the Waseda-University Tokio [Kat74] the construction of the humanoid robot Wabot 1 was started, and in the same year the German robotics pioneer KUKA built the worldfirst industrial robot known as FAMULUS [KUK73].

1997 the first mobile robot Sojourner landed on the Mars [Soj97].

Nowadays, robots’ domains are manifold. On the one hand there are highly specialized robots for industrial use. They often work in highly adapted environments, strictly separated from

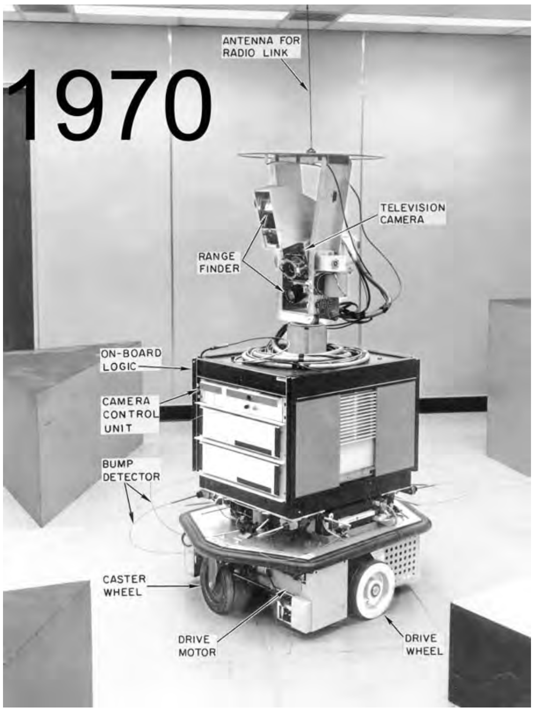

Figure 2.4.: The First Autonomous Mobile Robot (www.frc.ri.cmu.edu)

the humans' workingspace. The main activities of industrial robots are often handcrafts like assembly, painting, inspection or transportation tasks. However, on the other hand also in this context, there is a trend to collaboration with human workers [Ana13]. As a consequence, suiting methods for human-robot communication need to be provided, which is one of the main research goals of this work.

Other application areas, for example, are service robots, which provide attendances often directly to humans, and have to orient themselves in contexts together with human interaction partners, which again shows the importance of ways to communicate with each other in order to avoid misunderstandings or dangerous situations.

Furthermore, the toy industry has robots in their portfolio, for example the robotic dog Aibo from Sony (Figure 2.5), or the Lego Mindstorms, which were used as prototype platforms for the research in this thesis.

Looking like a game, but with a serious scientific background, robot soccer games between teams, consist of the same type of robot, are conducted. The objective of this research is to develop an autonomous humanoid soccer team, which is able to play and win against the current world champion until 2050 [Cup13].

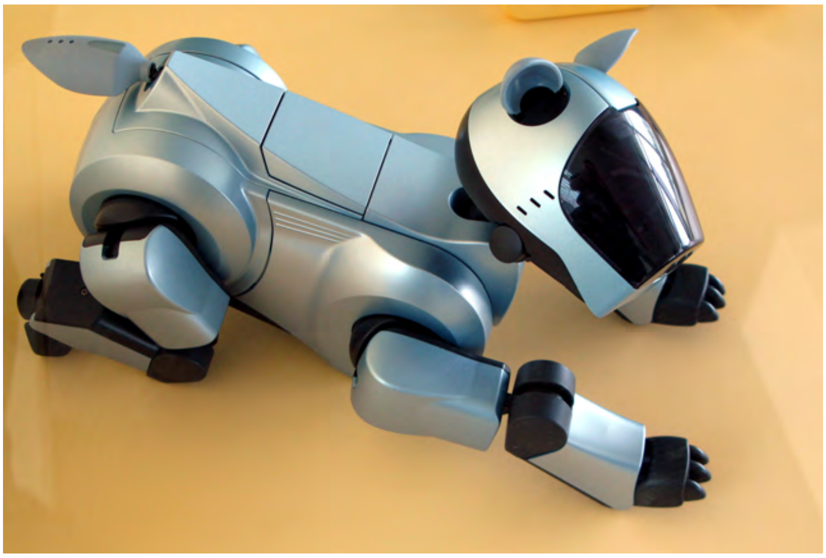

Figure 2.5.: Sony's Robotic Dog Aibo (From Wikimedia Commons)

Robots also help humans doing the housework, for example by mowing the lawn, or by vacuum cleaning. Exploration robots investigate dangerous areas such as planets in a far distance, as well as disaster areas. In medicine, robots support surgery or rehabilitation or simple monitoring tasks in a hospital.

In the military context robots are used too, for example uncrewd drones for spying out dangerous areas.

The fact that in the future robots will affect our living conditions a lot more than today, stresses the importance of enabling a clear communication, without misunderstandings. Additionally, also Hoffmann and Breazeal [HB04] propose, that collaboration between humans and robots has and will become much more important. The research field of Human-Robot Interaction (HRI) has dealt with such research questions for several years now, but there are still many questions to answer. For example, one important aspect in the exploration of human-robot cooperation is to find out which input modalities to communicate with the robot are best suited [RBO09], [RDO11], [GS08] ; especially concerning different levels of task complexity.

One severe problem of Human-Robot Collaboration is often, that not an appropriate input modality is provided for the interaction with the robot. The challenge thereby is that the ideal input modality can change between different situations or various tasks. For example, in a noisy context, a speech control may not be the most suiting tool to interact with a robot, whereas in a quiet surrounding where movement together with the robot is necessary, it could be the ideal choice. The ideal input modality is also dependant on the task which has to be fulfilled, for example if the user needs a hand free for a secondary task, a PC-remote control, where both hands are required for the interaction, would not be the optimal choice, whereas a gesture-based interface which requires just one hand for the interaction, could probably provide the functionality needed for such a situation.

In order to find the best input modality for collaboration, also the complexity level of the task plays an important role. An easy and superficial task like transporting a box straight ahead, could require other input possibilities, in comparision to a more complex task where more precision is needed like a surgery for example. As a consequence, findings of task complexity research were also included for the research in this thesis. This is essential in order to find the best mix of input modality in dependance of task complexity for human-robot collaboration.

2.1. Main Research Questions

Therefore, the main research goal of this thesis is to explore the ideal mix of input modalities depending on different levels of task complexities in terms of performance and user satisfaction measures. In other words, which input modality is the most appropriate for different levels of complexities. For example, a speech control system could possible be the best choice for easy tasks, whereas a gesture-based interface could be more appropriate for complex activities. Furthermore, it is investigated if minimal human-like cues, added to a purely functional robot in order to suggest anthropomorphism, have an effect on the interaction/collaboration.

2.2. Research Sub-Questions

1. What interdependencies between input modalities and task complexities can be identified?

2. How do users perceive the different input modalities in terms of user satisfaction measures?

3. Are the means of interaction provided by the different input modalities effective for the human and the robot?

4. Are the means of interaction provided by the different input modalities efficient for the human and the robot?

5. How do the different input modalities perform for different task complexities?

The research sub-questions are used in order to assess the main research goal and to provide a broader perspective about the research in this work. The research questions will be answered in Chapter 6.

However, the first step to reach the research goal, was an intense search for literature, which was used as a starting point for this thesis. The following chapter illustrates the most important work, which is related to this area of research.

3. Related Work

In order to enable successful human-robot collaboration, many factors have to be considered:

- The context in which the interaction takes place,
- the type of the robot,
- the experience of the user,
- the task which has to be fulfilled,
- and many other circumstances.

Especially the chosen input modality used for the cooperation is an essential aspect to enhance a satisfying interaction as there are many different possiblities to interact with robots nowadays. Much research effort has already been spent on investigating in different ways for commanding robotic systems:

- Different handheld devices [RDO11],
- tangible interfaces [GS08] or
- or different control modalities at the same time [BGR$^+$09].

Some researchers also investigate in using speech for commanding robotic systems [CSSW10] [AN04]. All in all, each input modality has advantages and disadvantages in specific situations, therefore further investigation is necessary.

Cantrell et al. [CSSW10] for example worked on the issue of teaching natural language to an artificial system. Therefore they conducted a human-human interaction experiment with the

purpose to collect data about verbal utterances and disfluencies of natural language. They mentioned, that most of the speech recognition systems are based on a sequential approach and just make limited use of goal structures, context and task knowledge, which is indeed essential for achieving a natural language like interaction. They identified a number of problems as disfluencies, omissions, grounding of referents and others which makes it complicated for an artificial system to understand natural language. Although they believed, that "no HRI architecture is currently capable of handling completely unrestricted natural language" they proposed a very promising integrated architecture for robust spoken instruction understanding using incremental parsing, incremental semantic analysis, disfluency analysis, and situated reference resolution besides the "classical" speech recognition, with the result, that they were able to handle most of the above mentioned problems quite well. The authors consider their work to be "an important step in the direction of achieving natural human-robot interactions". All in all, their system can handle a wide variety of spoken utterances, but it is still not free of errors.
Therefore, it was decided to use a sequential approach with strictly predefined commands for the speech recognition system in the user study in this thesis, to avoid some of the above mentioned problems, for example, the grounding of referents. Furthermore it is mentioned that just a small set of commands will improve the accuracy of speech recognition systems, which will totally meet the requirements for the studies in this thesis.

Ayres and colleagues [AN04] implemented a speech recognition system, which they studied by using Lego Mindstorm robots. They implemented a system, which could control a robotic device, with a JVM, via a Wifi/WLAN/TCPIP network from a PC workstation or a mobile handheld PC (iPaq). They mention, that traditionally speech recognition systems are mainly written in C, but due to recent advances in Java, especially in the Java Speech API, this programming language would also be suitable for speech and recognition engineering. In Figure 3.1 several implemented architectures are depicted. All of them have in common that firstly the voice signal is aquired and processed by a PC Workstation or the mobile handheld PC (iPaq), and secondly all of them are using different kinds of speech recognition systems like Sphinx 2 or the JSAPI Cloudgarden API. The aquired and recognised speech signals are sent to a Linux Server via WLAN then, which searches the grammar for possible commands and redirects it again via WLAN to the robot.

The results were quite satisfying, and they've proved their hypothesis, that Java has become a suiting tool for speech and recognition engineering, however, further implementation and testing work is required. Furthermore, they also mentioned the possibility of using Bluetooth for communication instead of WLAN, but due to the lower range and bandwith, and the higher difficulty to program, as Java APIs were not readily available, they claimed, that

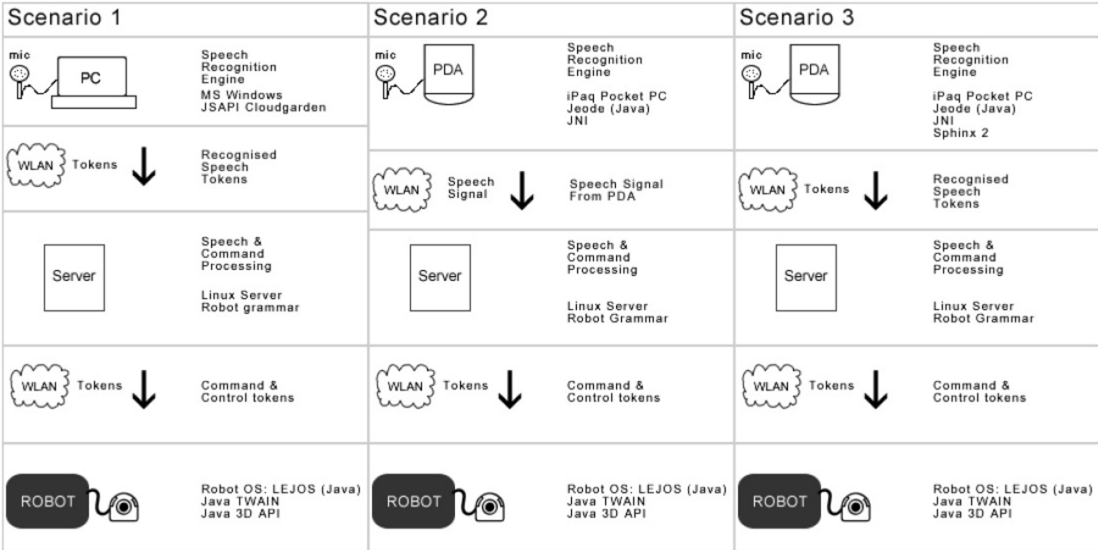

Figure 3.1.: The Different Speech Recognition Architectures (cf. Ayres [AN04], S. 3)

WLAN "provides a more effective networking solution".
However, for the studies in this thesis, the range and bandwith is not very important, therefore it was decided to explore a similar speech recognition approach, combining the advantages of Bluetooth with the plattform independent Java programming language.

Obviously, speech is just an excerpt of the huge amount of possibilites to interact with robots. Other input modalities such as gesture or keyboards etc. have shown promising results, too. Rouanet and colleagues [RBO09], for example, compared three different handheld input devices, in order to control Sony's Aibo robot and to show it different objects. For this, they inspected a keyboard-like and a gesture-based interface on the iPhone, and furthermore a tangible gesture-based interface using Nintendo's Wii. The reason why they chose the AIBO robots was that "Due to its zoomorphic and domestic aspect, non-expert users perceived it as a pleasant and entertaining robot, so it allows really easy and natural interaction with most users and keeps the study not too annoying for testers". They conducted an experiment with non-expert users in a domestic environment, where participants had to control the robot through two tracks, which can be seen in Figure 3.2, which differed in complexity. The results showed, that all input modalities were rather equally efficient and all considered as satisfying. Especially the iPhone gesture-based interface was preferred by the users, whereas the Wiimote was rather poorly rated, although especially for the hard course, the mean completion time was slightly lower with it. According to the authors, this advantage could be explained by the user's ability to always focus on the robot. All in all, their input modalities suffered from a little delay, which means that the reaction of the robot took a few milliseconds, until it actually happened, after a command was given. Therefore, they added visual feedback in

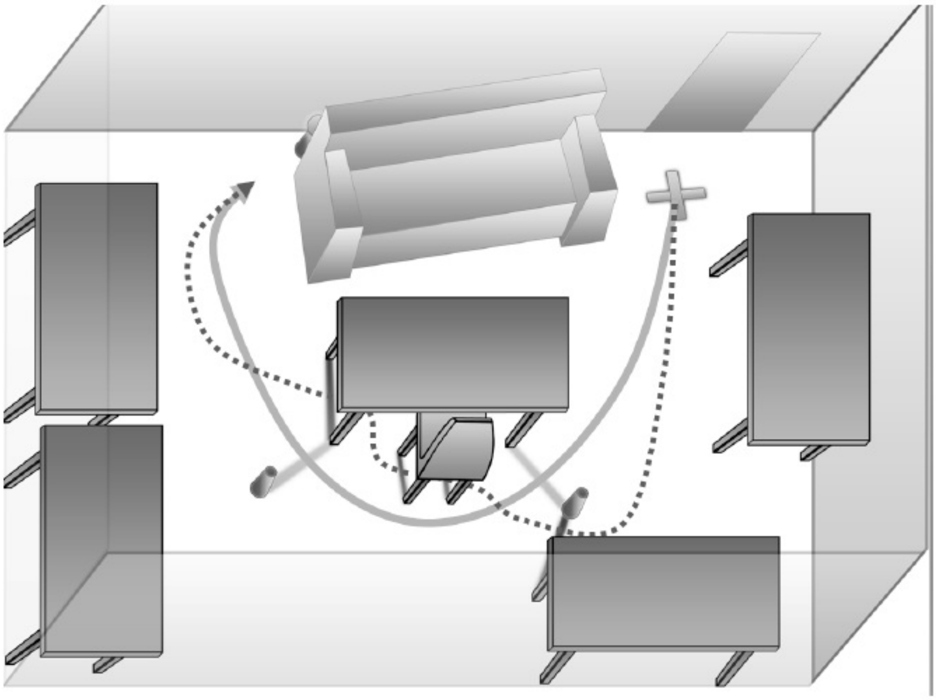

Figure 3.2.: The Two Obstacle Courses, the Dotted Line is the Hard One, the Other Line the Easy One (cf. Rouanet [RBO09], S. 5)

form of lights on the head of the robot, to add an abstraction layer to help users to enable a comparison of the interfaces themselves, instead of their underlying system.

Rouanet et al. [RDO11] continued their research and applied a real world user study with the aim, to teach robots new visual objects. Three input devices based on mediator objects with different kinds of feedback (Wiimote, Laser Pointer, and iPhone) where used. Furthermore, a gesture-based interface was simulated with a Wizard-of-Oz recognition system, to provide a more natural way of interaction. The iPhone interface was based on a video stream of the robot's camera, which allowed participants to monitor the view of the robot directly. However, it was mentioned that the splitting of direct and indirect monitoring of the robot increased the user's cognitive load. The Wiimote interface was based on the directional cross to move the robot. The laser pointer interface worked with the light to draw the robot's attention directly in one direction. Both of the interfaces (Wiimote and Laser Pointer) allowed participants to directly focus their attention on the robot. The gesture-based interface, using a Wizard-of-Oz framework, provided the possibility to guide the robot by hand or arm gestures (Figure 3.3). They designed their study as a robotic game, with the purpose to maintain the users' motivations. During their experiment, they found out that although the gesture-based interface had a lower usability, it was more entertaining for the users. It also increased the feeling of cooperation between the participants and the robot, whereas the usability was

Figure 3.3.: Robot Guided by Gestures (cf. Rouanet [RDO11], S. 4)

better when using the iPhone interface, especially for non-expert users, which could possibly be because of the visual feedback of the device.
All in all their work showed that although feedback increases usability, it could be negative for the cognitive workload on the other hand. Due to the fact that the gesture-based interface, which had the lowest usability was considered to be the most entertaining modality for participants, which could offer a better overall user experience, further investigation in this kind of input modality makes sense.

Cheng Guo et al. [GS08] suggested a tangible user interface for human-robot interaction with Sony's robotic dog AIBO. They used two input devices, namely a keypad interface and a gesture based interface with a Nintendo Wii controller. They believed, that a more natural way of interaction can be achieved by allowing users to focus their attention on more high level task planning in comparison to low level steps. Therefore a suitable input device is needed. They conducted a study, comparing the gesture-based interface and the keyboard interaction device, solving two different tasks with two difficulties each:

- Navigation task: Participants were asked to navigate the AIBO robot through two obstacle courses. For the easy task, participants could freely choose the combinations of actions they needed. When solving the hard task, they were forced to use rotation and strafing, in addition to walking and turning, for a successful completion of the course.

- Posture task: Users had to perform a number of different postures with the forelegs of the robot (cf. Figure 3.4). In this case, only one foreleg had to be manipulated, whereas for the hard tasks, both forelegs had to be transformed, in order to complete the transition.

One of their research questions was, if gesture-based input methods could provide a more efficient human-robot interaction, in comparision to more conservative devices like keyboard, mouse or joysticks. They mentioned, that the most important advantage of their suggested

Figure 3.4.: Possible Postures for each Foreleg of the AIBO Robot (cf. Guo [GS08], S. 6)

tangible user interface was that it provided affordances of physical objects. This advantage was explained because participants “do not need to focus on their hands while performing a posture. They are naturally aware of the spatial location of their hands.” The results of their user studies showed, that the gesture-based input outperformed the keyboard device in terms of task completion times in both types of tasks, and that it was a more natural way of interaction. The fact, that the keyboard interface required more attention shifts between the robot and the device, is another argument for the Wii control. Although most of the users mentioned, that they are more familiar with the keyboard - due to their computer game experiences - they would prefer the gesture-based interface because of the higher intuitiveness. On the other hand, it was mentioned, that if there had been a training session for both devices, the keypad would have outperformed the gesture-based interface. Furthermore, the authors argued, that their results are not always transferrable for these types of tasks, because they did not use the most intuitive mapping of keys on the keyboard, as in Figure 3.5, which could have effected their data.

Figure 3.5.: The Mapping of the Keys (cf. Guo [GS08], S. 6)

Thus, that participants believed, that a training session would have strengthened the results of the keypad interface. Furthermore the statement that not the most intuitive mapping of keys on the keyboard was used, and the fact that some kind of a tangible user interface (the Wiimote) allowed users a more high-level task planning, leaded to investigate in that kind of input modalities further in the studies.

Bannat et al. [BGR$^+$09] proposed that a more natural and flexible human-robot interaction could be achieved by providing different control modalities at the same time, namely speech, gaze and so-called soft buttons. So users can choose, which modality suits best in a specific situation, for example to use speech when both hands are needed for an assembly step (Cf. Figure 3.6). They conducted a study in a factory setup: The worker had to solve a hybrid assembly scenario, in which the assembly steps should be solved in collaboration with the robot serving as an assistant or also as a fully autonomous assembly unit. The aim of their research is to "establish the so-called Cognitive Factory as the manufacturing concept for the 21^{st} century." Their approach will be evaluated in experiments with humans in their future work.

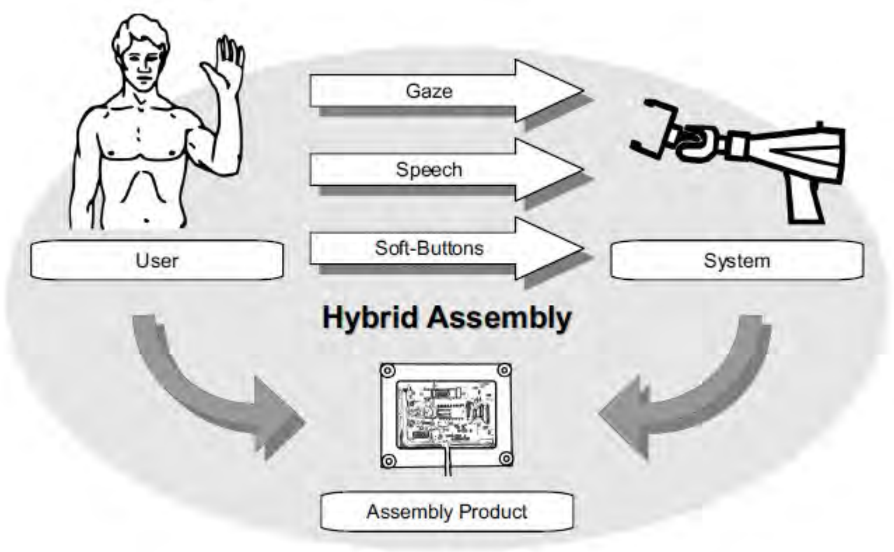

Figure 3.6.: Three Independent Communication Channels Between Human and Robot (cf. Bannat [BGR$^+$09], S. 2)

However, this concept has not been experimentally validated so far and the possibility of using more than one input modality at the same time could possibly lead to a higher cognitive workload.

Hayes et al. [HHA10] tried to develop an optimized multi-touch interface for human-robot interaction, using a map on a screen to control the robot. According to the researchers, static input devices like mouse and keyboard do often not result in a satisfying interaction form to command robots, especially in situations, where the user needs hands free for a secondary task, or if it is necessary to walk around together with the robot. Their multi-touch interface resulted in a better usability, faster completion times and in a lower overall workload and reduced frustration, in comparision to static input devices like mouse and keyboard. They believed, that their work was just a first step to developing a successful

multi-touch interaction paradigm for human-robot interaction and they are willing to develop additional touch capabilities. They argued, that it is essential to evaluate their interface with participants, who are on the move during the interaction.
Their work leads again to take more flexible input devices like a gesture-based mobile interface or a speech control system into account. To evaluate their usefulness, it is inevitable to compare them to static input modalites, such as a PC-remote.

Another approach was investigated by Pollard and colleagues [PHRA02]. They evaluated the possibility of using pre-recorded human motion and trajectory tracking as an input modality for anthropomorphic robots. They used a system with several cameras to gather data about the participants' gestures, and conducted an experiment with professionally trained actors to capture their motions, in order to transfer them to a Sarcos robot like in Figure 3.7. They came to the conclusion, that the biggest limitation of their approach was that the degree of freedom of each joint was inspected separately. This sometimes led to situations, where one joint exceeded its limit and another did not, which produced results, that the motion of the robot did not match the motion of the actor. They planned to make a second experiment, where they want to show participants different instances of actors, performing given motions and let the participants decide, whether the robot motion has been driven from the actors motion or not. An evaluation will show, if the robots motions have been retained successfully from the actors' movement.

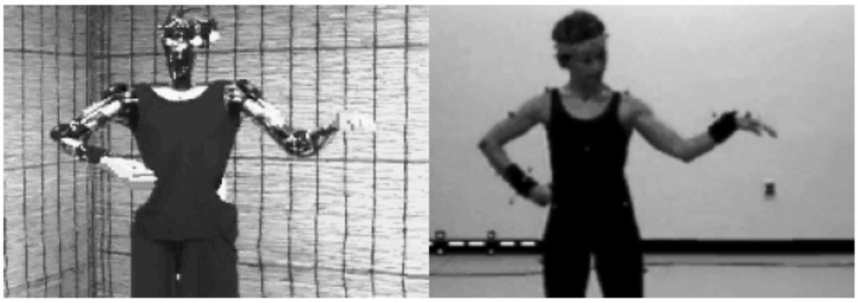

Figure 3.7.: The Sarcos Robot and a Professional Actor (cf. Pollard [PHRA02], S. 1)

On the one hand, the approach seemed to be very innovative and promising, on the other hand, it seems to take a lot of time to record and process human motions, which is furthermore a very complex activity. Therefore, this kind of interaction paradigm will not be used in the studies.

Although there has been a lot of research done in human-robot collaboration and especially on how to control robots, there are still many things to discover. Summarizing, many researchers mentioned that speech is the most natural way of interaction between humans, therefore, it could possibly be the most suiting control method for robots too. Several experiments on speech as an input modality for robots achieved quite positive results [AN04] [CSSW10] [WBS+09] [BGR+09], but the studies and measures were often not very detailed or not tested in a complete user study. Therefore further investigation makes sense, especially in collaboration scenarios where increasing usability and user experience could provide an immense outcome for companies in terms of time, money and overall satisfaction of the workers. For the studies presented in this thesis, it is not needed to implement an interface, which parses natural language, but just about five commands have to be understood, which increases the robustness of such a system a lot. Understandably, there have to be comparative input systems for the purpose to find arguments, whether a given input device is suitable for different scenarios like superficial or more complex tasks. The other two input modalities, which are investigated, are an Android application as a gesture-based interface and a PC remote control. All in all three very different devices, that offer the possibility to find well-founded arguments for comparing them, are investigated. Furthermore, it is a fact, that also different appearances of robots could have an impact on the results, like in the work of Groom et al. [GTON09]. As a consequence this factor will also be taken into account.

All in all, what has been - up to the literature review - left out so far, is a structured investigation of the interplay between: (1) input modality, (2) task complexity, and (3) appearance of the robot (functional vs. human-like). Which means that it is essential, not to investigate these factors just separately but combined in stuctured user studies. Therefore, especially interdependencies between them are in the focus of this research. The goal is to find the most appropriate mixture of input device and appearance of robots for different levels of task complexity, in terms of the resulting user satisfaction and the overall performance.

In order to assess these interdependencies in user studies, a robot prototype, as well as various input modalities to study were needed. The design as well as the implementation of the prototype and it's input modalities are described in the following chapter.

4. Prototype

4.1. The Purely Functional Robot

For the research presented in this thesis a robot was needed, which is able to fulfill pick and place and transportation tasks. These kind of jobs are often common in the context of a factory for example and needed for collaborative interaction scenarios. It was decided to use Lego Mindstorms NXT 2.0 [Wik13b], which were launched on August 5 2009, due to several reasons.

The Lego Mindstorms were developed through a partnership between Lego and the MIT Media Laboratory, also with the goal, to be used, besides as a toy, as teaching material for programming. It is a good example for an embedded system, which consists of a programmable brick computer (with the possibility to use four sensors and three motors concurrently, with a 32-bit microprocessor), three servomotors, different sensors (like an ultrasonic sensor to measure distances, a light sensor to gather information about six different colours, or the intensity of light, 2 touch sensors which react to button presses, a sound sensor which enables the robot to receive information about noise, and many more), and Lego technique parts. (cf. Figure 4.1) Furthermore, the robot is able to provide aural feedback by a 8-bit speaker and also visual output with the aid of a 100x64 pixel display.

It is possible to copy nearly any thinkable mechanic system and also to construct autonomous or interactive systems with Lego Mindstorms, which led 2008 to the acceptance for the robot hall of fame [Wik13c]. In addition to that, it is a more or less affordable robot, also available for the consumer market. Although it is primarily a toy, it provides the freedom to use nearly any of the most common programming languages, for the purpose to write routines for the robot. Summarizing, the flexibility, affordability, and quality has led to the decision to use Lego Mindstorms for this research.

A first functional prototype was constructed, which was able to drive around by using two

Figure 4.1.: The Programmable Brick Computer with Motors and Sensors Attached (cf. Lego.com [Leg13])

motors. The problem appeared, that there was just one motor left to enable the robot to pick and place things, which was essential for the user studies. As a consequence, an intensive search for literature and ideas followed, with the discovery of “The Snatcher” of Laurens. It is a functional robot with a grabbing arm, which was an inspiration for using this grabbing arm for the prototype too. In his book [Val10], Lauren provides a detailed description how to build this grabbing arm, with just one motor needed. This becomes possible by a clever combination of gears, beams, and transmission of power. The arm worked well, but after a few tests, strange noises were noticed, produced by the motor of the grabbing arm, when reaching the highest or lowest position of the arm. This was caused by the attempt of the motor, to overcome the resistance of physical barriers. In order to avoid damages to the motor, a minimum and a maximum point were defined, to restrict the movement of the grabbing arm motor. This was carried out by using the ultrasonic sensor and one touch sensor, to indicate the respective end points (cf. Figure 4.2). For example, when the ultrasonic sensor reported a distance lower than 1cm, the grabbing arm was on the lowest allowed position, and the motor should stop. On the opposite side, the touch sensor was pressed, when the arm reached the highest position and triggered the arm to stop.

Furthermore a sound sensor was added to the robot, to enable the robot to receive information about noise and of course sound commands. Unfortunately, the sound sensor was only capable of measuring the volume, and it was not managed to record actual sounds with it. The result of the prototyping process led to a purely functional robot design, which can be seen in Figure 4.3.

After a few short tests it was realised that the road grip was not adequate enough, as a consequence it was hard to precisely control the robot. This problem led to the decision, to change from wheels to chains which improved the handling a lot.

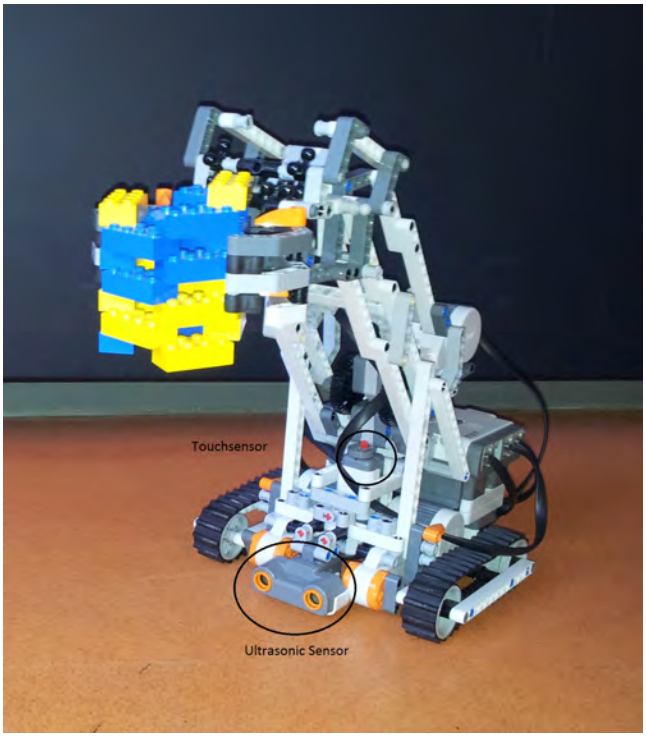

Figure 4.2.: The Ultrasonic Sensor Indicates the Minimum Point, the Touch Sensor the Maximum Point

Figure 4.3.: The Functional Prototype with the Grabbing Arm

4.2. The Anthropomorphic Robot

In order to get an insight into the impact of the appearance of the robot on the interaction, it was investigated, if minimal anthropomorphic cues, added to the robot, would lead to differences in the results. Therefore, the appearance of the robot was altered by adding a head to the purely functional robot, to indicate anthropomorphism. The question was, if minimalistic cues are sufficient to effect the results concerning user satisfaction measures, or even the performance. The head was an outcome of a 3D-printing workshop, which took place at the Computer Science Departement of the University of Salzburg and was provided by a member of Otelo [Ote13]. In the workshop, which lasted two days, three fully working 3D printers were built and tested and for the head of the robot a model, designed by Neophyte, was used.[1] The origin of the head can be seen in Figure 4.4.

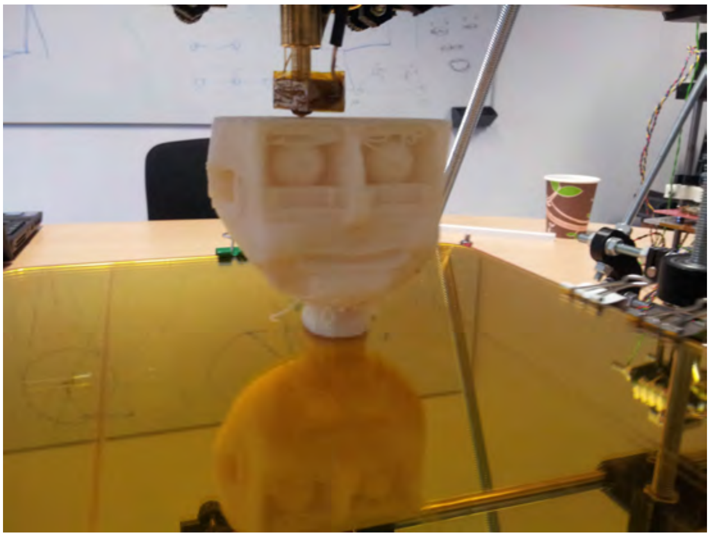

Figure 4.4.: The 3D Printer is Processing the Robot's Head

The head was then mounted on the functional prototype by using some lego bricks. In Figure 4.5, a first concept and the prototype with anthropomorphic cues is depicted.

4.3. The Input Modalities

After the robot prototype was finished, three different input modalities were implemented. All of them should have been able to control the robot to drive around in any desired direction, and furthermore to lift and release small objects by using the grasping arm of the prototype.

[1] http://www.thingiverse.com/thing:8075

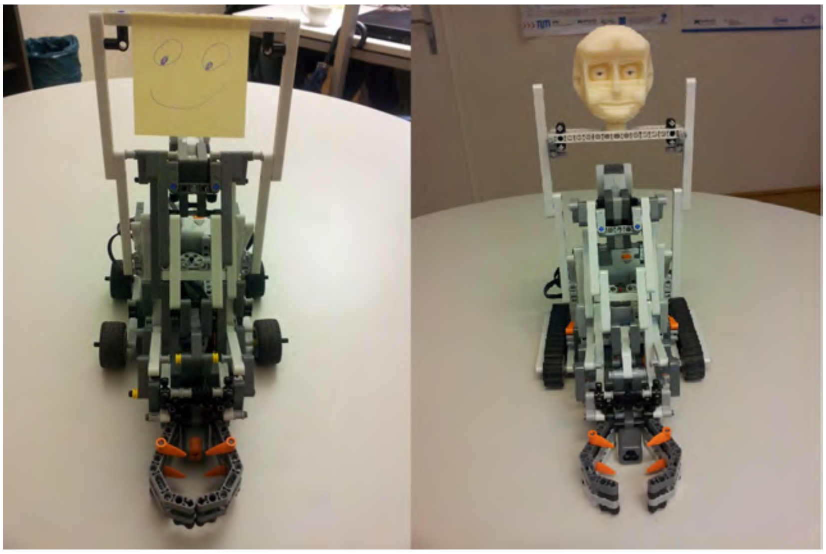

Figure 4.5.: A first Concept and the Prototype with Anthropmorphic Cues

Following the literature review in Chapter 3, it was concluded to provide

- a gesture-based interface on an Android mobile phone,
- a speech control system,
- and, last but not least a “classical” PC-remote interface, using a standard keyboard for the interaction with the robot.

In the following section, an overview about how the different input modalities were implemented, and what and especially why some technologies were used, is given.

4.3.1. The Mobile Gesture-based Interface

During the conception of the different input possibilities, the MINDdroid application was conquered, which offered nearly all of the functionalities, which were needed for the gesture-based interface. The executable and also the source code is freely available under the GPL v3 license [2]. The source code can be downloaded at the github webpage [HÏ3].

The robot is controlled by using the data of the accelerometer sensor of the smartphone, in other words, the direction of the robot is triggered by the relative angle of the mobile to the

[2]http://www.gnu.org/licenses/gpl-3.0.html

ground. By using this data, it is possible to control two motors of Lego Mindstorm robots. Furthermore, the application provides visual feedback on the screen and haptic feedback by vibrations. The application also provides the possibility to start programs, which are directly installed on the Lego Mindstorm using the standard firmware or others. This happens by using the touchscreen interface via the action button. As it can be seen in Figure 4.6, the relative position of the yellow square indicates the direction, which the robots is moving to. If the yellow square is within the grey one, it stops. For the communication to the robot and sending the commands, Bluetooth is used.

Figure 4.6.: The MINDdroid Application (cf. github.com [HÏ3])

In order to allow a more exact comparision between the different input modalities, the source code was used and modified to meet specific requirements. The modification was covering a limitation of the maximum speed in order to make it equal to the other input modalities - the speech control interface and the PC-remote control. Performance measures would have had hardly any value, if one input modality allows the robot to go faster or slower, than the others.

Due to the fact, that a Samsung Galaxy S2 mobile phone was used for the studies, the app was compiled for Android 4.1 Jelly Bean with API level 16 [And13], which was at that time the newest version. For all the programming work, the integrated development environment Eclipse 4.2.0 Juno [Ecl13] and the ADT plugin [ADT13] was used, which provides an integrated environment for developers of Android applications within Eclipse.

At this stage, the robot was just able to drive around, therefore, the functionality for the grasping arm was added. For this purpose two small routines were implemented, namely to lift and release the grasping arm. For the routines, the programming language NXT-G, which was provided by the NXT 2.0 standard base toolkit, was used. It features programming by providing an graphical and interactive drag and drop environment, which can be seen in

Figure 4.8. NXT-G is based on LabVIEW, an industrial programming standard, which was created by National Instruments using data flow programming (cf. [Wik13b]). The programs ran directly on the Mindstorm robot, and no change of the standard firmware was necessary, which was indeed another argument for using the NXT-G language.

4.3.2. The Speech Control

Several reasons led to the conclusion, that further investigation in speech control was promising. According to the work of Hearst et al. [Hea11], users prefer to speak rather than type. A good interface can bridge the gap in the back end, which makes systems more usable for less experienced people. Additionally, such an input modality provides the advantage, that it is possible to move along with the robot for a better supervision.

When it came to the implementation of speech control, the first idea was to implement the system directly on the robot, in order to avoid the necessity of other equipment. The flexibility of a directly embedded speech control seemed to be a high advantage, therefore, the soundsensor was attached to the robot and programming with NXT-G was started. The sound sensor in Figure 4.7 is, according to the manifacturer, able to detect noise levels in decibels (frequency range from 3 to 6 khz), as well as to identify sound patterns and tone differences (cf. Lego Shop [LSh13]).

The robot should "understand" 7 commands, turning in each direction, stop, and lift or release the grasping arm. Following Alan Dix [DFAB98], humans are just able to remember 7+-2 chunks of information, due to the limited capacity of short term memory. As a consequence it was decided to use the same number of commands for the robot, which was colloquial sufficient to perform the required navigation and grasping tasks.

Firstly thresholds were defined. The noise values were divided from 1 (silent) to 100 (loud) in order to strain off some of the background noise. Everything more quiet than a noise value of 20 was identified as background noise, and therefore ignored. Noises between 20 and 59 were recognized as sound commands, and for security reasons, any sound louder than 60 triggered the stop command on the robot. This was implemented because in a noisy context the robot could become probably uncontrollable and as a consequence the safest possibility for the robot as well as for the user is to stop the robot automatically.

When the sensor detected a sound within the volume level of sound commands, the robot recorded the sound for half a second, which means that for each millisecond the sound level

Figure 4.7.: The Lego Mindstorms Sound Sensor (cf. [LSh13])

was stored. The outcome was a graph, where the x-value indicated the time and the y-value the noise volume level. By analyzing the peaks of the graph, a differentiation of the sound commands became possible pretty exact. After the result of the analysis of the sound command the respective action on the robot was triggered. After the action was started, the programm started again because of an infinite loop, the whole programm was included.

All in all, the speech recognition worked quite well, but unfortunately just for controlling the robot forward, left, right, and to stop. When it came to adding more commands, which were needed for the grasping arm, the accuracy of the speech recognition decreased a lot. A small excerpt of the NXT-G program can be seen in Figure 4.8, just to get an idea of how an NXT-G program looks like.

As a consequence it was decided to try an alternative approach, namely a program based on Java on a laptop, which meant on the one hand giving up the advantage of a directly embedded system on the robot, but on the other hand provided much more possibilities in programming, and hopefully a higher accuracy and robustness of the speech recognition system. A divison of the interaction between the user and the robot into necessary steps was made from the speech control system's point of view:

1. Aquire the spoken signal
2. Process and recognize the signal
3. Transform the recognized token into a command, the robot is able to handle
4. Send the command to the robot

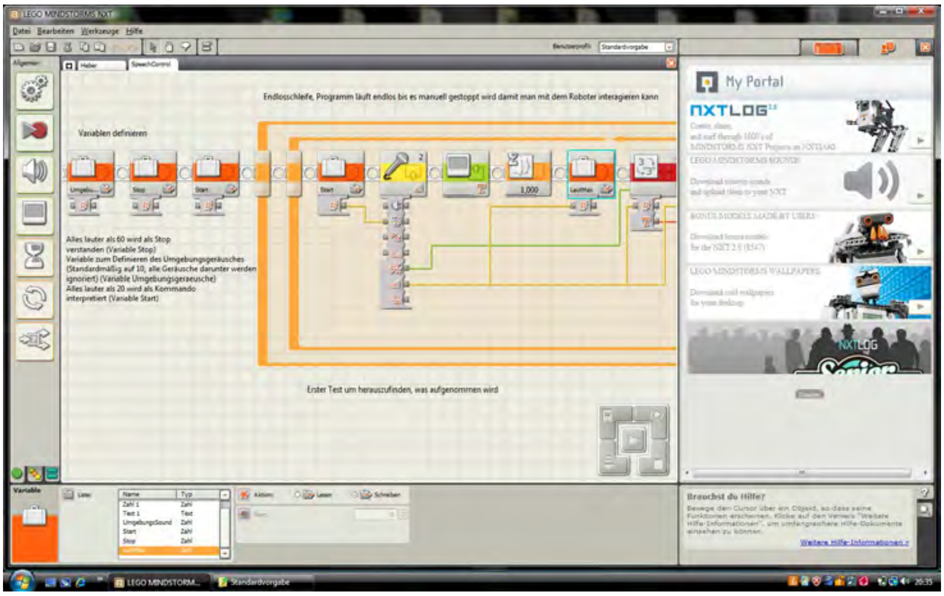

Figure 4.8.: An Excerpt of the First Speech Recognition Approach written in NXT-G

5. Start again with 1.

Step 1 was simply done by using the built-in microphone of a Lenovo Thinkpad R61 Notebook, and the standard Windows microphone. The speech signal was then assigned to the Java programm, where it was processed in **Step 2** by a Speech recognition framework.

There are several open-source speech recognition engines available on the Internet, but it was decided to use Sphinx [Sph13], which was released under the BSD style license by the Carnegie Mellon University and others, and seemed to be very promising [WLK+04].

Sphinx is available in different versions:

- Sphinx 2 is a speaker-independent, real-time large vocabulary especially for mobile applications based on the C programming language.
- Sphinx 3 is more accurate than Sphinx 2, but needs more ressources; also based on C.
- Sphinx 4 is a Java version of the toolkit, therefore, the first choice.
- Pocketsphinx, C-based recognizer library, especially suited for modern smartphones, and newest Sphinx release.
- there are some more versions available which are not further discussed in this thesis.

In 2004 Sphinx 4 was currently in development and suffered many compile errors and no documentation was provided. As a consequence Ayres et al. [AN04] failed to achieve satisfying results. In the last years, Sphinx 4 has made huge advances, hence it was reasonable to use it for the speech recognition system.

The speech recognition relies on a BNF-style grammar[3] in the Java Speech API Grammar Format (JSGF) [JSG13], a platform-independent grammar in textual form, expecially for the use in speech recognition and engineering systems. In speech recognition systems, this grammar needs to be specified, in order to determine, what the system should be capable to understand, and of course to describe the utterances users may and should use. Due to the fact, that 7 commands were necessary, a simple JSGF compatible grammar for the commands "forward", "backward", "pause", "lift", "release", "turn left" and "turn right" was created:

```
#JSGF V1.0;

/**
 * JSGF Grammar for commanding the robot to move
 * around and to lift and release objects
 */

grammar commandRobot;
public <command> = (Forward| Backward| Pause| Lift| Release);
public <turn> =  (turn) (left | right);
```

At the beginning an experimentation with a few possibilities for the words was done. Within the grammar, example given "stop" instead of pause, or "go forward" instead of "forward", was defined, but it was concluded, that the accuracy of the speech recognition was best with the words used in the end. The outcome of **Step 2** was a string, which matched one of the commands which were defined in the grammar.

This particular string was used to decide, which command should be triggered on the robot in **Step 3**. To this end, the program applied the Icommand API v.0.7 [Ico13], which offers the possibility of using Bluetooth for communication with the robot. For completing **Step 4**, Icommand provides several possibilites for Bluetooth communication with Lego Mindstorms, like RXTX, Bluez or Bluecove. The advantage of RXTX is, that it should be possible, to simply adress the robot by the com port (e.g. COM4). Unfortunately the connection

[3]http://en.wikipedia.org/wiki/Backus-Naur_Form

sometimes got lost during the interaction due to unknown reasons. As a consequence, it was switched to BlueCove [Blu13], a Java library which uses directly the Microsoft Bluetooth stack, which worked out really well.

Summarizing, the system consisting of a notebook and the Lego Mindstorm robot, offered the possiblity to control the robot by speech commands.

4.3.3. The PC-remote Control

The third input modality which was implemented, in order to compare them in a human-robot collaboration task, was a "classical" PC-remote, using the keyboard on a notebook, which was a coproduct of the speech control system and also written in Java. During the development of the speech control, a key-listener was implemented for testing and debugging reasons, which was also suitable for the PC-remote. The communication to the robot worked in exactly the same way as in the speech control modality, in other words the Icommand API v.0.7 [Ico13] for parsing the command understandable for the robot, and furthermore Bluecove [Blu13] for the transmission of the command. The only differences were:

- Commands were not triggered by speech, but by button presses on the keyboard.
- Buttons needed to be kept pressed, in order to make the robot move. When the keys were released, the robot stopped, unlike, by using the speech modality, it moved, until it actually received the stop command. For this functionality the functions keyPressed() and keyReleased() from the type keyevent were used:

```
public void keyPressed(KeyEvent e)
{
        int kc = e.getKeyCode();

        switch (kc)
        {
        case java.awt.event.KeyEvent.VK_UP:
                command = COMMAND_FORWARDS;
                break;
        case java.awt.event.KeyEvent.VK_DOWN:
                command = COMMAND_BACKWARDS;
                break;
```

```
            case java.awt.event.KeyEvent.VK_LEFT:
                command = COMMAND_LEFT;
                break;
            case java.awt.event.KeyEvent.VK_RIGHT:
                command = COMMAND_RIGHT;
                break;
            case java.awt.event.KeyEvent.VK_G:
                command = COMMAND_GRAB;
                break;
            case java.awt.event.KeyEvent.VK_R:
                command = COMMAND_RELEASE;
                break;
            default:
                command = COMMAND_NONE;
                break;
        }
    }

    public void keyReleased(KeyEvent e)
    {
        command = COMMAND_NONE;
    }
```

For directing the robot, the arrows on the keyboard were used, whereas for the grasping arm the buttons G ("Grab") and R ("Release") were chosen.

This chapter explained the process and design of the robot prototype which is needed for the two user studies and, moreover, described the planning and the implementation of the three input modalities. Also some problems during the design process were described, such as the capability of the sound sensor or the problem with the Bluetooth communication using RXTX. In the following chapter, the two user studies (one in a public context, and one in a controlled laboratory setup) which were conducted in order to achieve the research goals and to answer the research questions are illustrated. At first, the study setup is described, followed by the results of each study and a short summary of the most important findings.

5. User Studies

To enable a better understanding of human-robot interaction/collaboration, especially about interdependencies between:

- different input modalities,
- different task complexities,
- and different appearances of the robot prototype,

two user studies were conducted. The aim was to identify, which combination of input modality, and appearance of the robot works best in terms of performance and user satisfaction for different human-robot collaboration tasks, which differ in complexity. This knowledge would make it possible to define suggestions and guidelines in terms of suitable input modalities for specific types of tasks and difficulty levels. During the two studies, users were invited to solve different tasks together with the robot.

- At the beginning of June 2012 during the 50th anniversary of the Paris Lodron University of Salzburg, at the ICT&S Center, a preliminary study was conducted, were participants were invited to compete against each other in a Lego Mindstorm race.
- Furthermore, in a controlled laboratory study, users had to perform a collaborative task of building a house out of Lego bricks, together with the robot.

5.1. Preliminary Study in a Public Context

In order to explore how input modality, task difficulty, and the appearance of the robot interplay in terms of performance and user satisfaction, a user study was set-up at the ICT&S Center, during the 50^{th} year anniversary of the Paris Lodron University of Salzburg.

[Results of the preliminary study have already been published as a workshop position paper [SWT12] and also a late-breaking report [SWT13b] was successful submitted and accepted. The publications can be found in Appendix A.1 and Appendix A.2]

5.1.1. Methodology

For the purpose of exposing errors in the study setup and to refine the tasks, a pilot study was arranged with volunteers from the ICT&S Center. Due to the fact that the pilot study was quite successful, in other words, no severe flaws were identified in the study-setup, a study was arranged, in which visitors competed against each other in a Lego Mindstorm race, which was set-up as a 2x2 between-subject design.

The conditions were:

- Task difficulty (Easy / Hard)
- Used input modality (Gesture-based interface / PC-remote control)

Participants firstly had to sign an informed consent [which can be found in Appendix A.3] , in order to give their allowance to use the outcoming data for scientific purposes. After that, they were provided with a short description of the input modalities and the task they had to fulfill. The overall goal was to navigate their robot through a race track, which was filled with obstacles, which the participants had to avoid. There were always two equal tracks in parallel, one for each person. At the end of each track, a Lego box was placed, which had to be transported to the starting point (cf. Figure 5.1).

One participant used the gesture-based interface, whereas the other one the PC-remote control, which offered the possibility to compare the two input devices in terms of efficiency and effectiveness. The participant, who managed to transport the box with the help of the robot first, was the winner of the race. In order to gather data about user satisfaction measures, participants were asked to fill in a questionnaire after the race.

During the day the difficulty of the tracks was changed from easy to hard by adding more obstacles to the track, in order to gain insights on the impact of different task complexities (cf. Figure 5.2).

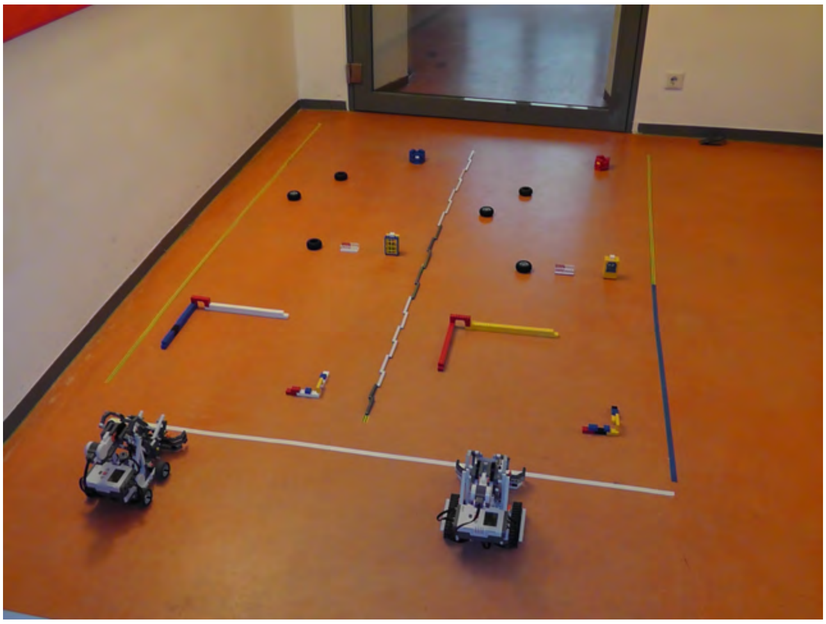

Figure 5.1.: The Two Tracks with the Mindstorm Robots

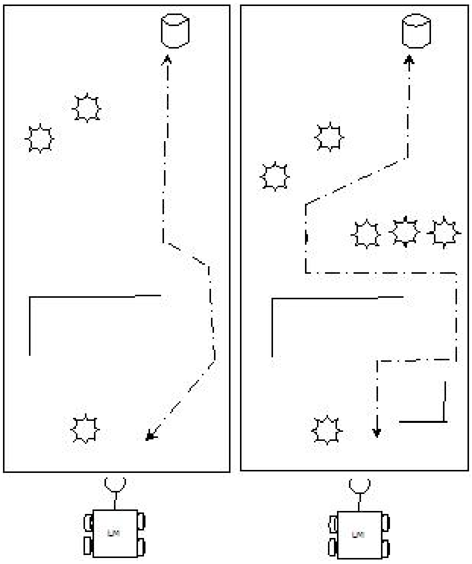

Figure 5.2.: The Easy (left) and the Hard Track (right) was Different in the Number and Placement of Obstacles

5.1.2. Measures

The measured information was driven from the EN ISO 9241-11 Standard [ISO99] which defines three main measures for usability, also proposed by Jakob Nielsen [Nie01].

Based on a further literature review the measures were divided into two categories.

5.1.2.1. Performance Measures (Behavioral Level)

In order to evaluate the potential of productivity with the different input modalities we measured performance concerning two factors:

- Efficiency - time, participants needed to accomplish the course from starting to the end point.

- Effectiveness - determined by counting the number of errors during the navigation based on collisions with other objects or barricades.

5.1.2.2. User Satisfaction Measures (Attitudinal Level)

Besides performance also the attitudinal level is important to strenghten the results of the studies. Therefore, several factors were assessed:

- The perceived task complexity, in order to check, if the manipulation of the track difficulty was successful

- Acceptance, intuitiveness and satisfaction of the used device

The user satisfaction measures were captured by a questionnaire, which was inspired from a survey which was used by a colleague [MBF+09] for the purpose to evaluate the intuitiveness of the Nabaztag robot as an ambient interface. On the other hand, the USUS evaluation framework [Wei10] was used as a source for the questionnaire, which consisted of 15 questions with a 5-point Likert scale. Table 5.1 shows the item of the questionnaire concerning the perceived task complexity. The participant had to check one of the boxes which ranged from "not at all" to "very"; the last option "k.A." meant no answer. Furthermore, participants where asked why they decided their rating, indicated by the line under the checkboxes.

Table 5.1.: The Item "Perceived Task Complexity" with a 5-point Likert Scale

	nicht	wenig	mittel	ziemlich	sehr	k.A.
Die Strecke war schwer zu befahren	☐	☐	☐	☐	☐	☐

Warum? ________________________________

The questionnaire consisted of:

- 1 item for perceived task complexity
- 5 items for intuitiveness
- 4 items for satisfaction
- And 5 items for the acceptance scale

Apart from the user satisfaction measures, the survey also contained information about demographic data, like gender or age, and furthermore which device was used. The survey was filled in within the study area to avoid the participants to be distracted from the change of the context, which could possibly influence the results of the questionnaire. The full questionnaire can be found in Appendix A.4.

Moreover, the study was videotaped, in order to recheck the measured time and number of collisions as well as to identify observable problems with the input devices.

5.1.3. Results

Although it was a challenge to conduct a user study in such an open space context, because of the frequent change of visitors, it offered - on the other hand - the opportunity of studying lots of people with varying socio-demographic background in a short time.

All in all, there were 24 participants participating in the study. 10 of them were female, and 14 male. The youngest participant was 11 years old, whereas the oldest one was 66 years old. The mean age was 36.73 yeas with a standard deviation of 14.63.

For the data analysis, 4 data sets were excluded because of two reasons:

- 2 participants passed the race without an opponent, with the consequence, that there was no real race condition. This would likely have effected the results, therefore these two participants were ignored for data analysis.
- Furthermore, a woman did not want to insult her son, therefore, she let him win the race. The two data records were also removed because that situation would have influenced the data as well.

5.1.3.1. Performance Measures (Behavioral Level)

To begin with, it was checked, if the task complexity was successfully diversified by the placement and number of obstacles, in other words, if the hard track was really more complex, than the easy one. For this purpose, a Mann-Whitney-U test on the number of collisions concerning the different task complexities was run. The test revealed, that the number of collisions was significantly higher for the more complex track, $z = 2.899$, $p = 0.004$. The mean rank for the hard task was 13.50, whereas for the easy one 6.00. Also the track solution time was significantly different for the two tracks, $z = 2.162$, $p = 0.031$. The average rank for the easy tasks was 7.00 and for the hard one 12.83. The absolute means concerning the performance can be seen in Table 5.2

Table 5.2.: Efficiency and Effectiveness in Dependency of the Track Difficulty

	Difficulty of Track			
	Easy		Hard	
	Mean	Standard Deviation	Mean	Standard Deviation
Number of Collisions	0.50	0.76	2.58	1.73
Track Solution Time in Seconds	54.50	15.22	81.75	27.75

Summarizing, the task completion time, as well as the number of collisions was higher for the hard task, as a consequence the manipulation of the task difficulty could be considered successful.

Regarding the two input modalities, the PC-remote and the gesture-based interface, the PC-remote outperformed the gesture-based interface:

- The mean track solution time, which indicated the efficiency, was 68 seconds when using the PC-remote and 73 seconds for the gesture-based interface.

- For effectiveness of the input modalities, the mean number of collisions was 2 for both devices, but in absolute values the gesture-based interface caused 19 collisions, whereas the PC-remote just raised 16 at all.

5.1.3.2. User Satisfaction Measures (Attitudinal Level)

Regarding the user satisfaction measures, the scales for (1) intuitiveness, (2) satisfaction and (3) the overall acceptance, gathered by the above explained questionnaire, were computed. In order to evaluate the internal reliability of the questionnaire for these 3 factors, a Cronbach's Alpha test was conducted, on the three scales. Internal reliability means hereby, if all items which measured intuitiveness for example, are related to each other.

The result of the Cronbach's Alpha can be between -∞ and 1, but just positive values can be meaningful interpreted. A common rule of the explanatory power can be seen in Table 5.3.

Table 5.3.: The Explanatory Power of the Cronbach's Alpha Value cf. [Wik13a]

α	Internal reliability
$\alpha >= 0.9$	Excellent
$0.8 <= \alpha <= 0.9$	Good
$0.7 <= \alpha <= 0.8$	Acceptable
$0.6 <= \alpha <= 0.7$	Questionable
$0.5 <= \alpha <= 0.6$	Poor
$\alpha <= 0.5$	Unacceptable

The scale intuitiveness, which consisted of 5 items in the questionnaire, scored a Cronbach's Alpha of 0.706 at first, which was acceptable according to the Table 5.3. In order to improve the Alpha value, one item of the intuitiveness scale ("The used device was hard to use") was deleted, with the result, that the alpha value reached 0.733.

For the satisfaction scale, which consisted of 4 questions, the internal reliability score was 0.635. After removing one item ("I was satisfied with my own performance"), the score was 0.646.

Regarding the last of the user satisfaction scales, acceptance consisted of 5 items, reached a value of 0.629. Also in this case, a deletion of an item brought an improvement to the score. The item "I would not be able to solve a task with the robot, with this input device, without help" was deleted, and as a consequence, the Cronbach's Alpha of acceptance was 0.741.

After the computation of the user satisfaction scales intuitiveness, satisfaction, and acceptance was done, the descriptive data revealed a trend, that the user satisfaction measures were strongly dependent on the task difficulty (cf. Table 5.4). When solving the hard course, the intuitiveness, satisfaction, and acceptance of the device was rated higher, although participants suffered more collisions, and they needed significantly more time to finish the track. At first, this seems paradox, but it also shows the strong coherency between input modalities and task complexities. One possible explanation of this tendency could be the fact, that people are more satisfied with a system, and also with themselves, when they succeed in more challenging tasks.

Table 5.4.: User Satisfaction Scales Regarding the Track Difficulty. (Higher Values Indicate Better Scales)

	Difficulty of Track			
	Easy		Hard	
	Mean	Standard Deviation	Mean	Standard Deviation
Intuitiveness	4.43	0.72	4.73	0.23
Satisfaction	3.96	0.84	4.24	0.67
Acceptance	4.50	0.88	4.58	0.53

For all the three scales (intuitiveness, satisfaction and acceptance) a Mann-Whitney-U test was conducted, also concerning for the used device (PC-remote and gesture-based interface), in order to identify, if one was perceived more intuitively for example. Unfortunately, there was no significance in the results. However, a trend was identified, that the PC-remote was perceived on the one hand to be more satisfying and accepted, but on the other hand less intuitive, in comparision to the gesture-based interface, which can be seen in Table 5.5.

Table 5.5.: The PC-remote Scored a Better Acceptance and Satisfaction, whereas the Gesture-Based Interface was Considered to be More Intuitive. (Higher Values Indicate Better Scales)

	Used Device			
	PC Control		Mobile Control	
	Mean	Standard Deviation	Mean	Standard Deviation
Intuitiveness	4.55	0.61	4.67	0.37
Satisfaction	4.28	0.56	3.97	0.88
Acceptance	4.72	0.34	4.38	0.87

Furthermore, the Mann-Whitney U test revealed interesting results for some of the single scale items of the questionnaire. As already pointed out in Table 5.5, the gesture-based interface was rated to be more intuitive by the participants of the study. However, for one item of the intuitiveness scale ("The used device was hard to use") it was the opposite. The PC-remote scored a higher score for this item. The result of the test was significant, $z = -2.195$, $p = 0.028$. The mean rank for the PC-control was 12.55 and the average rank for the gesture-based interface 7.17. A possible explanation would be, that using the gesture-based interface requires all in all more movement, therefore, it was perceived to be harder to use, than the PC-remote control. Anyway summarizing, the gesture-based interface reached a higher intuitiveness score.

In addition to that, a correlation between the used input device and one item of the satisfaction scale could be identified. The item "I was satisfied with my own performance" was rated significantly higher for the PC-remote, which reached a mean rank of 13.20; whereas the gesture-based interface just had an average rank of 7.80 ($z = -2.160$, $p = 0.031$). Obviously the kind of the used input device had direct consequences on how satified participants were with themselves. In that case they were more satisfied with the PC-remote in comparison to the gesture-based interface.

The last significant difference revealed by a Mann-Whitney-U test, was for one item of the satisfaction scale ("I would like to use the device often") with regard to the task complexity, $z = 2.438$, $p = 0.015$. The average rank for the hard track was 12.55 and for the easy one 6.50. This is another indication, that people are more satisfied when they manage to solve more complex tasks.

5.1.4. Summary of the Preliminary Study

All in all the PC-remote outperformed the gesture-based interface, especially concerning the performance measures efficiency and effectiveness, but the differences were much lower for the easy task. Also the differences in task completion time and number of collisions were statistically significant comparing the easy and the hard track. Which means, that in general for the hard task, differences between the two input modalities were much higher, in terms of user satisfaction measures as well as in performance results. On the other hand, the hard track was not perceived much harder than the easy one, which implicates, that real and perceived difficulty not always directly correlate.

It was also identified, that there is a strong relationship between the used input device and the perceived task complexity [cf. Figure 5.3]. There was no statistical prove, but a trend in the descriptive data supported this assumption. Participants who used the PC-remote control rated the easy track easier than people using the gesture-based interface. For the hard course it was the opposite, which means, that the gesture-based interface users thought, that the hard track was easier than for PC-remote users.

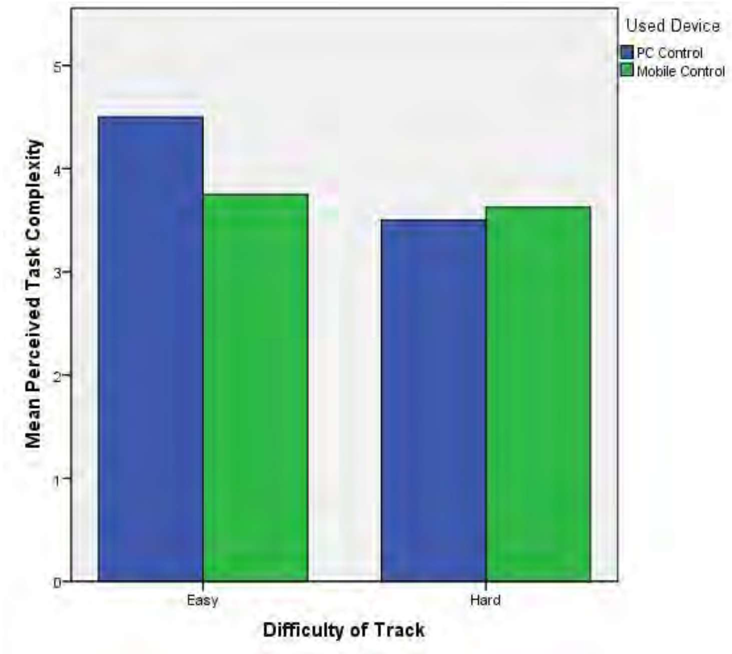

Figure 5.3.: Perceived Task Complexity in Relationship to the Difficulty of the Track and the Used Input Device

One explanation for this effect could be the fact, that especially people using the PC-remote suffered problems, when they controlled the robot back to the start. Some of the participants mentioned, that it was very challenging driving the robot back to the starting point, because they had to steer left, in order to let the robot go to the right, and vice versa. On the easy track, there was not such a strong need in turning, therefore this effect had a higher impact on the hard track.

Furthermore, there is another big advantage when using the gesture-based interface, five participants profited from. During the race they stood up in order to move along with the robot, which provided a better visibility of the race track. This behavior demonstrated the advantage of such an input modality for that type of task.

Summarizing, it could be shown, that different task complexities, as well as various input modalities have an impact on the performance and user satisfaction results. It is clear, that the results cannot be generalized for all contexts, because a race situation for example cannot be compared to a working context. Also the assumptions cannot be generalized for all variations of gesture-based interfaces or PC-remote controls, but just for the ones used in the study, but it can be supposed, that for most variations of them, the results would be similar.

The descriptive data, as well as the results of the reliability analysis and the nonparametric tests can be found in Appendix A.5.

5.2. Controlled Laboratory Study

In order to strenghten the assumptions already made, and to further investigate the interdependency of input modalities and task complexities, a follow-up study in a controlled laboratory setting was conducted. The preliminary study was focused on solving a task as fast as possible. Now an investigation in more complex tasks, which require a more exact controlling of the robots was done. Within the laboratory environment in the experience laboratory of the ICT&S center, the impact of different appearances of robots was inspected too. In the study, participants had to solve predefined transportation tasks by assistance of the robot with three different input modalities. They had to navigate the robot through a course with obstacles - which was not varied this time - to get the required parts, needed for assembling a Lego house.

[Results of the laboratory study have already been published successfully as a full paper [SWT13a] for the 2013 IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2013). The publication can be found in Appendix A.11.]

5.2.1. Methodology

The start was again a pilot study with an employee of the ICT&S Center which was necessary for detecting errors in the study setup. After refining the study-plan, a study was conducted which was set-up as a 3x3x2 mixed experimental design with the three conditions:

1. Three different input modalities, namely speech and the two modalities already used in the preliminary study (gesture-based interface, PC-remote control).

2. Three building tasks with different complexities (easy, medium, hard).

3. Two different appearances of the robot prototype (functional vs more human-like).

The input modalities were tested within-subject in the study, which means that all participants worked with all the input modalities. In order to avoid carry over effects (e.g., learning effects) [ER08], the sequence of the used input modalities was counterbalanced as shown in Figure 5.4. In other words, one participant began with the speech control, followed by the gesture-based interface and the PC-remote, another one started with the PC-remote follwed by the speech control and the gesture-based interface and so on. Each participant used each input modality once.

Gesture-Based	PC -Remote	Speech
Speech	Gesture-Based	PC –Remote
PC –Remote	Speech	Gesture-Based
Gesture-Based	Speech	PC –Remote
Speech	PC –Remote	Gesture-Based
PC -Remote	Gesture-Based	Speech

Figure 5.4.: Possible Sequences of the Used Input Modalities

In order to take the different levels of task complexities into account, findings of task complexity research were used, with regard to assembly tasks by Stork et al. [SSS08]. There they found some criteria which effect task complexity, especially for Lego assembly tasks. In their contribution “Optimizing Human-Machine Interaction in Manual Assembly” [SSS08], they were engaged with a concept for human-machine interaction, using a foot-pedal-based interface in a working environment. The main goal was to investigate “capabilities and bottlenecks in human information processing” and “the user controlled timing of relevant information presentation”. Therefore an experimental study took place where they varied difficulty of manual assembly tasks with a view to use respective processing resources of the participants. All in all, the data was interpreted for the general optimization of human-machine interaction, and to develop an “assistive system in production environments”. However, they pointed out, that there are several important factors for assistive systems in manual assembly considering the user’s state and task properties. Workers have to process visual information, store what parts are needed, and to find, grasp and mount the parts.

They mentioned, that people are not able to process an unlimited amount of information or perform large numbers of actions in parallel, therefore the tasks should be structured in a way that they can be done in a more linear way. (One step is followed by one another). It means that the robot should not disrupt the worker, therefore most activities of the robot should be controlled by the user in cases of timing, to create a kind of turn taking interaction. In their study they varied task complexity through:

- total amount of parts

- amount of different parts
- object classes to be built (frame, gear, rotor, roof and group)

They found out that the amount of different parts seemed to have the weakest influence on the results, whereas the object class had the highest impact. The most difficult one was the class roof, whereas the easiest one was the class frame. Group tasks were considered to be medium difficult.

According to these findings it was decided to let participants build a Lego house as in Figure 5.5 in the study, which could perfectly be divided into three subtasks which differed in complexity.

- Construction Task 1 (Easy): The frame of the house which was the easiest class according to [SSS08].
- Construction Task 2 (Medium): A group task, where participants had to add the prebuilt door to the frame.
- Construction Task 3 (Hard): To finish the Lego house, the roof was added which was the hardest type of assembly task follwing the results of Stork et al. [SSS08].

Figure 5.5.: Construction Task 1 to 3 and the Finished Lego house (From Left to the Right)

The factor task complexity was also studied as a within-subject design, because each participant had to complete the Lego house, therefore all subtasks had to be fulfilled.

The last of the three conditions was the appearance of the robot, which was studied as a between-subject design, which means, that the total number of participants was split between the two different appearances, in that case the purely functional robot, and the robot with the 3D printed head, to suggest anthropomorphism.

Before the study, participants had to sign a consent form, in order to allow us to use the data for scientific purpose. The data consent form can be found in Appendix A.6. After that, an overview about this master's thesis and the research topic was given. They were assured, that the collected data will only be used for scientific purposes, and that they could not do anything wrong during the study, because the system was studied, and not them. At the beginning, the tasks and the input modalities were explained and shown to the users. Furthermore, they got a guidance sheet, on which all the necessary commands for the PC-remote and the speech control system were depicted, in case they forgot some of the commands. On Figure 5.6 the guidance sheet for the speech command interface is depicted, for the PC-remote on Figure 5.7. For the gesture-based interface, participants got no guidance sheet.

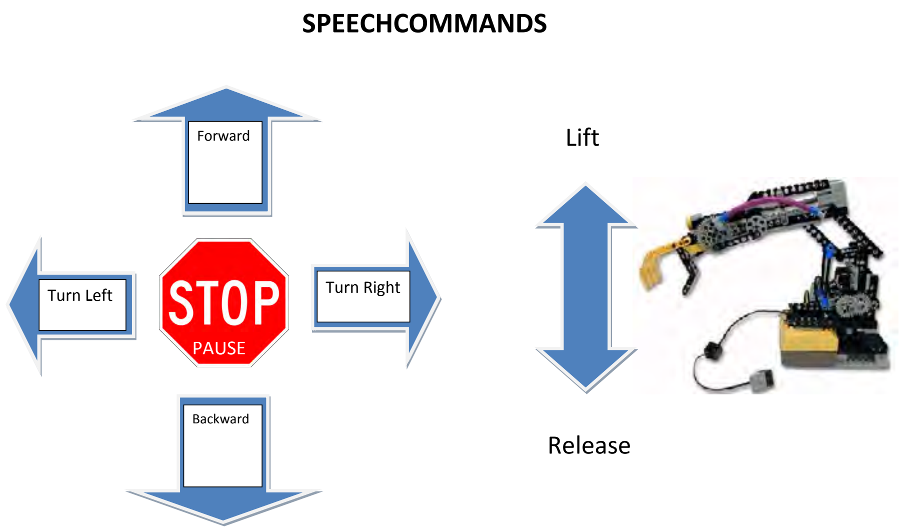

Figure 5.6.: Guidance Sheet for the Speech Command Interface

After that, participants had to navigate the robot through the track with one of the three input modalities and transport a box, which contained the required parts for the assembly tasks, from the end point of the track back to the start. The track can be seen in Figure 5.8

Shortly after getting the required parts, participants started with the first building task. Then they filled in all the surveys, except the questionnaire concerning the appearance of the robot, which only had to be filled in at the end of the study. The whole process was done three times. This means, that participants navigated the robot three times through the track, finished the Lego house, and filled in the surveys three times. At the very end they had to fill in the questionnaire concerning the appearance of the robot, and were asked one question about it, in order to investigate, if the appearance of the robot was suitable for these tasks.

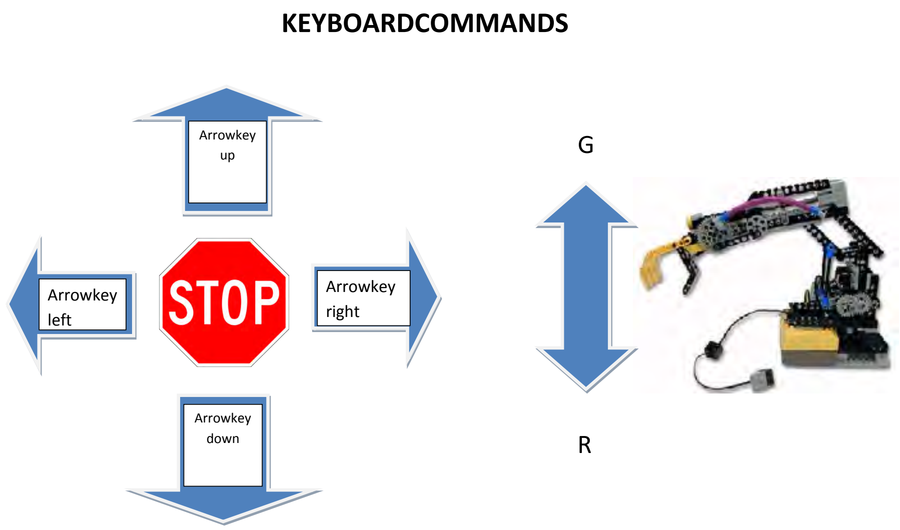

Figure 5.7.: Guidance Sheet for the PC-Remote Control

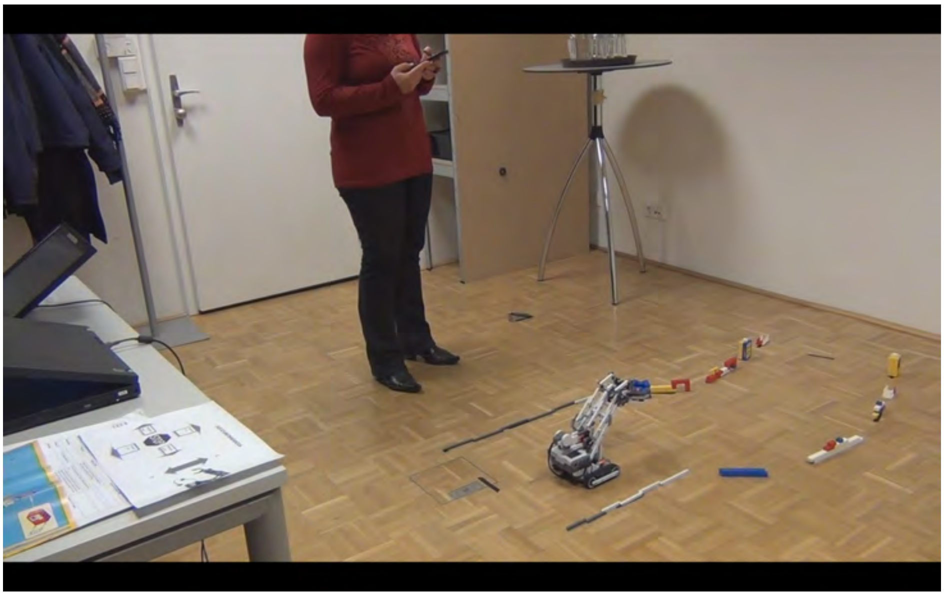

Figure 5.8.: The Track of the Laboratory Study

Moreover, the study was videotaped, in order to recheck the performance measures, and to identify problems with the input modalities. The studyplan can be found in Appendix A.7, but just in German, because the study was conducted in Austria.

5.2.2. Measures

Similiar to the preliminary study, the measures were divided into the same two categories.

5.2.2.1. Performance Measures (Behavioral Level)

The performance measures were nearly the same as in the preliminary study, with the only difference, that efficiency consisted of two variables, not only the time for navigating the robot through the track was measured, but also the solution time for the building tasks. Effectiveness remained the same, that means counting the numbers of collisions while controlling the robot.

5.2.2.2. User Satisfaction Measures (Attitudinal Level)

The same questionnaires as in the preliminary study were used in order to assess:

- Perceived task complexity concerning the difficulty of the track, and additionally regarding the complexity of the building task.
- Intuitiveness, satisfaction and acceptance with the input modality.

In addition to that trust, people had in the different input modalities, was inspected, specifically for different task complexities. Maybe trust is higher for more "classical and traditional" input modalities which are commonly known such as a PC-remote for example, compaired to more innovative ones such as a gesture-based device or a speech control interface. In order to assess trust, the guidelines of McKnight et al. [MCTC11] were used, who divided trust into three subcategories and explained it as follows:

- Functionality: "refers to whether one expects a technology to have the capacity or capability to complete a required task"
- Helpfulness: "excludes moral agency and volition (i.e., will) and refers to a feature of the technology itself - the help function, i.e., is it adequate and responsive?"
- Reliability: "suggests one expects a technology to work consistently and predictably."

Due to the fact that the used input modalities did not provide any help function at all, just questions about functionality and reliabilty were used in the studies. So the trust questionnaire which was used, consisted of seven items (4 regarding reliability and 3 concerning functionality) with a five-point Likert scale, as for the other user satisfaction measures. The complete questionnaire can be found in Appendix A.8.

Moreover, the cognitive workload of the participants after using each input modality was assessed. It was also thinkable, that for one input modality, for example, the performance is better than for another one, but on the other hand the cognitive load could be higher, which could be at the expense of user satisfaction, for example. In the work of Hart [Har06] the term workload is referred to as "the cost of accomplishing mission requirements for the human operator." In order to assess these costs or the cognitive workload, the Nasa-task load index or shortly NASA-TLX was created. The NASA-TLX "is a multi-dimensional scale designed to obtain workload estimates from one or more operators while they are performing a task or immediately afterwards." It consists of six different subscales, which represent in combination the "workload" which was experienced by the participants during a task:

- Mental demand: The level of cognitive activity required for the task.
- Physical demand: The level of physical activity required for the task.
- Temporal demand: The level of time pressure participants experienced during the task.
- Frustration: How stressed or insecure participants felt.
- Effort: How hard participants had to work to solve a task.
- Performance: How successful participants were.

The six subscales had to be rated between 1 which meant low, and 20 which indicated a high value on the questionnaire (cf. Figure 5.9).

Mit **Zeitdruck** ist die Geschwindigkeit gemeint, in der die Aufgabe ausgefüllt werden musste (langsam oder schnell, ausreichend Zeit zum Fertigwerden oder zuwenig Zeit).
Der **Zeitdruck** war:

Niedrig — Hoch

Figure 5.9.: The Item Temporal Demand of the German Version of the NASA-TLX

The NASA-TLX has often been used in several studies, and has also been translated to various languages. Due to the fact, that the participants in the studies were all Austrian or

German natives, the German version of the questionnaire was used. This was also used in the doctoral thesis of Wölber [Wï0]. Furthermore, the original NASA-TLX contains a weighing scheme, in order to take into account the subjective preferences concerning the different scales. One very common modification of the questionnaire is the NASA RAW TLX (NASA RTLX) which is simplier to apply, because of the elimination of the whole weighing process. Data analysis is done simply by averaging the ratings, in order to get an estimated workload. According to the work of Hart [Har06] who investigated and compared many studies which wheter or not applied the weighing scheme, the NASA RTLX was either found to be more sensitive, less sensitive, or equally sensitive. As a consequence, it was decided to use just the NASA RTLX in the studies, which can be found in the Appendix A.9.

As already stated, it can be assumed, that minimal human-like cues which are added to a robot, could suggest anthropomorphism, and furthermore have an impact on performance and user satisfaction. Possibly an anthropomorphic robot could result in a higher user satisfaction due to the more human-like appearance, whereas a purely functional robot could provide a better performance because of the lack of "unnecessary" parts which could distract users. To get an insight into the impact of different appearances of robots, the German version of the godspeed questionnaire concerning the anthropomorphism was applied. The questionnaire was implemented by Bartneck et al. [BCK09] and tested in various studies, in order to create a standartised measurement tool for researchers in the field of human-robot interaction. They provided a set of standardised godspeed questionnaires for many key concepts of human-robot interaction like perceived intelligence, animacy, perceived safety, anthropomorphism, and likeability. The authors provide their series of questionnaires in various languages freely available on the internet [Bar08].

The godspeed questionnaire series is based on semantic differential scales, and consists of five items for the factor anthropomorphism. Semantic differential means that pairs of oppositional words have to be rated, which can be seen in Figure 5.10.

The only survey, which the participants in the laboratory study had to fill in just once, at the very end of each session, was the godspeed questionnaire concerning anthropomorphism. The complete questionnaire can be found in Appendix A.10.

Furthermore, the study was videotaped, in order to recheck the measured time and number of collisions as well as to identify problems with the input devices.

	1	2	3	4	5	
Unecht	☐	☐	☐	☐	☐	Natürlich
Wie eine Maschine	☐	☐	☐	☐	☐	Wie ein Mensch
Hat kein Bewusstsein	☐	☐	☐	☐	☐	Hat ein Bewusstsein
Künstlich	☐	☐	☐	☐	☐	Realistisch
Bewegt sich steif	☐	☐	☐	☐	☐	Bewegt sich flüssig

Figure 5.10.: The German Version of the Godspeed Questionnaire Concerning Anthropomorphism

5.2.3. Results

All in all, there were 24 participants participating in the study. 11 of them were female, and 13 male.

The youngest participant was 15 years old, whereas the oldest one was 61 years old. The mean age was 29.46 yeas with a standard deviation of 12.19.

5.2.3.1. Performance Measures (Behavioral Level)

To begin with, it was checked, if the task complexity was successfully diversified by the different building tasks, in other words if the frame task was the easiest one, the group task medium, and the roof task the hardest one as intended.

For this purpose, a Kruskal-Wallis test on the time was conducted, participants needed to solve the different building tasks, which revealead a highly significant difference in the distribution between the three building tasks ($H(2) = 36.530$, $p = 0.000$), with a mean rank of 58.26 for the hardest task (roof), 32.22 for the medium one (group), and 22.22 for the class frame (easy).

As a consequence, it can be said, that the manipulation of the task complexity was successful. The study design for the building tasks followed the work of Stork et al. [SSS08], where they proposed, that roof tasks are the most complex Lego building tasks. This assumption could be supported by the fact, that in the laboratory study, the building solution time, which can be seen in Figure 5.11 was significantly higher than for the other tasks. Two participants did not even manage it to finish the roof task at all and gave up.

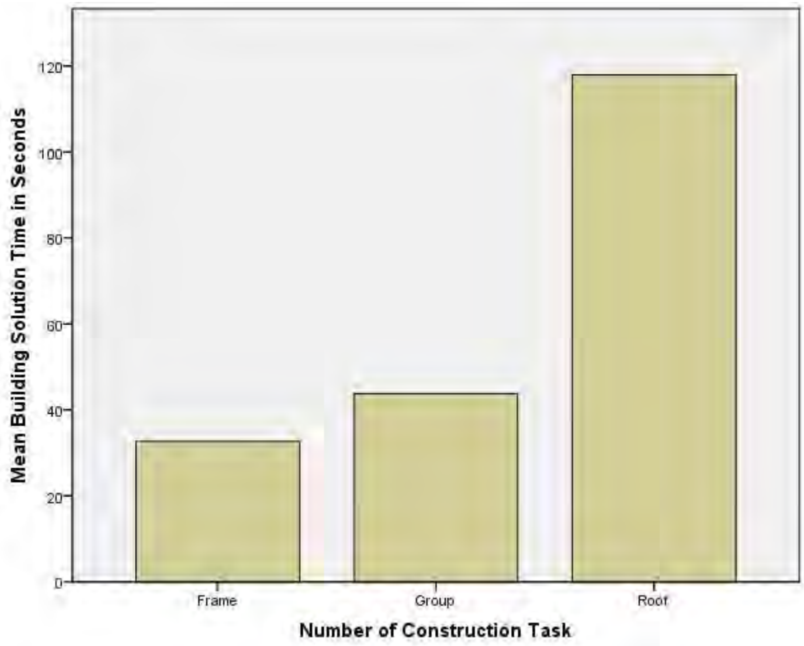

Figure 5.11.: The Mean Building Solution Time in Seconds for the Three Different Building Tasks Frame (Easy), Group (Medium) and Roof (Hard)

Furthermore, also the perceived building complexity matched the intended complexity for the different tasks (H(2) = 9.188, p = 0.010) with a significant result. The mean rank for the roof task was 30.88, for the medium group building task 36.78, and 46.34 for the frame class, which was the easiest one. In this case, a low value indicates a higher perceived complexity.

Summarizing, the difficulty of the building tasks was varied and perceived as intended, and the findings of Stork et al. [SSS08] concerning the three classes of building types used in this study could be reproduced.

Regarding performance of the different input modalities speech, the gesture-based interface, and the PC-remote control, the findings of the preliminary study could be reproduced too. The PC-remote outperformed the other two input modalities in terms of track solution time and also the number of collisions. The mean values can be seen in Figure 5.12. The speech remote was the poorest input modality concerning the performance measures.

This fact was also supported by a Kruskal-Wallis test (H(2) = 49.853, p = 0.000), which revealed highly significant differences in the distribution of the track solution time concerning the used device. The average rank for the speech interface was 58.20, followed by 37.56 for the gesture-based interface and 15.28 for the PC-remote control.

	Used Device					
	PC Control		Gesture-based		Speech	
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Track Solution Time in Seconds	55.20	9.10	79.96	19.19	145.65	61.58
Number of Collisions	0.16	0.37	0.24	0.72	2.10	1.29

Figure 5.12.: The Track Solution Time in Seconds as well as the Number of Collisions for the Three Different Input Modalities Ppeech, the Gesture-Based Interface and the PC-Remote Control

Also the differences in the distribution of the number of collisions was highly significant (H(2) = 39.204, p = 0.000) with a mean rank of 56.08 for the speech control, 27.22 for the gesture-based interface, and 27.32 for the PC-remote control.

5.2.3.2. User Satisfaction Measures (Attitudinal Level)

Regarding the user satisfaction measures, the scales were computed for:

1. intuitiveness, satisfaction, and acceptance
2. overall trust and its subcategories reliability and functionality
3. cognitive workload gathered by the NASA RTLX
4. and the level of anthropomorphism gathered by the godspeed questionnaire.

In order to again evaluate the internal reliability of the questionnaires for these factors, a Cronbach's Alpha test on all the scales was conducted.

The scale intuitiveness, which consisted of 5 items in the questionnaire, scored a Cronbach's Alpha of 0.792. This time a deletion of an item would have brought no improvement to the score.

For the satisfaction scale, which consisted of 4 questions, the internal reliability score was 0.912 which was an excellent score for internal reliablity according to the Table 5.3.

Regarding the acceptance scale, which consisted of 5 items, a score of 0.639 was achieved.

The complete trust scale with its 7 items resulted in a value of even 0.955 and its subcategory reliabilty consisting of 4 items resulted in a score of 0.948. The subcategory functionality with 3 items produced a Cronbach's Alpha of 0.904, which was also an excellent result.

For the cognitive workload, which was gathered by the NASA RTLX the value was 0.733 and for the appearance questionnaire 0.745 was calculated.

For all of these scales no deletion of items would have brought an improvement to the score, therefore all items were used for computing the different scales.

In order to further investigate the impact of different task complexities, another Kruskal-Wallis test on all the scales grouped by the different construction tasks was conducted. Unfortunately, there was no significant difference in the results besides of the already mentioned performance measures. However, interesting significances were revealed for single scale items.

For example, the distribution of the item "The used input device was fast to learn" was significant ($H(2) = 8.166$, $p = 0.017$), with a mean rank of 32.90 for class roof, 33.66 for class group, and 47.44 for the frame building task. Another significant difference could be found for a second item of the intuitiveness scale "The used control device needs much to learn" ($H(2) = 6.594$, $p = 0.037$) with the ranks 35.67 for class roof, 33.30 for class group, and 44.94 for the frame class, which was another indication for this assumption. These results could be explained by the fact that participants perceived the used input modalities to be faster and easier to learn, when they worked on an easier building task, which strongly shows the relationship between input modalities and task complexity, although the two actions had to be fulfilled independently within the study. In other words, participants had to control the robot first, and solve the building task after that.

Furthermore, also the distribution of the physical demand, which was one item of the cognitive workload scale was significant concerning the task complexity. The result of the NASA RTLX ($H(2) = 6.044$, $p = 0.49$) scored an average rank of 37.72 for the roof building task, 45.38 for the medium task (group), and 30.90 for the easiest one, which was the frame class. In addition to that, a second item of the cognitive workload scale - time pressure - was also significant ($H(2) = 5.993$, $p = 0.05$) with an average rank of 43.02 for the roof, 41.46 for group, and 29.52 for the frame of the Lego house. That means, the cognitive workload was always lower when building the frame part of the Lego house.

Combined, all of these facts imply that the manipulation of the building complexity was successful, and furthermore, that the findings of Stork et al. [SSS08]- which were used for the building tasks - could be reproduced.

In order to get an insight, which of the different input modalities (1) speech, (2) the gesture-based interface, or (3) the PC-remote control scored the best values in user satisfaction measures, a Kruskal-Wallis test grouped by the used device, was conducted. The test revealed highly significant results for many of the user satisfaction scales used in the study.

At first, a Kruskal-Wallis test regarding the perceived control complexity revealed a highly significant result ($H(2) = 23.060$, $p = 0.000$) with a mean rank of 47.52 for the PC-remote control, 40.50 for the gesture-based interface, and 21.76 for the speech interface. In other words, the PC-remote was considered to be easier to control, whereas the speech interface was considered to be harder to use, in comparision with the other two input modalities.

As in performance measures, the PC-remote outperformed the other two input modalities regarding the user satisfaction scales:

- The distribution of the scale intuitiveness which consisted of 5 questions ($H(2) = 25.238$, $p = 0.000$) revealed a mean rank of 52.74 for the PC-remote control, followed by 38.76 for the gesture-based interface, and 22.50 for speech, which made the PC-remote to be considered beeing the most intuitive device, which was a contradiction to the preliminary study, where the gesture-based interface was rated to be the most intuitive one. One possible explanation could be that for tasks which require speed the gesture-based interface was considered to be more intuitive, but for tasks which require a more exact control the PC-remote was perceived to be more intuitive. This again shows a strong interdependency between input modality and the type of task.

- The PC-remote was also rated to be the most satisfying input modality with a highly significant result ($H(2) = 31.947$, $p = 0.000$). 52.66 was the mean rank for the PC-remote control, 42.44 for the gesture-based interface, and speech was again considered to be the poorest device with an average rank of 18.90.

- Furthermore also the acceptance scale revealed a highly significant result ($H(2) = 16.467$, $p = 0.000$) with average ranks of 48.74 for the PC-remote, 40.64 for the gesture-based interface and 24.62 for the speech control.

- Concerning trust, the result was similar ($H(2) = 45.001$, $p = 0.000$) with 55.08 for the PC-remote, 43.78 for the gesture-based interface, and 15.14 for speech.

- The same was true for the subscales of trust reliability and functionality.

  - Reliabilty scale (H(2) = 47.850, p = 0.000); 55.90 for the PC-remote, 43.46 for the gesture-based interface, and 14.64 for the speech control.

  - Functionality scale (H(2) = 34.895, p = 0.000); 51.46 for the PC-remote, 44.66 for the gesture-based interface, and 17.88 for the speech control.

- Also the cognitive workload scale, where high values indicate a high workload offered a highly significant result (H(2) = 17.924, p = 0.000) with a mean rank of 26.36 for the PC-remote, 35.44 for the gesture-based interface and 52.10 for the speech control.

Summarizing, the PC-remote again outperformed the other two input modalites concerning user satisfaction scales followed by the gesture-based interface. This assumption was also supported by most of the single items. The complete data analysis results can be found in Appendix A.12.

Concerning the appearance of the robot, it was investigated, if a small human-like feature suggesting anthropomorphism - like the head used in the user study - could have an influence on the overall interaction. Interestingly, even this small change of the appearance of the functional robot influenced the results, although as expected, the robot was not considered much more humanlike than the purely functional one. Many of the participants mentioned, that the head had no functionality, and was therefore not necessary. Nevertheless, it's even the more astonishing that a Mann-Whitney U test showed, that for the single item "I would not be able to solve a task with the robot without help" more participants tended to disagree when using the robot with the 3D printed head (z = 2.054, p = 0.040), with an average rank of 42.04 for the robot with the head, and 33.62 for the purely functional one. In other words, participants who were in collaboration with the purely functional robot were less self-confident in beeing able to solve assignments on their own. Thus, it can be assumed, that even minimalistic human-like cues added to a robot, can increase participants' positive experience when collaborating with it. Three participants commented that the head made the robot more appealing and funny and one participant who interacted with the purely funtional prototype proposed to add eyes or similar cues to the robot in order to make it easier to forgive errors. One participant even mentioned that the head was necessary in order to identify the front and the back of the robot.

At the end of the study participants were asked, if they think, that the appearance of the robot was adequate for such kind of tasks.

- Nearly all persons concluded that the appearance of the robot was absolutely suiting for such kind of tasks and compared the robot to a digger or a pallet transporter.

- They mentioned, that at first sight the mechanism to drive and also the grabber was easily noticeable and as a consequence it was clear what can be done with the robot.
- On the other hand some participants thought that the head was not necessary and that the appearance of the robot did not matter at all but just the functionality.

In addition to the comparision of the different input modalities and task complexities, some differences concerning the gender of the participants could be identified. It was not the primary focus of the studies to identify gender differences, therefore, just a small overview of the most important results of the Mann-Whitney U test is given:

- The distribution of the single item "The used control device is easy comprehensible" revealed a significant result (z = 2.448, p = 0.014) with an average rank of 43.79 for female participants and 33.45 for male ones. In other words, women considered the different input modalities to be easier comprehensible, in contrast to male participants.
- For the item "I would not be able to solve a task with the robot without help" (z = -2.100, p = 0.036) men (average rank was 41.81) were more self-confident than women (average rank was 33.15).
- For the cognitive workload, where high values indicated a high workload, the test also revealed a significant result (z = 2.829, p = 0.005) with a mean rank of 46.03 for women and 31.69 for men. Obviously the cognitive workload was higher for female participants in this study.
- Also a difference in the perceived level of anthropomorphism was revealed (z = 2.212, p = 0.027). Female persons with a mean rank of 16.64 significantly perceived the robot in general to be more human-like than male persons, with an average rank of 10.14.

Due to the fact that the focus of this work was not on gender differences, these findings will not be further discussed.

In order to get a deeper insight into the interdependencies between input modality and task complexity, the results grouped by task complexity and input modality were further analyzed. Many interdependencies could be identified within the results.

Participants, especially after using the speech control system, which caused the highest mental workload - according to the questionnaire - perceived the building tasks more complex, although the building assignment was independent from the input modality. As already stated in the study setup, people had always to control the robot first, in order to get the required parts, and then to assembly them. In Figure 5.13, it can be seen, that especially for the speech control system, the building complexity was perceived harder. In addition to that, the difference in the perception for the easy and the medium building task (frame and group) was lower than for the roof task, which was the most complex assembly task in the study. It is obvious, that there is a strong inderdependency between the used input modality and task complexity, which had a high impact on the results.

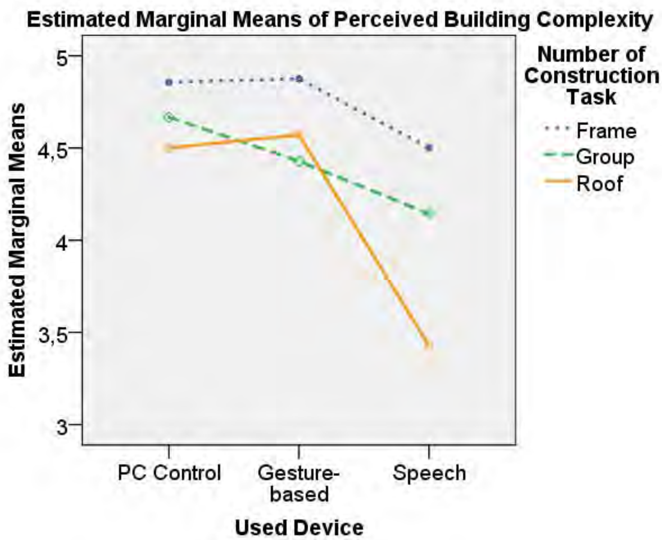

Figure 5.13.: The Buildings Tasks were Perceived More Complex, when the More Complex Input Modalities were used Before (5 = Easy, 1 = Hard)

Also for the item “The control device was needlessly complex” the result was nearly the same, for the frame assembly task, which was the easiest one, but differed much more for the classes group and roof, which again emphasizes the strong relationship between the two factors input modality and task complexity. In Figure 5.14, it can be seen, that for the easiest building task (frame), there is nearly no difference concerning the item “the used control device was needlessly complex” between the different input modalities, but for the harder tasks roof and frame the difference is huge.

In general, it can be said, for all the used user satisfaction measures and also for the cognitive workload results, that especially when the most complex input modality (in this case speech control) and the more complex assembly tasks (group and roof class) converged, the user satisfaction measures were rated much lower. One example regarding the intuitiveness scale

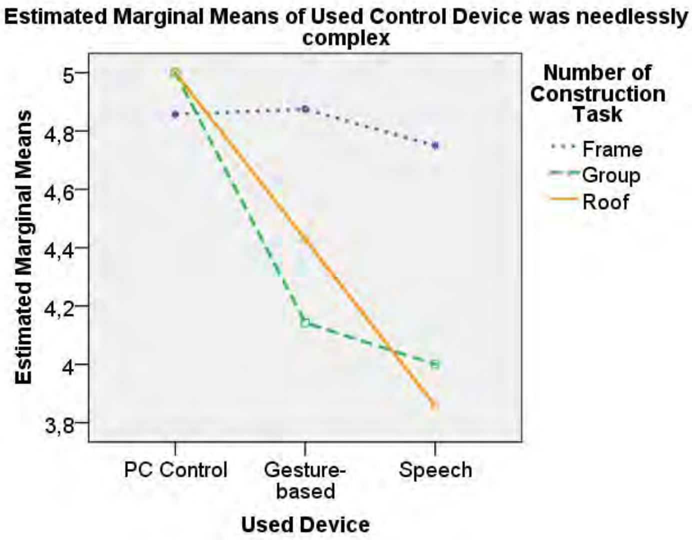

Figure 5.14.: The Perceived Control Complexity in Dependency of the Different Assembly Tasks (5 = Totally disagree, 1 = Totally agree)

can be seen in Figure 5.15. This trend could be identified for all of our user satisfaction measures, as well as for the cognitive workload.

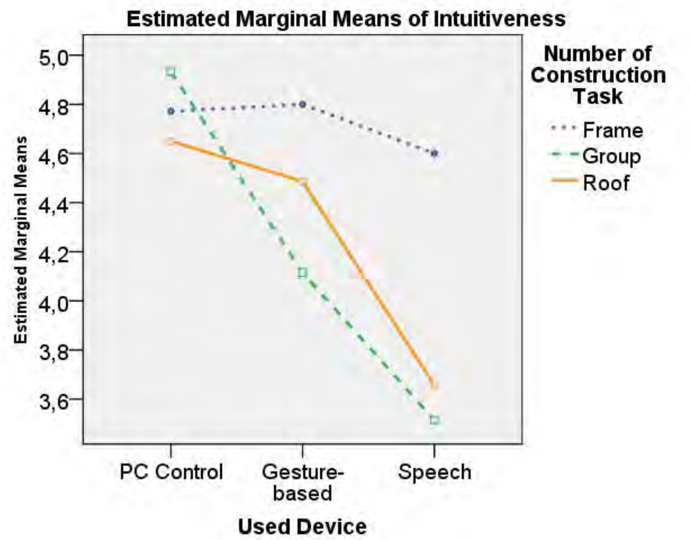

Figure 5.15.: The Intuitiveness Scale was Rated Much Lower when Complex Modality and Complex Assembly Tasks Converged (5 = Very high, 1 = Very low).

5.2.4. Summary of the Laboratory Study

Summarizing, similar to the preliminary study, the PC-remote again outperformed the other two input modalities in terms of performance, as well as in the user satisfaction measures. However, the differences for simple task were much lower than for hard tasks, matching again the findings of the preliminary study.

The manipulation of the task complexity worked out perfectly, even for both the perceived as well as for the real task complexity. The findings of Stork et al. [SSS08] could be reproduced. The frame task was considered to be the easiest one, and the roof task to be the most difficult assignment. Some participants even gave up building the roof of the Lego house.

The study also showed, that it is possible to enhance human-robot collaboration by just adding minimal human-like cues to the robot. The head which was added in the second user study, made participants significantly more self-confident in solving tasks on their own. An effect, which should not be underestimated because if such minimal cues improve the interaction in terms of user satisfaction, this is a simple way to improve the interaction as such for functional robots.

All in all, it could be shown, that different task complexities, as well as various input modalities have a severe impact on the results. It is clear, that the results cannot be generalized for all variations of gesture-based interfaces, PC-remote controls or speech control systems, but just for the ones used in the study. But again this trend should be similar for other variations for these types of input modalities.

In general, the study showed that especially when the most complex input modality (in that case speech control) and the more complex assembly tasks (group and roof class) converged, the user satisfaction measures were rated much lower. This trend could be identified for all of the user satisfaction measures, as well as for the cognitive workload. This points out the need of a careful selection of input modalities, especially for hard tasks in order to avoid to overburden users, and as a consequence handicap a performant and satisfying human-robot collaboration. The descriptive data, as well as the results of the reliability analysis and the nonparametric tests can be found in Appendix A.12.

This chapter dealt with the two user studies (one in a public context, and one in a controlled laboratory setup) which were conducted in order to achieve the research goals and to answer the research questions. In the following part, the research questions are answered an discussed.

6. Discussion

From the technical point of view, all the three input modalities which were implemented to control the robot worked out sufficiently, so that naive users could perform all tasks successfully. Similarly the Mindstorm prototype suited for the tasks in the user studies. Therefore the prototyping and implementation phase can be considered to be quite successful, although some problems appeared while implementing the different modalities. A detailed description of the prototyping phase can be found in Chapter 4.

In most cases, when humans and robot work together shoulder to shoulder, there is only one possibility to control a robot, but many different tasks to solve in collaboration. Due to the strong interdependency of input modality and task complexity, which could be identified in the studies, the design process for planning human-robot collaborative systems should specifically include the tasks which have to be fulfilled.

In order to answer research sub-question 1 (What interdependencies between input modalities and task complexities could be identified?) it could be seen, that it is not really important which input modality is provided and used for simple tasks. The differences in user satisfaction rankings and performance measures were small and not statistically significant. To put it another way, users were always satisfied when solving easy tasks, and performed them equally well, no matter which input modality they used. However, for hard assignments, the differences in performance and user satisfaction were larger and need to be taken into consideration because of their statistical significance.

Concerning research sub-question 2 (How do users perceive the different input modalities in terms of user satisfaction measures?) it could be shown, that users perceived all the different input modalities equally well, but just when solving easy tasks. For hard tasks the differences in user satisfaction measures were statistically significant. The rating of the PC-remote was best in both studies except in the preliminary study, where the gesture-based interface was rated to be more intuitive. Nevertheless, the PC-remote outperformed the other two input modalities followed by the gesture-based interface.

Generally, the speech control interface was considered to be the poorest device in all categories of the study, regarding research sub-question 2. One possible explanation for this fact could be the latency, which participants had to take into account. Latency is often a problem for speech control systems, which means that if a participant wanted to give a command to the robot, it took a few milliseconds until the robot actually reacted. In other words, if a participant wanted the robot to stop, it went a few centimeter further until it actually stopped. According to some of the participants, this was a huge disadvantage of this kind of modality, and made it very complex to control the robot. Furthermore, due to the used speech recognition system [Sph13] and the corresponding dictionary, the commands for controlling the robot were implemented in English, which was another problem for some participants, who were all Austrians. One participant even rejected to try the speech control, because of the lack of English language skills. These two facts had probably made a contribution for the speech control system, to be the poorest device in this study. On the other hand, the accuracy of the speech control was satisfying for most of the participants in the study, and many of them mentioned after the test run that the disadvantage of the latency could easily be compensated with a little experience using this kind of modality. Besides, they suggested, that especially for very simple tasks, like going forward or backward, the speech control system was absolutely suitable. In addition to that, the speech control was the only input modality, which provided the possibility to use both hands for a secondary or even a primary task while collaborating with the robot. Furthermore it was possible to move along with the robot, which provided a better overview over the situation, and could especially be useful in situations, where the robot could get out of sight. This is indeed a huge advantage compared to more static input devices like the PC-remote, for example.

The third input modality, which is investigated in research sub-question 2 was the gesture-based interface. It did not achieve the best values in terms of performance, nor in the user satisfaction scales, but many of the participants mentioned that it was a pleasant experience to use it, which can also be considered as an advantage of this modality. This fact could provide a better long-term motivation and satisfaction for users, which possibly leads to a higher internal motivation, what makes people more receptive to the information according to Brown [Bro88]. As a consequence, the learning process for using such a kind of input modality will be faster and with a higher output. Furthermore, also the advantage to be on the move could be capitalised, but in contrary to the speech control system, the user needs one hand for controlling the gesture-based interface. As a consequence just one hand is available for other tasks, which is anyway one more than when using the PC-remote.

Research sub-questions 3 (Are the means of the interaction provided by the different input modalities effective for the human and the robot?), 4 (Are the means of the interaction provided by the different input modalities efficient for the human and the robot?) and 5 (How do the different input modalities perform for different task complexities?) dealt with the performance of the different input modalities. Again for easy tasks the difference was minimal, but for challenging tasks the PC-remote performed best in terms of efficiency and effectiveness. Probably because it was considered the most accurate, reliable, and familiar input modality. In challenging situations users probably prefer to use control possibilites which are familiar, compared to ones, which are more innovative or experimental. Concerning performance, the speech control was the poorest device for hard tasks too.

Regarding one of the main questions if minimal human-like cues added to a purely functional robot in order to suggest anthromorphism have an effect on the interaction/collaboration, it can be said, that even with this small change of the functional robot by adding the 3D printed head, a positive effect could be identified. For one item of the acceptance scale a significant difference could be detected: People collaborating with an anthropomorphised robot with head were more self-confident in solving tasks on their own than participants who interacted with the purely functional one. If even this small human-like cue could evoke a positive effect on the overall interaction, this offers many quick and often cheap possibilities to enhance Human-Robot Collaboration.

Concerning the main research goal of finding the best mixture of input modalities and different levels of task complexities in terms of performance and also user satisfaction measures, it can be stated that for easy and simple tasks it does not matter which input modality is provided. However, for challenging tasks it cannot be said, for example, that a PC-remote is always the best choice. It depends on so many other factors and circumstances. Like already mentioned one factor is if the user has to be on the move or can be static during the interaction with the robot. Furthermore, does the user need one or two hands free for a secondary task during the interaction? What is the type of task or the surrounding? A speech control system, for example, is not really suitable in a very noisy context and a gesture-based interface possibly useless in a workingplace with very little space.

All in all, the two user studies proved the assumption about interdependencies of input modality and task complexity in Human-Robot Collaboration. Pros and cons of the different input modalities were illustrated, especially concerning different complexities. Furthermore, the outcome indicates, that none of these elements should be inspected separately. Only a combined investigation of all of these factors, such as input modalities, task complexity and appearance of the robot (and probably additional factors) allows well-founded implications for Human-Robot interaction. This is a fact, any researcher who is working in this area should be aware of. It is clear, that this work cannot be generalized to all variations of speech, gesture, and point-and-click input, only assumptions for the input modalities used in this studies can be made. However, these studies are just a starting point for a series of controlled experiments, to further decode this interdependency and propose adaptive multimodal Human-Robot Collaboration scenarios. Next steps are illustrated in the following future work section.

7. Future Work

Due to the fact, that in this work the interaction was only studied with naive users, the aim - in a next step- is to investigate the possibility to reproduce the findings if participants had a longer training phase and got used to all input modalities. Therefore, it is planned to give participants the robot and the input modalities for usage and training in their private home for one week, before conducting a controlled user study in the lab again.

It can be assumed, as an initial tendency was found in the user studies, that a humanoid design of robots generates a more positive feeling in the user. Therefore, further investigation is needed of how to produce minimal cues on the robot, to enhance the overall cooperation from the user's point of view. Moreover, a deeper insight into the impact of different appearances is needed; The human-like cue has a positive effect on the interaction, although the robot is not considered to be more human-like, therefore, a more human-like robot (e.g., the Nao or the Darwin robot) should be compared to the functional Lego Mindstorm prototype. Similarly, due to the fact that no subject worked with both types of robots, in a next study, a within-subject approach for further investigation is considered.

In order to further investigate in the main research goal of finding the best mixture of input modality, task complexity, and appearance of robot, the concept of an "intelligent" multimodal interface should be explored, where people can freely choose which input modality they need in specific situations. In comparison to most typical multimodal interfaces, which provide many different input modalities at the same time, is probably not the ideal solution for successful human-robot collaboration. Users should be able to always concentrate on their primary task and not think about which input possibility would be suiting for different kind of tasks and complexities. In that case "intelligent" means, that the interface reacts also in accordance to the context and other circumstances like light, noise, and temperature and provides a proper input modality and an according feedback modality (e.g., visual, haptic, and auditive). The resulting combinations of input modality and feedback mechanism for different task complexities should enable context-specific human-robot cooperation.

Therefore, the overall goal is to further explore the most appropriate input modality for a set of tasks, categorized according to their level of complexity. Additional to complexity, other factors have to be taken into account, such as whether or not the robot is physically collocated with the user, ambient noise, light conditions, and other factors, such as, if the robot could get out of sight during the interaction or the user needs one or both hands free for another task, like the necessity of carrying anything, for example. Moreover, it has to be clear, if the user needs to have the possibility of mobility during the interaction. This classification will enable the possibility to provide the ideal complementing input modality for each type of task.

I strongly believe that a more adaptive and “intelligent“ multimodality is needed, which could only be achieved if the tasks have been well classified before. This kind of multimodality concerning input modalities could be the key to context-specific human-robot collaboration and could enhance the relationship between human and robot itself.

List of Figures

Bibliography

[ADT13] Adt plugin for eclipse. http://developer.android.com/tools/sdk/eclipse-adt.html, 2013. [Online; accessed 25-March-2013].

[AN04] T. Ayres and B. Nolan. Voice activated command and control with java-enabled speech recognition over wifi. In *Proceedings of the 3rd international symposium on Principles and practice of programming in Java*, PPPJ '04, pages 114–119. Trinity College Dublin, 2004.

[Ana13] T. Anandan. The end of separation: Man and robot as collaborative coworkers on the factory floor. *Robotics Online*, 2013.

[And13] Android api for developers. http://developer.android.com/about/versions/android-4.1.html, 2013. [Online; accessed 25-March-2013].

[Arc05] Archytos of tarentum. http://www-history.mcs.st-andrews.ac.uk/Biographies/Archytas.html, 2005. [Online; accessed 11-November-2013].

[Asi42] I. Asimov. Runaround. Street and Smith, March 1942.

[Bar08] C. Bartneck. The godspeed questionnaire series. http://www.bartneck.de/2008/03/11/the-godspeed-questionnaire-series/, 2008. [Online; accessed 9-June-2013].

[BCK09] C. Bartneck, E. Croft, and D. Kulic. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, pages 71–81, 2009.

[Bed64] S. A. Bedini. The role of automata in the history of technology. *Technology and Culture*, pages 24–42, 1964.

[BGR+09] A. Bannat, J. Gast, T. Rehrl, W. Rösel, G. Rigoll, and F. Wallhoff. A multimodal human-robot-interaction scenario: Working together with an industrial robot. In *Proceedings of the 13th International Conference on Human-Computer Interaction. Part II: Novel Interaction Methods and Techniques*, pages 303–311, Berlin, Heidelberg, 2009. Springer-Verlag.

[Blu13] Bluecove jsr-82 project. `http://bluecove.org/`, 2013. [Online; accessed 28-March-2013].

[Bro88] Ann L. Brown. Motivation to learn and understand: On taking charge of one's own learning. *Cognition and Instruction*, pages 311–321, 1988.

[CSSW10] R. Cantrell, M. Scheutz, P. Schermerhorn, and X. Wu. Robust spoken instruction understanding for hri. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, HRI '10, pages 275–282, Piscataway, NJ, USA, 2010. IEEE Press.

[Cup13] Official robocup website. `http://www.robocup.org/about-robocup/objective/`, 2013. [Online; accessed 05-November-2013].

[Dav06] The da vinci robot. `http://www.ncbi.nlm.nih.gov/pubmed/17206888`, 2006. [Online; accessed 11-November-2013].

[DFAB98] A. Dix, J. Finley, G. Abowd, and R. Beale. *Human-computer interaction (2nd ed.)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1998.

[Ecl13] The eclipse juno project. `http://www.eclipse.org/juno/projects.php`, 2013. [Online; accessed 25-March-2013].

[ER08] Experiment-Resources.com. Counterbalanced measures design, 2008. [Online; accessed 21-Mai-2013].

[GS08] C. Guo and E. Sharlin. Exploring the use of tangible user interfaces for human-robot interaction: a comparative study. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pages 121–130, New York, NY, USA, 2008. ACM.

[GTON09] V. Groom, L. Takayama, P. Ochi, and C. Nass. I am my robot: the impact of robot-building and robot form on operators. In *Proceedings of the 4th ACM/IEEE* *international conference on Human robot interaction*, HRI '09, pages 31–36, New York, NY, USA, 2009. ACM.

[HÏ3] G. Hölzl. Lego-mindstorms-minddroid. https://github.com/NXT/LEGO-MINDSTORMS-MINDdroid, 2013. [Online; accessed 24-March-2013].

[Har06] S. G. Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 50, pages 904–908. Sage Publications, 2006.

[HB04] G. Hoffman and C. Breazeal. Collaboration in human-robot teams. In *In Proceedings of the AIAA 1st Intelligent Systems Technical Conference*, 2004.

[Hea11] M. A. Hearst. 'natural' search user interfaces. *Communications ACM*, 54(11):60–67, November 2011.

[Her99] Heron of alexandria. http://www-history.mcs.st-and.ac.uk/Biographies/Heron.html, 1999. [Online; accessed 11-November-2013].

[HHA10] S. T. Hayes, E. R. Hooten, and J. A. Adams. Multi-touch interaction for tasking robots. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, HRI '10, pages 97–98, New York, NY, USA, 2010. ACM.

[Ico13] Icommand api. http://lejos.sourceforge.net/p_technolog\ies/nxt/icommand/api/index.html, 2013. [Online; accessed 28-March-2013].

[ISO99] Ergonomische anforderungen für bürotätigkeiten mit bildschirmgeräten teil 11: Anforderungen an die gebrauchstauglichkeit - leitsätze, Januar 1999.

[JSG13] Jspeech grammar format. http://www.w3.org/TR/jsgf/, 2013. [Online; accessed 28-March-2013].

[Kat74] I. Kato. Information-power machine with senses and limbs (wabot 1). In *First CISM-IFToMM Symp. on Theory and Practice of Robots and Manipulators*, volume 1, pages 11–24. Springer-Verlag, 1974.

[KUK73] The first kuka robot. http://www.kuka-robotics.com/austria/de/company/group/milestones/1973.htm, 1973. [Online; accessed 26-March-2013].

[Leg13] Lego mindstorms official website. http://mindstorms.lego.com/en/history/default.aspx, 2013. [Online; accessed 16-March-2013].

[LSh13] Official lego shop website. http://mindstorms.lego.com/en-us/products/9845.aspx, 2013. [Online; accessed 26-March-2013].

[Mat07] M. J. Matarić. *The robotics primer.* MIT Press, 2007.

[MBF+09] T. Mirlacher, R. Buchner, F. Förster, A. Weiss, and M. Tscheligi. Ambient rabbits likeability of embodied ambient displays. In M. Tscheligi, B. Ruyter, P. Markopoulus, R. Wichert, T. Mirlacher, A. Meschterjakov, and W. Reitberger, editors, *Ambient Intelligence*, volume 5859 of *Lecture Notes in Computer Science*, pages 164–173. Springer Berlin Heidelberg, 2009.

[MCTC11] D. H. Mcknight, M. Carter, J. B. Thatcher, and P. F. Clay. Trust in a specific technology: An investigation of its components and measures. *ACM Trans. Manage. Information Systems*, pages 12:1–12:25, July 2011.

[Nie01] T. Nielsen. Usability metrics, 2001. [Online; accessed 21-Mai-2013].

[Ote13] Official website of otelo. http://www.otelo.or.at/, 2013. [Online; accessed 23-March-2013].

[PHRA02] N.S. Pollard, J.K. Hodgins, M.J. Riley, and C.G. Atkeson. Adapting human motion for the control of a humanoid robot. In *Robotics and Automation, 2002. Proceedings. ICRA '02. IEEE International Conference on Robotics and Automation*, volume 2, pages 1390 – 1397, 2002.

[RBO09] P. Rouanet, J. Bechu, and P.-Y. Oudeyer. A comparison of three interfaces using handheld devices to intuitively drive and show objects to a social robot: the impact of underlying metaphors. In *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on Robot and Human Interactive Communication*, pages 1066 –1072, 2009.

[RDO11] P. Rouanet, F. Danieau, and P-Y. Oudeyer. A robotic game to evaluate interfaces used to show and teach visual objects to a robot in real world condition. In *Proceedings of the 6th international conference on Human-robot interaction*, HRI '11, pages 313–320, New York, NY, USA, 2011. ACM.

[Soj97] Sojourner. http://spacepioneers.msu.edu/robot_rovers/sojourner.html, 1997. [Online; accessed 26-March-2013].

[Sph13] Cmu sphinx official website. http://cmusphinx.sourceforge.net/, 2013. [Online; accessed 27-March-2013].

[SSS08] S. Stork, C. Stössel, and A. Schubö. Optimizing human-machine interaction in manual assembly. In *17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN, 08)*, pages 113 –118, August 2008.

[SWT12] G. Stollnberger, A. Weiss, and M. Tscheligi. The effect of input modalities and different levels of task complexity on feedback perception in a human-robot collaboration task. In *Workshop: Feedback in HRI at RO-MAN 2012*, 2012.

[SWT13a] G. Stollnberger, A. Weiss, and M. Tscheligi. The harder it gets: Exploring the interdependency of input modalities and task complexity in human-robot collaboration. In *RO-MAN 2013: 22th IEEE International Symposium on Robot and Human Interactive Communication*, 2013.

[SWT13b] G. Stollnberger, A. Weiss, and M. Tscheligi. Input modality and task complexity: Do they relate? In *HRI 13: Proceedings of the 8th ACM/IEEE International Conference on Human Robot Interaction*, 2013.

[Val10] L. Valk. *The LEGO MINDSTORMS NXT 2.0 Discovery Book: A Beginner's Guide to Building and Programming Robots.* No Starch Press, San Francisco, CA, USA, 1st edition, 2010.

[Ẅ10] J. P. Wölber. *Evaluation zweier computergestützter Lernprogramme in der Parodontologie.* dissertation, Albert-Ludwigs-Universität, 2010.

[WBS+09] A. Weiss, R. Bernhaupt, D. Schwaiger, M. Altmaninger, R. Buchner, and M. Tscheligi. User experience evaluation with a wizard of oz approach: Technical and methodological considerations. In *Humanoids 2009. 9th IEEE-RAS International Conference on Humanoid Robots*, pages 303 –308, dec 2009.

[Wei10] A. Weiss. Validation of an evaluation framework for human-robot interaction. the impact of usability, social acceptance, user experience, and societal impact on collaboration with humanoid robots. *Unpublished doctoral dissertation, University of Salzburg*, 2010.

[Wik13a] Wikipedia. Cronbach's alpha — Wikipedia, the free encyclopedia, 2013. [Online; accessed 26-Mai-2013].

[Wik13b] Wikipedia. Lego mindstorms nxt 2.0 Wikipedia, the free encyclopedia. `http://en.wikipedia.org/wiki/Lego_Mindstorms_NXT_2.0`, 2013. [Online; accessed 16-March-2013].

[Wik13c] Wikipedia. Robot hall of fame Wikipedia, the free encyclopedia. `http://en.wikipedia.org/wiki/Robot_Hall_of_Fame`, 2013. [Online; accessed 18-March-2013].

[WLK+04] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel. Sphinx-4: a flexible open source framework for speech recognition. Technical report, Mountain View, CA, USA, 2004.

[YVR11] D. R. Yates, C. Vaessen, and M. Rouppret. From leonardo to da vinci: the history of robot-assisted surgery in urology. *BJU International*, pages 1708–1713, 2011.

A. Appendix

A.1. Workshop Position Paper accepted for the 2012 IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)

G. Stollnberger, A. Weiss, M. Tscheligi

The effect of input modalities and different levels of task complexity on feedback perception in a human-robot collaboration task.

Abstract— **In the research field of Human-Robot Collaboration choosing the right input modality and the according feedback modality are crucial aspects for successful cooperation for different levels of task complexity. In this position paper we present a preliminary study we conducted in order to investigate the correlations between input modalities, feedback and task complexity. We assume that specific input devices are suitable for specific levels of task complexity and that these input modalities differ in the ideal complementing feedback modality. Moreover, we assume that through the ideal mix, user satisfaction and overall experience of the human-robot cooperation can be enhanced. We conducted a study in which participants could choose between two different input modalities to drive a race with a Lego Mindstorm robot against the other participant. One of the main findings was that all of these factors have a severe impact on the results but the effect of task complexities and input devices was more significant than the different feedback mechanisms which on the other hand strongly correlate with the used input device. Therefore none of these factors should be inspected separately. Furthermore, we found out that user's perceptions of their performance in some cases differed from reality.**

I. Introduction

Human-Robot-Collaboration becomes more and more important in different contexts, such as in modern homes or hospitals as assistive systems for example and in the factory context as assembly, painting, inspection or transportation robots.

One important aspect in the exploration of human-robot cooperation is to find out which input modalities should be used for cooperation. Especially concerning different levels of task complexity and the according feedback, which needs to be provided. Therefore, we wanted to investigate the impact and correlation of input devices with different levels of task complexity and different feedback modalities. Based on literature review we built two robot Mindstorm robots, which can be controlled with two different input devices (a gesture-based input system on a mobile phone and a PC-remote control interface) and studied these input modalities in a field trail with 24 participants.

In the following paper we want to provide a short summary of related work, after that our research aim and the experimental setup is described followed by the results of the study which include a discussion part and finally a few sentences about future work are given.

II. Related Work

One relevant aspect for successful human-robot cooperation is the input modality, which is used to control and interact with the robot. Clearly different input modalities suit better with specific feedback modalities, e.g. if we talk to a robot we assume that the robot can talk back. Speech is still considered the most natural input modality for human-robot cooperation, but it is still a problem for robots to process spoken commands with high accuracy.

In addition to that there are other input modalities to interact with robots than speech such as gesture, keyboards etc. Rouanet and colleagues for example compared a keyboard-like and a gesture-based interface to communicate and control the system driving through two courses which differed in complexity and results showed that different task complexities have a significant impact on the user performance. [1]

Further studies with a Wiimote, Laser pointer, Iphone and a simulated gesture-based interface were compared as input modalities in HRI, which also showed the importance of according feedback modalities. [2]

Bannat et al. proposed that a more natural and flexible human-robot interaction can be achieved by providing different control modalities at the same time namely speech, gaze and so-called soft buttons and that users can choose which modality suits best in a specific situation. However, this concept is not experimentally validated so far. [3]

Another interesting approach are multi-touch interfaces for commanding robots. According to Hayes et al. it is not satisfying commanding robots with static input devices, such as a mouse or a keyboard, above all in situations where the user is on the move or needs the hands free for a secondary or even for the primary task. They could show that multi-touch interfaces provide a good usability and a lower overall workload for users controlling the robot. [4]

Furthermore, experiments with a tangible user interface for human-robot interaction with Sony's AIBO had been done using also a keypad and a gesture-based interface. The researchers believed that a more natural way of interaction could be achieved if users can focus their attention on a more high-level task planning in comparison to low-level steps. Their tangible user interface outperformed the keyboard interface, but they mentioned that this could possibly be because they did not use the most intuitive mapping of keys on the keyboard. [5]

Another approach was investigated where pre-recorded human motion and trajectory tracking as an input modality was used for anthropomorphic robots with the aim to compare motions of the participants and the robots. They mentioned that their results were limited because they only measured the degree of freedom of each joints separately therefore they have to conduct a second study. The process of recording human motions seems to be very complex and time consuming therefore this possibility will be ignored for our research. [6]

G. Stollnberger, A. Weiss, and M. Tscheligi are with the HCI and Usability Unit, ICT&S Center, University of Salzburg,
5020 Salzburg, Austria (firstname.lastname@sbg.ac.at)

To summarize, a lot of research has already been done to compare input modalities for Human-Robot collaboration in order to identify their advantages and disadvantages in terms of performance and user experience/ usability. In some cases the input modalities were compared according to their task-suitability and their impact on feedback modalities. However, what is up to our knowledge missing so far is a structured investigation of the interplay between: (1) input modality, (2) task complexity, and (3) feedback modality. In order to investigate this interplay we conducted a preliminary small user study in which we compared two input modalities (gesture and keyboard) with two different levels of task complexity [Figure 1] with the according feedback modalities.

Figure 1: Track with hard difficulty (contains more barricades than the easy track)

III. Research Aim

The overall aim of investigating the combination of input modality with task complexity and feedback modality is to identify which combination works best for different human-robot collaboration tasks. The outcome should be recommendations for specific task types (and difficulty levels) in terms of suitable input and output modalities. We expect that the input modality will have the highest impact on the overall user performance in a collaborative task, however the task difficulty and the feedback will also effect the results. Moreover, we consider that the feedback modality will influence aspects, such as user satisfaction and system trust more than the input modality.

For our first study we built two identical robots driven by “The Snatcher” from Laurens. [7] [Figure 2], which could be controlled by (1) a gesture-based interface and (2) a PC-remote control. [Figure 3] Both robots were constructed in the same way and both provided aural feedback driven by the sounds of the motors and optical feedback from the robot’s movement. The feedback of the different input modalities was haptic feedback from the keyboard of the PC-remote control and visual and haptic feedback from the mobile phone, which was used for the gesture-based interface.

We assumed that the PC-remote control will achieve the best results in terms of performance (efficiency and effectiveness), but that the gesture-based interface will strengthen the feeling of collaborating with the robot and will be perceived as more intuitive.

Figure 2: The Lego Mindstorm robot.

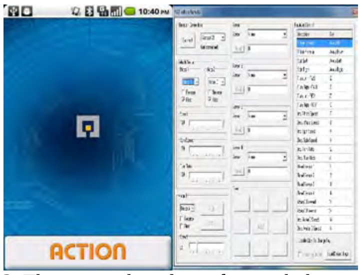

Figure 3: The gesture-based interface with the action button to lift the grabber and on the right the PC-remote with the assignment of keys for each action.

IV. Experimental Setup

In order to explore how input modality, task difficulty, and feedback modality interplay in terms of performance and user satisfaction we set-up a study at an open-house university event, in which visitors were invited to compete each other in a Lego Mindstorm race.

The experiment was set-up as a 2x2 between-subject experiment with the conditions task difficulty (easy/ hard) and input device (gesture-based/ PC-remote)

After a short description of the input devices and the task they had to fulfill they had to navigate their robot through two equal courses and avoid some obstacles with the aim to lift and transport a box to a given goal. One of the participants always used the PC remote the other one the gesture based interface which offers the possibility to compare the two input devices and the according feedback in terms of efficiency and effectiveness. After the race they were asked to fill in a short questionnaire to gather further data. During the day we changed the difficulty of the tracks from easy to hard by adding more obstacles to gain insights on the impact of the different task complexity.

Although the open space context was not easy to control because of the frequent change of visitors, it provided the advantage of studying many participants with different socio-demographic background in a short time.

MEASURES:

Based on a literature review we decided to divide our measures into two categories.

Performance Measures:

- Efficiency: The time users needed to accomplish the course from starting to the end point.
- Effectiveness: The number of errors during the navigation based on collisions with other objects or barricades.

User Satisfaction Measures:

- Perceived task complexity, intuitiveness, satisfaction and acceptance of the device: A questionnaire was used to gather information about these measures. It consisted of a 5-point Likert scale and 15 questions, which were driven from the USUS evaluation framework [8] and from a survey used to assess the intuitiveness of the Nabaztag Robot as an ambient interface. [9]

Moreover, the whole study was videotaped to recheck the measured time and number of collisions as well as to identify problems with the input devices and the according feedback channel.

V. RESULTS

We conducted the study with 24 participants, 10 female and 14 male, age from 11 to 66 years, the average age was 36,73 years and the standard deviation was 14,63. For further analysis four data records were removed because some participants solved the track without an opponent, in that case there was no race condition therefore we didn't use the data. Furthermore one woman let her son win in order to raise his mood which would have influenced the data as a consequence this data record was also removed.

At first we conducted a manipulation check if the task complexity was successfully varied. Therefore, we ran a Mann-Whitney-U Test on the track solution time and number of collisions to check if they significantly differ for the assumed track difficulty. The track solution time differs for the track difficulty. The results of the test were in the expected direction and significant, $z=2{,}162$, $p<.05$. The mean rank for easy track was 7.00 while the average rank for hard one was 12.83.

As expected also the number of collisions was higher for the hard track, $z=2{,}899$, $p<.05$. The mean rank for easy track was 6.00 while the average rank for hard one was 13.50.

After that we computed the scales for the items intuitiveness, satisfaction and the overall acceptance of the questionnaire. Therefore we ran a Cronbach's Alpha test to check the internal reliability for these 3 factors.

For the intuitiveness scale which consists of 5 questions we reached a Cronbach's Alpha of .706, after deleting the item "the used device was hard to use" the Cronbach's Alpha was .733.

The Alpha value for satisfaction with 4 questions was .635 and after deleting "I was satisfied with my own performance" .646.

For acceptance with 5 questions the value was .629 and without "I would not be able to solve a task with the robot with this input device without help" even .741.

Regarding performance the gesture-based interface enabled a less efficient control because the mean track solution time was 73 seconds and for the PC-remote 68 seconds. Concerning effectiveness the absolute number of collisions was 19 for the gesture-based interface and 16 for the PC-remote but the mean number of collisions was 2 for both devices.

Regarding the three scales intuitiveness, satisfaction and acceptance there was a trend that all of them were perceived better when solving the hard track and experienced worse practicing on the easy one even for both devices. [Figure 4]

	Difficulty of Track			
	Easy		Hard	
	Mean	Std Dev	Mean	Std Dev
Intuitiveness	4,43	,72	4,73	,23
Satisfaction	3,96	,84	4,24	,67
Acceptance	4,50	,88	4,58	,53

Figure 4: The three scales grouped by track difficulty

Regarding intuitiveness the gesture-based interface was considered more intuitive but for terms of satisfaction and acceptance the PC-remote was perceived better. [Figure 5]

	Used Device			
	PC-Remote		Gesture-Based	
	Mean	Std Dev	Mean	Std Dev
Intuitiveness	4,55	,61	4,67	,37
Satisfaction	4,28	,56	3,97	,88
Acceptance	4,72	,34	4,38	,87

Figure 5: The three scales grouped by used device

We conducted Mann-Whitney U tests for the three scales (intuitiveness, satisfaction, and overall acceptance) for the types of device and the task difficulty, but unfortunately there was no significance in the results. However, interesting results could be found for single scale items.

A Mann-Whitney U test revealed that one of the two input devices was less intuitive to use, which was in that case the gesture-based interface but only for one item in the intuitiveness scale ("the used device was hard to use") the results of the test were in the expected direction and significant, $z=-2{,}195$, $p<.05$. The mean rank for the mobile control was 7.17 while the average rank for the pc control was 12.55.

Furthermore, the used input device correlates with the own satisfaction of the performance of the participants. The results were also significant, $z=-2{,}160$, $p<.05$. The mean rank for satisfaction with the gesture-based interface was 7.80 while the average rank for the PC-remote was 13.20.

Another interesting result was a significant difference in terms of user satisfaction for one item on likability (I would like to use the device for often) according to the task difficulty (z=2,438, p<.05). The mean rank for the easy track was 6.50 while the average rank for hard one was 12.55.

VI. DISCUSSION

Although the PC-remote outperformed the gesture-based interface in terms of efficiency and effectiveness the difference concerning efficiency was much lower for the hard course. Apart from that the difference in task completion times and number of collisions was huge comparing easy and hard track, but not that much for the perceived task complexity. This means that real task complexity and perceived complexity do not always correlate directly.

Apart from that the perceived task complexity was also dependent on the used input device, which could not be proved statistically but in the descriptive data there is a trend showing this fact. People who used the PC-remote thought that the easy course was more easy than participants who used the gesture-based interface, but for the hard course it was the opposite which means that gesture-based interface users perceived the hard track more easy than participants who used the PC-remote. [Figure 6] One possible reason for that phenomenon could be the visual feedback provided by the gesture-based interface, which was used a lot by the participants and especially people using the PC-remote had problems when they steered the robot back to the starting point of the track because of the fact that they had to steer left when they wanted to turn the robot to the right. Adding visual feedback to the PC-remote could minimize the effect for this device.

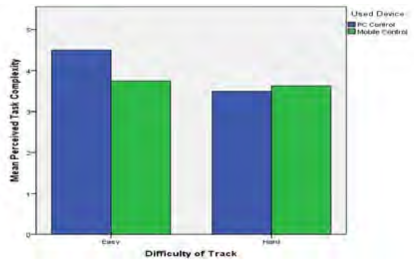

Figure 6: Perceived task complexity dependant on used device and difficulty of track

Moreover 5 participants using the gesture-based interface stood up during the race and moved with the robot, which demonstrates the advantage of such an input modality for this type of task, but also demonstrates the lack of interface-specific feedback which has to be enhanced by context-feedback ("user workaround").

Although the PC-remote was much better in terms of efficiency and effectiveness the differences in intuitiveness, satisfaction and acceptance were not significant, possibly also because of the fact that the gesture-based interface provided more feedback to the participants.

All in all different input devices as well as task complexities have a severe impact on the results. Also the effect of feedback must not be underestimated. Every input device needs specific feedback depending on the device, task complexity and different situations.

VII. FUTURE WORK

In a next step we want to investigate speech input. We plan to conduct a more controlled experiment in a closed laboratory setting adding speech as an input modality for comparison and furthermore different embodiments (functional vs. anthropomorphic) of the robots, which could have an impact on the findings too. It is planned to let a higher number of participants solve given assembly tasks (which will differ in complexity) together with a Lego Mindstorm robot which helps in principle the users to transport Lego parts from A to B to get further insight of the correlation of input methods, task complexities and feedback modalities as well as the appearance of the robot. Finally this basic research study should inform the interaction design for a robotic arm in the context of a factory in which operators have to cooperate also with robots in transportation tasks as simulated in our preliminary study.

REFERENCES

[1] Rouanet, P.; Bechu, J.; Oudeyer, P.-Y.; , "*A comparison of three interfaces using handheld devices to intuitively drive and show objects to a social robot: the impact of underlying metaphors*," Robot and Human Interactive Communication, 2009. RO-MAN 2009.

[2] Pierre Rouanet, Fabien Danieau, and Pierre-Yves Oudeyer. 2011. A robotic game to evaluate interfaces used to show and teach visual objects to a robot in real world condition. In Proceedings of the 6th international conference on Human-robot interaction (HRI '11). ACM, New York, NY, USA, 313-320.

[3] Alexander Bannat, Jürgen Gast, Tobias Rehrl, Wolfgang Rösel, Gerhard Rigoll, and Frank Wallhoff. 2009. *A Multimodal Human-Robot-Interaction Scenario: Working Together with an Industrial Robot*. In Proceedings of the 13th International Conference on Human-Computer Interaction. Part II: Novel Interaction Methods and Techniques, Julie A. Jacko (Ed.). Springer-Verlag, Berlin, Heidelberg, 303-311.

[4] Sean Timothy Hayes, Eli R. Hooten, and Julie A. Adams. 2010. Multi-touch interaction for tasking robots. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction* (HRI '10). IEEE Press, Piscataway, NJ, USA, 97-98.

[5] Cheng Guo and Ehud Sharlin. 2008. *Exploring the use of tangible user interfaces for human-robot interaction: a comparative study*. In Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems (CHI '08). ACM, New York, NY, USA, 121-130.

[6] Pollard, N.S.; Hodgins, J.K.; Riley, M.J.; Atkeson, C.G.; , "*Adapting human motion for the control of a humanoid robot*," Robotics and Automation, 2002. Proceedings. ICRA '02. IEEE International Conference on , vol.2, no., pp. 1390- 1397 vol.2, 2002

[7] Laurens Valk. 2010. *The LEGO MINDSTORMS NXT 2.0 Discovery Book: A Beginner's Guide to Building and Programming Robots* (1st ed.). No Starch Press, San Francisco, CA, USA.

[8] Weiss, A. *Validation of an Evaluation Framework for Human-Robot Interaction. The Impact of Usability, Social Acceptance, User Experience, and Societal Impact on Collaboration with Humanoid Robots.* PhD thesis, University of Salzburg, Feburary 2010.

[9] Mirlacher, T., Buchner, R., Förster, F., Weiss, A., and Tscheligi, M. *Ambient Rabbits – Likeability of Embodied Ambient Displays.* In AmI2009: Proceedings of the 3rd European Conference on Ambient Intelligence (Berlin/Heidelberg, 2009), M. Tscheligi and B. de Ruyter, Eds., Springer, pp. 164–173.

A.2. Late-breaking Report accepted for ACM/IEEE International Conference on Human-Robot Interaction 2013

Input Modality and Task Complexity: Do they Relate?

Gerald Stollnberger, Astrid Weiss, Manfred Tscheligi
CDL on Contextual Interfaces
University of Salzburg
Salzburg, Austria
firstname.lastname@sbg.ac.at

***Abstract*— In the research field of Human-Robot Collaboration (HRC) choosing the right input modality is a crucial aspect for successful cooperation, especially for different levels of task complexity. In this paper we present a preliminary study we conducted in order to investigate the correlation between input modalities and task complexity. We assume that specific input devices are suitable for specific levels of task complexity in HRC tasks. In our study participants could choose between two different input modalities to perform a race with Lego Mindstorm robots against each other. One of the main findings was that both factors (input modality / task complexity) have a severe impact on task performance and user satisfaction. Furthermore, we found out that users' perceptions of their performance differed from reality in some cases.**

***Index Terms*—Human-Robot collaboration; input modalities; task complexity; gesture; speech; remote; mindstorm**

I. Introduction and Motivation

In order to investigate which input modalities should be used for cooperative tasks, especially for different levels of task complexity, we built a Lego Mindstorm robot, which can be controlled by two different input devices: gesture-based interface on a mobile phone and a PC-remote control. These input modalities were studied in a field trial with 24 participants.

Different input modalities suit better for different task complexities. E.g. for simple tasks there is no strong need in a very precise and robust control mechanism. There are many different input modalities to interact with robots, such as speech, gesture, keyboards, etc. which were tested in several studies. Research has shown that different task complexities have a significant impact on the user performance [1]. Furthermore, a more natural and flexible human-robot interaction can be achieved by providing different control modalities at the same time [2].

A lot of research has already been done to compare input modalities for HRC in order to identify their advantages in terms of performance and user experience/usability. In some cases also according to their task-suitability [1],[4]. However, what hasn't been explored so far is a structured investigation of the interplay between: (1) input modality, and (2) task complexity. In order to investigate this interplay, we conducted a preliminary user study in which we compared two input modalities (gesture-based, PC-remote) with two levels of task complexity.

The overall aim of this investigation is to identify which combination works best for different HRC tasks. The outcome should be recommendations for specific task types (and difficulty levels) in terms of suitable input modalities for robots. For a first investigation of this assumption we built two identical robots driven by "The Snatcher" of Laurens [3], which were controlled by (1) a gesture-based interface and (2) a PC-remote control.

II. Experimental Setup

Our aim was to find out how input modality and task difficulty interplay in terms of performance and user satisfaction. We set-up a study at an open-house university event, in which visitors were invited to compete in a Lego Mindstorm race. The experiment was set-up as a 2x2 between-subject experiment with the conditions task difficulty (easy/hard) and input device (gesture-based/ PC-remote).

After a short description of the input devices and the task, they had to navigate their robot through two equal courses and avoid obstacles. The aim was to lift and transport a box to a given goal. After the race the participants had to fill in a questionnaire. During the day we changed the task level of the tracks from easy to hard, by adding more obstacles to gain insight on the impact of different task complexity.

Although the open space context was not easy to control, because of the frequent change of visitors, it provided the advantage of studying many participants with different socio-demographic background in a short time.

Based on various sources we decided to divide our measures into two categories. (1) Performance Measures, such as efficiency (i.e. task completion time) and effectiveness (e.g. number of collisions) and (2) User Satisfaction Measures: perceived task complexity, intuitiveness, satisfaction, and acceptance. We gathered this information with a questionnaire consisting of 15 items, which had to be rated on a 5-point Likert scale. Moreover, the whole study was videotaped to recheck the measured time and number of collisions, as well as to identify problems with the input devices.

III. Results

We conducted the study with 24 participants, 10 female and 14 male, aged from 11 to 66 years (M = 36.73, SD = 14.63). For further analysis four data records were removed because some participants solved the track without an opponent. Therefore, we didn't use the data. Furthermore, one woman let her son win, in order to raise his mood, which would have influenced the data as well. As a consequence this data record was also removed.

At first we conducted a manipulation check, i.e. the task complexity was successfully varied. Therefore, we ran a Mann-Whitney-U Test on the track solution time and number of collisions. We checked if they significantly differ from the assumed track difficulty. The track solution time differed for the two tracks. The results of the test were as expected and significant, $z = 2.162$, $p = .031$. The mean rank for easy track was 7.00, while the average rank for hard one was 12.83. The number of collisions was higher for the hard track, $z = 2.899$, $p = .004$. The mean rank for easy track was 6.00, while the average rank for hard one was 13.50.

Regarding performance the gesture-based interface enabled a less efficient control. The mean track solution time was 73 seconds (SD = 23) and for the PC-remote 68 seconds (SD = 31). Concerning effectiveness the absolute number of collisions was 19 for the gesture-based interface and 16 for the PC-remote.

Then we computed the scales for intuitiveness, satisfaction, and overall acceptance. The Cronbach's Alpha for these factors was between .646 and .741. After deleting unreliable items, the internal reliability for our scales was satisfying.

Regarding the three scales intuitiveness, satisfaction, and acceptance, there was a trend that all of them were perceived better, when solving the hard track, and experienced worse, when practicing on the easy one for both devices. [Fig. 1]

	Difficulty of Track			
	Easy		**Hard**	
	Mean	***Std Dev***	***Mean***	***Std Dev***
Intuitiveness	4.43	.72	4.73	.23
Satisfaction	3.96	.84	4.24	.67
Acceptance	4.50	.88	4.58	.53

Fig. 1. The three scales grouped by track difficulty

Regarding intuitiveness the gesture-based interface was considered more intuitive, but in terms of satisfaction and acceptance the PC-remote was perceived better.

We conducted Mann-Whitney U tests for the three scales (intuitiveness, satisfaction, and overall acceptance) for the types of device and the task difficulty, but no significant differences could be identified in the results.

Another test revealed that the used input device correlates with the own satisfaction of the performance, $z = -2.160$, $p = .031$. The mean rank for satisfaction with the gesture-based interface was 7.80, while the mean rank for the PC-remote was 13.20. Moreover, for the item likability, (I would like to use the device often) a difference, regarding task difficulty ($z = 2.438$, $p = .015$), could be identified. The mean rank for the easy track was 6.50, while the average rank for the hard one was 12.55. This implies that the likability was higher for harder tasks.

IV. Discussion

Although the PC-remote outperformed the gesture-based interface in terms of efficiency and effectiveness, the difference concerning efficiency was much lower for the hard course.

Apart from that, the difference in task completion times and number of collisions was huge, comparing easy and hard track, but not that much for the perceived task complexity. This means: real task complexity and perceived complexity does not always directly correlate. Apart from that, the perceived task complexity was also dependent from the used input device. It could not be proved statistically, but in the descriptive data we found a trend supporting this assumption. People using the PC-remote perceived the easy course easier than participants using the gesture-based interface, but for the hard course it was the opposite. [Fig. 2]

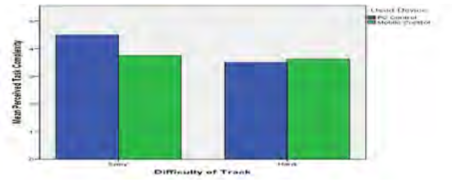

Fig. 2. Perceived task complexity dependant on used device and difficulty of track

One possible reason for that phenomenon could be that people using the PC-remote, had problems steering the robot back to the starting point. They had to steer left when they wanted the robot to go right. Moreover, five participants, who used the gesture-based interface, stood up during the race and moved with the robot. This demonstrates the advantage of such an input modality for this type of task.

To sum up, we could show that different input modalities and task complexity levels influenced our measures. Every task needs specific input modalities, depending on the complexity and situation to reach higher performance and user satisfaction. A next step will be to investigate speech input and the effect of the embodiment (functional vs. anthropomorphic). We assume this could also affect task performance and user satisfaction.

Acknowledgment

We greatly acknowledge the financial support by the Federal Ministry of Economy, Family and Youth and the National Foundation for Research, Technology and Development (Christian Doppler Laboratory for “Contextual Interfaces”).

References.

[1] Rouanet, P.; Bechu, J.; Oudeyer, P, "A comparison of three interfaces using handheld devices to intuitively drive and show objects to a social robot: the impact of underlying metaphors," Robot and Human Interactive Communication. RO-MAN 2009.

[2] A. Bannat, J. Gast, T. Rehrl, W. Rösel, G. Rigoll, F. Wallhoff. A Multimodal Human-Robot-Interaction Scenario: Working Together with an Industrial Robot. Proc of the 13th Int. Conf. on HCI, Part 2: Novel Interactionmethods and Techniques, J. Jacko (Ed.) Springer-Verlag Berlin, Heidelberg, 303-311.

[3] Laurens V. The Lego Mindstorms NXT 2.0 Discovery Book: A Beginner's Guide to Building and Programming Robots (1st ed.). No Starch Press, San Francisco, CA, USA.

[4] Cheng G. Ehud S. 2008. Exploring the use of tangible user interfaces for human-robot interaction: a comparative study. In Proc. of the 26th annual SIGCHI conference on Human factors in computing systems (CHI '08). ACM, New York, USA.

A.3. Data Consent Form for the Preliminary Study

ICT&S Center

UNIVERSITÄT SALZBURG

Control ME Nutzerstudie

Teilnehmer Nr.: ___ Salzburg, Juni 2012

Datenverwertungserlaubnis

Liebe/r Teilnehmer/in!

Diese Studie wird von der HCI und Usability Unit des ICT&S Center (Advanced Studies and Research in Information and Communication Technologies & Society) der Universität Salzburg durchgeführt; die/der verantwortliche Projektleiterin ist Astrid Weiss. Um die Interaktion eines Roboters mit Benutzern zu optimieren, wird in dieser Vorstudie die Interaktion zwischen Mensch und Roboter untersucht.

Das Material (Text, Audio, Video, Fotos), das während der Studie erstellt wird, wird in weiterer Folge zur Auswertung bzw. Erarbeitung der entsprechenden Untersuchungsergebnisse verwendet.

Sie erklären sich einverstanden, dass das Material für diese Auswertungen verwendet wird. Das anonymisierte Rohmaterial kann für Präsentationen und wissenschaftlichen Publikationen im Rahmen der Studie verwendet werden, wird aber nicht an Dritte weitergegeben. Fotos und Videoausschnitte können selektiv für die Präsentationen und wissenschaftlichen Ergebnisse bzw. der beispielhaften Darstellung des wissenschaftlichen Tätigkeitsfeldes der HCI und Usability Unit des ICT&S Centers verwendet werden.

Mit dieser Erklärung gebe ich die Erlaubnis, als Teilnehmer/in an der Studie wie oben beschrieben, gefilmt bzw. aufgenommen zu werden.

Ich verpflichte mich weiters dazu, sämtliche Informationen zu dieser Studie vertraulich zu behandeln und nicht an Dritte weiterzugeben.

Name in Blockbuchstaben: ______________________________

Unterschrift: ______________________________

Bestätigung: Gerald Stollnberger (Diplomand ICT&S Center).

Rückfragen: Tel. +43-662-8044-4836, Email: gerald.stollnberger@sbg.ac.at

Unterschrift: ______________________________

A.4. Questionnaire used in the Preliminary Study

ICT&S Center   UNIVERSITÄT SALZBURG

Das Roboter Rennen

Fragebogen

Geschlecht: ❑ m ❑ w   **Alter:** _______   **TP Nr.:** _______

Welche Steuerung wurde verwendet? ❑Handy ❑PC

	nicht	wenig	mittel	ziemlich	sehr	k.A.
Die Strecke war schwer zu befahren	☐	☐	☐	☐	☐	☐

Warum? ____

Wie sehr treffen die folgenden Aussagen zu?

	nicht	wenig	mittel	ziemlich	sehr	k.A.
Ich fand die verwendete Steuerung unnötig komplex.	☐	☐	☐	☐	☐	☐
Ich fand die verwendete Steuerung leicht verständlich.	☐	☐	☐	☐	☐	☐
Ich kann mir vorstellen, dass die meisten Leute sehr schnell lernen würden, mit dieser Steuerung umzugehen.	☐	☐	☐	☐	☐	☐
Ich würde sehr viel lernen müssen, bevor ich mit dieser Steuerung umgehen könnte.	☐	☐	☐	☐	☐	☐
Es war für mich schwierig, den Roboter zu steuern.	☐	☐	☐	☐	☐	☐

Wie sehr treffen die folgenden Aussagen zu?

	nicht	wenig	mittel	ziemlich	sehr	k.A.
Ich denke, dass ich diese Steuerung gerne häufig benutzen würde.	☐	☐	☐	☐	☐	☐
Ich war zufrieden mit der Leistung des Roboters.	☐	☐	☐	☐	☐	☐
Ich war zufrieden mit der Leistung der verwendeten Steuerung.	☐	☐	☐	☐	☐	☐
Ich war zufrieden mit meiner persönlichen Leistung.	☐	☐	☐	☐	☐	☐

ICT&S Center   UNIVERSITÄT SALZBURG

Wie sehr treffen die folgenden Aussagen zu?

	nicht	wenig	mittel	ziemlich	sehr	k.A.
Ich habe nicht die notwendigen Fähigkeiten, um Roboter mithilfe dieser Steuerung bedienen zu können.	☐	☐	☐	☐	☐	☐
Ich habe das notwendige Wissen um mit dieser Steuerung umzugehen.	☐	☐	☐	☐	☐	☐
Ich könnte keine Aufgabe mit Hilfe dieser Steuerung und den Roboter lösen, wenn niemand da wäre, den ich fragen kann.	☐	☐	☐	☐	☐	☐
Unter Zeitdruck könnte ich niemals Erfolg bei der Zusammenarbeit mit Robotern haben wenn ich diese Steuerung verwende.	☐	☐	☐	☐	☐	☐
Ich werde nie eine Aufgabe gemeinsam mit Robotern mit dieser Steuerung lösen können.	☐	☐	☐	☐	☐	☐

Gibt es noch Anmerkungen? Wenn ja, welche?

____

DANKE!

A.5. Output of the Data Analysis of the Preliminary Study

```
GET
  FILE='C:\Users\Stollnberger\Desktop\Diplomarbeit\50JahrFeier\AuswertungBilder50jFeier\50JFeierSPSS.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
*Table of all Variables

DESCRIPTIVES VARIABLES=PA_NR Gender Age Device TrackDiff SolTime NumOfColl PerceivedTComplexity
    IComplexity IComprehensible ILearn IHLearn IHard SUse SatRobot SatDevice SatSelf ASkill AKnowledge
    AHelp APressure ANever
  /STATISTICS=MEAN STDDEV MIN MAX.
```

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Number of Participant	20	1	24	13,60	7,221
Gender	20	1	2	1,45	,510
Age of Participant	19	11	66	35,63	15,075
Used Device	20	1	2	1,50	,513
Difficulty of Track	20	1	2	1,60	,503
Track Solution Time in Seconds	20	36	139	70,85	26,812
Number of Collisions	20	0	5	1,75	1,743
Perceived Task Complexity	20	2	5	3,90	1,071
Used Control Device was needlessly complex	20	3	5	4,55	,686
Used Control Device was easy comprehensible	20	3	5	4,75	,550
Used Control Device was fast to learn	19	1	5	4,32	,946
Used Control Device needs much to learn	20	4	5	4,80	,410
Used Control Device was hard to use	19	1	5	3,79	1,084
Like to use device often	19	2	5	3,74	,933
Satisfied with rob ots' performance	20	2	5	4,35	,988
Satisfied with devices' performance	20	2	5	4,30	,979
Satisfied with own performance	20	1	5	3,85	1,387
Own required skill for commanding robots with device	20	3	5	4,65	,745

Own required knowledge for commanding robots with device	20	3	5	4,65	,587
Would not be able to solve a task without help	20	1	5	3,75	1,446
Would not be able to solve a task under time pressure	20	2	5	4,30	1,031
Would never be able to solve a task	20	1	5	4,60	1,095
Valid N (listwise)	16				

```
DATASET ACTIVATE DataSet1.
* Custom Tables. All Variables grouped by Used Device and Task Complexity

CTABLES
  /VLABELS VARIABLES=Age SolTime NumOfColl PerceivedTComplexity IComplexity IComprehensible ILearn
    IHLearn IHard SUse SatRobot SatDevice SatSelf ASkill AKnowledge AHelp APressure ANever
    Intuitiveness Satisfaction Acceptance TrackDiff Device
    DISPLAY=LABEL
  /TABLE Age [MEAN] + SolTime [MEAN] + NumOfColl [MEAN] + PerceivedTComplexity [MEAN] + IComplexity
    [MEAN] + IComprehensible [MEAN] + ILearn [MEAN] + IHLearn [MEAN] + IHard [MEAN] + SUse [MEAN] +
    SatRobot [MEAN] + SatDevice [MEAN] + SatSelf [MEAN] + ASkill [MEAN] + AKnowledge [MEAN] + AHelp
    [MEAN] + APressure [MEAN] + ANever [MEAN] + Intuitiveness [MEAN] + Satisfaction [MEAN] + Acceptance
    [MEAN] BY TrackDiff > Device
  /CATEGORIES VARIABLES=TrackDiff Device ORDER=A KEY=VALUE EMPTY=INCLUDE.
```

Custom Tables

	Difficulty of Track			
	Easy		Hard	
	Used Device		Used Device	
	PC Control	Mobile Control	PC Control	Mobile Control
	Mean	Mean	Mean	Mean
Age of Participant	44	28	38	32
Track Solution Time in Seconds	45	64	84	80
Number of Collisions	0	1	3	3
Perceived Task Complexity	5	4	4	4
Used Control Device was needlessly complex	5	4	5	5
Used Control Device was easy comprehensible	4	5	5	5
Used Control Device was fast to learn	4	5	5	4

Used Control Device needs much to learn	5	5	5	5
Used Control Device was hard to use	4	3	4	3
Like to use device often	4	3	4	4
Satisfied with rob ots' performance	5	4	4	4
Satisfied with devices' performance	5	4	5	4
Satisfied with own performance	5	2	5	4
Own required skill for commanding robots with device	5	5	5	4
Own required knowledge for commanding robots with device	5	4	5	5
Would not be able to solve a task without help	4	4	3	4
Would not be able to solve a task under time pressure	5	4	4	4
Would never be able to solve a task	5	4	5	4
Intuitiveness	4,25	4,60	4,75	4,71
Satisfaction	4,42	3,50	4,19	4,28
Acceptance	4,69	4,31	4,75	4,42

```
*Cronbach Alpha for intuitiveness

RELIABILITY
  /VARIABLES=IComplexity IComprehensible ILearn IHLearn IHard
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /SUMMARY=TOTAL.
```

Reliability

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	18	90,0
	Excluded[a]	2	10,0
	Total	20	100,0

Reliability Statistics

Cronbach's Alpha	N of Items
,706	5

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Used Control Device was needlessly complex	17,61	4,840	,603	,607
Used Control Device was easy comprehensible	17,44	5,320	,587	,631
Used Control Device was fast to learn	17,89	4,810	,351	,719
Used Control Device needs much to learn	17,39	5,546	,733	,624
Used Control Device was hard to use	18,33	4,235	,390	,726

```
*Cronbach Alpha for Satisfaction
RELIABILITY
  /VARIABLES=SUse SatRobot SatDevice SatSelf
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /SUMMARY=TOTAL.
```

Reliability

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	19	95,0
	Excluded[a]	1	5,0
	Total	20	100,0

Reliability Statistics

Cronbach's Alpha	N of Items
,635	4

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Like to use device often	12,53	6,708	,313	,630
Satisfied with rob ots' performance	11,89	5,988	,420	,564
Satisfied with devices' performance	11,89	5,322	,649	,415
Satisfied with own performance	12,47	4,930	,356	,646

```
*Cronbach Alpha for Acceptance

RELIABILITY
  /VARIABLES=ASkill AKnowledge AHelp APressure ANever
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /SUMMARY=TOTAL.
```

Reliability

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	20	100,0
	Excluded[a]	0	,0
	Total	20	100,0

Reliability Statistics

Cronbach's Alpha	N of Items
,629	5

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Own required skill for commanding robots with device	17,30	7,695	,563	,519
Own required knowledge for commanding robots with device	17,30	8,326	,562	,547
Would not be able to solve a task without help	18,20	7,116	,177	,741
Would not be able to solve a task under time pressure	17,65	5,818	,743	,373
Would never be able to solve a task	17,35	8,029	,217	,660

```
*Compute Variable Intuitiveness without Hard to use for better Cronbach Alpha

COMPUTE Intuitiveness=MEAN(IComplexity,IComprehensible,ILearn,IHLearn).
EXECUTE.

*Compute variable Satisfaction without own performance for better Cronbach Alpha

COMPUTE Satisfaction=MEAN(SUse,SatRobot,SatDevice).
EXECUTE.

*Compute Variable Acceptance without Help for better Cronbach Alpha

COMPUTE Acceptance=MEAN(ASkill,AKnowledge,APressure,ANever).
EXECUTE.

*Tables for new variables

DESCRIPTIVES VARIABLES=Intuitiveness Satisfaction Acceptance
  /STATISTICS=MEAN STDDEV MIN MAX.
```

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Intuitiveness	20	3,00	5,00	4,6083	,49567
Satisfaction	20	2,33	5,00	4,1250	,73523
Acceptance	20	2,50	5,00	4,5500	,66689
Valid N (listwise)	20				

```
*Nonparametric Tests: Independent Samples. grouped by Device, only Single questions

NPTESTS
  /INDEPENDENT TEST (Age SolTime NumOfColl PerceivedTComplexity IComplexity IComprehensible ILearn
    IHLearn IHard SUse SatRobot SatDevice SatSelf ASkill AKnowledge AHelp APressure ANever) GROUP
    (Device)
  /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
  /CRITERIA ALPHA=0.05  CILEVEL=95.
```

Nonparametric Tests

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Age of Participant is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	.219	Retain the null hypothesis.
2	The distribution of Track Solution Time in Seconds is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	.520	Retain the null hypothesis.
3	The distribution of Number of Collisions is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	.608	Retain the null hypothesis.
4	The distribution of Perceived Task Complexity is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	.812	Retain the null hypothesis.
5	The distribution of Used Control Device was needlessly complex is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	.687	Retain the null hypothesis.
6	The distribution of Used Control Device was easy comprehensible is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	.255	Retain the null hypothesis.
7	The distribution of Used Control Device was fast to learn is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	1.000	Retain the null hypothesis.
8	The distribution of Used Control Device needs much to learn is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	1.000	Retain the null hypothesis.
9	The distribution of Used Control Device was hard to use is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	.028	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
10	The distribution of Like to use device often is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	.830	Retain the null hypothesis.
11	The distribution of Satisfied with robots' performance is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	.796	Retain the null hypothesis.
12	The distribution of Satisfied with devices' performance is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	.229	Retain the null hypothesis.
13	The distribution of Satisfied with own performance is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	.031	Reject the null hypothesis.
14	The distribution of Own required skill for commanding robots with device is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	.212	Retain the null hypothesis.
15	The distribution of Own required knowledge for commanding robots with device is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	.888	Retain the null hypothesis.
16	The distribution of Would not be able to solve a task without help is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	.319	Retain the null hypothesis.
17	The distribution of Would not be able to solve a task under time pressure is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	.864	Retain the null hypothesis.
18	The distribution of Would never be able to solve a task is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	.466	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

```
*Nonparametric Tests: Independent Samples. grouped by Difficulty of the task, only Single questions

NPTESTS
  /INDEPENDENT TEST (Age SolTime NumOfColl PerceivedTComplexity IComplexity IComprehensible ILearn
    IHLearn IHard SUse SatRobot SatDevice SatSelf ASkill AKnowledge AHelp APressure ANever) GROUP
    (TrackDiff)
  /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
  /CRITERIA ALPHA=0.05  CILEVEL=95.
```

Nonparametric Tests

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Age of Participant is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	.804	Retain the null hypothesis.
2	The distribution of Track Solution Time in Seconds is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	.031	Reject the null hypothesis.
3	The distribution of Number of Collisions is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	.004	Reject the null hypothesis.
4	The distribution of Perceived Task Complexity is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	.145	Retain the null hypothesis.
5	The distribution of Used Control Device was needlessly complex is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	.293	Retain the null hypothesis.
6	The distribution of Used Control Device was easy comprehensible is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	.108	Retain the null hypothesis.
7	The distribution of Used Control Device was fast to learn is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	.775	Retain the null hypothesis.
8	The distribution of Used Control Device needs much to learn is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	.119	Retain the null hypothesis.
9	The distribution of Used Control Device was hard to use is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	.894	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
10	The distribution of Like to use device often is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	.015	Reject the null hypothesis.
11	The distribution of Satisfied with robots' performance is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	.356	Retain the null hypothesis.
12	The distribution of Satisfied with devices' performance is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	1.000	Retain the null hypothesis.
13	The distribution of Satisfied with own performance is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	.369	Retain the null hypothesis.
14	The distribution of Own required skill for commanding robots with device is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	.740	Retain the null hypothesis.
15	The distribution of Own required knowledge for commanding robots with device is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	.102	Retain the null hypothesis.
16	The distribution of Would not be able to solve a task without help is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	.569	Retain the null hypothesis.
17	The distribution of Would not be able to solve a task under time pressure is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	.335	Retain the null hypothesis.
18	The distribution of Would never be able to solve a task is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	.385	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

```
*Nonparametric Tests: Independent Samples. grouped by Device, computed Variables

NPTESTS
  /INDEPENDENT TEST (Intuitiveness Satisfaction Acceptance) GROUP (Device)
  /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
  /CRITERIA ALPHA=0.05  CILEVEL=95.
```

Nonparametric Tests

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Intuitiveness is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	.969	Retain the null hypothesis.
2	The distribution of Satisfaction is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	.468	Retain the null hypothesis.
3	The distribution of Acceptance is the same across categories of Used Device.	Independent-Samples Mann-Whitney U Test	.571	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

```
*Nonparametric Tests: Independent Samples. grouped by Task Complexity, computed Variables

NPTESTS
  /INDEPENDENT TEST (Intuitiveness Satisfaction Acceptance) GROUP (TrackDiff)
  /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
  /CRITERIA ALPHA=0.05  CILEVEL=95.
```

Nonparametric Tests

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Intuitiveness is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	.575	Retain the null hypothesis.
2	The distribution of Satisfaction is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	.483	Retain the null hypothesis.
3	The distribution of Acceptance is the same across categories of Difficulty of Track.	Independent-Samples Mann-Whitney U Test	.869	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

```
*Cronbach Alpha Test Intuitiveness without the item hard to use!

RELIABILITY
  /VARIABLES=IComplexity IComprehensible ILearn IHLearn
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /SUMMARY=TOTAL.
```

Reliability

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	19	95,0
	Excluded[a]	1	5,0
	Total	20	100,0

Reliability Statistics

Cronbach's Alpha	N of Items
,733	4

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Used Control Device was needlessly complex	13,84	2,585	,486	,696
Used Control Device was easy comprehensible	13,68	2,561	,705	,592
Used Control Device was fast to learn	14,11	1,988	,474	,768
Used Control Device needs much to learn	13,63	3,023	,650	,662

```
*Cronbach's Alpha Test Satisfaction without the item Satisfied with own performance

RELIABILITY
  /VARIABLES=SUse SatRobot SatDevice
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /SUMMARY=TOTAL.
```

Reliability

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	19	95,0
	Excluded[a]	1	5,0
	Total	20	100,0

Reliability Statistics

Cronbach's Alpha	N of Items
,646	3

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Like to use device often	8,74	3,316	,218	,832
Satisfied with rob ots' performance	8,11	2,322	,514	,463
Satisfied with devices' performance	8,11	2,099	,693	,195

```
*Cronbach's Alpha Test Acceptance without the item Never be able to solve a task without help

RELIABILITY
  /VARIABLES=ASkill AKnowledge APressure ANever
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /SUMMARY=TOTAL.
```

Reliability

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	20	100,0
	Excluded[a]	0	,0
	Total	20	100,0

Reliability Statistics

Cronbach's Alpha	N of Items
,741	4

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Own required skill for commanding robots with device	13,55	4,576	,622	,645
Own required knowledge for commanding robots with device	13,55	5,208	,583	,688
Would not be able to solve a task under time pressure	13,90	3,463	,675	,590
Would never be able to solve a task	13,60	4,147	,396	,790

A.6. Data Consent Form for the Laboratory Study

ICT&S Center

UNIVERSITÄT SALZBURG

Roboter Interaktions Nutzerstudie

Teilnehmer Nr.: ___ Salzburg, November 2012

Datenverwertungserlaubnis

Liebe/r Teilnehmer/in!

Diese Studie wird von der HCI und Usability Unit des ICT&S Center (Advanced Studies and Research in Information and Communication Technologies & Society) der Universität Salzburg durchgeführt; die/der verantwortliche Projektleiterin ist Astrid Weiss. Um die Interaktion eines Roboters mit Benutzern zu optimieren, wird in dieser Studie die Interaktion zwischen Mensch und Roboter untersucht.

Das Material (Text, Audio, Video, Fotos), das während der Studie erstellt wird, wird in weiterer Folge zur Auswertung bzw. Erarbeitung der entsprechenden Untersuchungsergebnisse verwendet.

Sie erklären sich einverstanden, dass das Material für diese Auswertungen verwendet wird. Das anonymisierte Rohmaterial kann für Präsentationen und wissenschaftlichen Publikationen im Rahmen der Studie verwendet werden, wird aber nicht an Dritte weitergegeben. Fotos und Videoausschnitte können selektiv für die Präsentationen und wissenschaftlichen Ergebnisse bzw. der beispielhaften Darstellung des wissenschaftlichen Tätigkeitsfeldes der HCI und Usability Unit des ICT&S Centers verwendet werden.

Mit dieser Erklärung gebe ich die Erlaubnis, als Teilnehmer/in an der Studie wie oben beschrieben, gefilmt bzw. aufgenommen zu werden.

Ich verpflichte mich weiters dazu, sämtliche Informationen zu dieser Studie vertraulich zu behandeln und nicht an Dritte weiterzugeben.

Name in Blockbuchstaben: ______________________________

Unterschrift: ______________________________

Bestätigung: Gerald Stollnberger (Diplomand ICT&S Center).

Rückfragen: Tel. +43-664-2830-569, Email: gerald.stollnberger@sbg.ac.at

Unterschrift: ______________________________

A.7. Studycycle

Testphases

Introduction and Briefing

Guten Tag! Mein Name ist Gerald Stollnberger und ich bin am ICT&S Center der Universität Salzburg beschäftigt. Diese Studie findet im Rahmen meiner Diplomarbeit statt, welche ich am ICT&S Center zum Thema „Der Effekt von Eingabegeräten, verschiedenen Aufgabenschwierigkeiten und die Optik von Robotern auf die Performance in Mensch-Roboter Zusammenarbeit." schreibe. Vielen Dank für Ihre Bereitschaft an unserer Studie teilzunehmen. Sie helfen uns wertvolle Einblicke zu gewinnen und so unsere Forschungsarbeit voranzutreiben.

Als erstes darf ich Sie bitten, diese Datenverwertungserlaubnis zu unterschreiben da der gesamte Test auf Video aufgezeichnet wird.. Selbstverständlich werden alle Daten vertraulich behandelt und anonymisiert, und nur für unsere Studienzwecke verwendet.

Vielen Dank! Damit wir die Studie optimal auswerten können, müssen wir diesen Test auf Video aufzeichnen. Damit wir diese Daten auswerten dürfen, brauche ich Ihre schriftliche Einwilligung und bitte Sie diese Datenverwertungserlaubnis zu unterzeichnen. Selbstverständlich werden auch diese Daten vertraulich behandelt, anonymisiert und nur für unsere Studienzwecke verwendet.

[Datenverwertungserlaubnis unterschreiben] Appendix D

Explanations

In meiner Studie geht es darum, die Interaktion zwischen Mensch und Roboter zu analysieren. Zu diesem Zweck habe ich hier eine kleine Strecke mit Hindernissen aufgebaut, durch die Sie in weiterer Folge dann den Roboter steuern werden.

[Der LegoMindstorms Roboter wird hergezeigt.]

Damit wir die Daten im Anschluss an die Studie auswerten können zeichnen wir den Ablauf der Studie mit dieser Kamera auf.

[Die Kamera wird nochmal kurz gezeigt.]

Dies ist der Bautask den sie ausführen sollen, die Teile dafür sind in der Box die sie da vorne sehen.

[Auf die Box zeigen] Ihre Aufgabe ist es den Roboter zu dieser Box zu steuern, diese aufzuheben und hier bei Ihnen abzulegen. Die verfügbaren Kommandos hierfür sind auf diesem Zettel vermerkt.

[Kommandos zeigen]

Bitte lassen Sie sich von der Aufzeichnung nicht stören. Es geht in der Studie nicht darum Ihr Verhalten zu analysieren, sondern das Zusammenspiel zwischen einem Roboter und einem Menschen. Sie können absolut nichts falsch machen! Haben Sie noch Fragen vor Beginn des Tests?

[evtl. Beantwortung von Fragen]

Test run

[Lego Mindstorm wird im Szenario platziert]

[Der Teilnehmer nimmt am Tisch platz]

Wenn sie bereit sind, geben sie mir Bescheid. [Programm, Videoaufzeichnung und Stoppuhr starten]

Teilnehmer steuert Roboter durch den Kurs und holt die Box

Teilnehmer baut die Lego Anleitung nach

Das Szenario ist beendet sobald der Bautask komplett ist. [Zeiten für Robotersteuern und Bautask werden separat gemessen]

[Analog für PC-Remote und Gesture-based Interface, also insgesamt 3 Durchläufe.]

[Taskfragebogen ausfüllen lassen] Appendix A

[Workload Fragebogen ausfüllen lassen] Appendix C

[Diese Phase findet 3mal statt mit 3 Bauanleitungen und 3 Input Devices]

Final interview

Zum Abschluss füllen Sie mir bitte noch einen letzten Fragebogen aus, danach möchte ich Ihnen noch eine Frage stellen.

[Questionnaire on embodiment] Appendix B

Frage: Glauben Sie, dass das äußere Erscheinungsbild des Roboters für diese Aufgabe geeignet war?

[Eventuell nachfragen warum, etc..]

Debriefing

Wir sind nun am Ende des Test angelangt. Ich bedanke mich sehr herzlich für Ihre Teilnahme und Ihr Engagement. Kann ich Ihnen noch Fragen zur Studie beantworten?

[Researcher answers the questions of the participant]

Vielen Dank, auf Wiedersehen!

A.8. Questionnaire concerning the User Satisfaction Measures used in the Laboratory Study

ICT&S Center — UNIVERSITÄT SALZBURG

Roboter Interaktionsstudie
Aufgabenfragebogen

Geschlecht: ❑ m ❑ w **Alter:** ______ TP Nr.: ______ CT Nr.:______

Welche Steuerung wurde verwendet? ❑Handy ❑PC ❑Sprache

	nicht	Wenig	mittel	ziemlich	sehr	k.A.
Die Strecke war schwer zu befahren	☐	☐	☐	☐	☐	☐

Warum? ____

	nicht	Wenig	mittel	ziemlich	sehr	k.A.
Die Bauaufgabe war schwer zu vervollständigen	☐	☐	☐	☐	☐	☐

Warum? ____

Wie sehr treffen die folgenden Aussagen zu?

	nicht	wenig	mittel	ziemlich	sehr	k.A.
Ich fand die verwendete Steuerung unnötig komplex.	☐	☐	☐	☐	☐	☐
Ich fand die verwendete Steuerung leicht verständlich.	☐	☐	☐	☐	☐	☐
Ich kann mir vorstellen, dass die meisten Leute sehr schnell lernen würden, mit dieser Steuerung umzugehen.	☐	☐	☐	☐	☐	☐
Ich würde sehr viel lernen müssen, bevor ich mit dieser Steuerung umgehen könnte.	☐	☐	☐	☐	☐	☐
Es war für mich schwierig, den Roboter zu steuern.	☐	☐	☐	☐	☐	☐

Wie sehr treffen die folgenden Aussagen zu?

	nicht	wenig	mittel	ziemlich	sehr	k.A.
Ich denke, dass ich diese Steuerung gerne häufig benutzen würde.	☐	☐	☐	☐	☐	☐
Ich war zufrieden mit der Leistung des Roboters.	☐	☐	☐	☐	☐	☐
Ich war zufrieden mit der Leistung der verwendeten Steuerung.	☐	☐	☐	☐	☐	☐
Ich war zufrieden mit meiner persönlichen Leistung.	☐	☐	☐	☐	☐	☐

ICT&S, Human-Computer Interaction & Usability Unit www.icts.sbg.ac.at

Wie sehr treffen die folgenden Aussagen zu?

	nicht	wenig	mittel	ziemlich	sehr	k.A.
Ich habe nicht die notwendigen Fähigkeiten, um Roboter mithilfe dieser Steuerung bedienen zu können.	☐	☐	☐	☐	☐	☐
Ich habe das notwendige Wissen um mit dieser Steuerung umzugehen.	☐	☐	☐	☐	☐	☐
Ich könnte keine Aufgabe mit Hilfe dieser Steuerung und den Roboter lösen, wenn niemand da wäre, den ich fragen kann.	☐	☐	☐	☐	☐	☐
Unter Zeitdruck könnte ich niemals Erfolg bei der Zusammenarbeit mit Robotern haben wenn ich diese Steuerung verwende.	☐	☐	☐	☐	☐	☐
Ich werde nie eine Aufgabe gemeinsam mit Robotern mit dieser Steuerung lösen können.	☐	☐	☐	☐	☐	☐

	Stimme sehr zu	Stimme zu	Weder noch	Stimme nicht zu	Stimme gar nicht zu
Das System ist sehr betriebssicher.	☐	☐	☐	☐	☐
Das System hat die Fähigkeit das umzusetzen, was ich umsetzen will.	☐	☐	☐	☐	☐
Das System ist extrem zuverlässig.	☐	☐	☐	☐	☐
Das System hat die Funktionalität die ich brauche.	☐	☐	☐	☐	☐
Für mich funktioniert das System.	☐	☐	☐	☐	☐
Das System hat die Eigenschaften, die ich für die Erfüllung meiner Aufgaben brauche.	☐	☐	☐	☐	☐
Das System lässt mich nicht im Stich.	☐	☐	☐	☐	☐

Gibt es noch Anmerkungen? Wenn ja, welche?

____

DANKE!

A.9. The NASA Raw Task Load Index used in the Laboratory Study

ICT&S Center

UNIVERSITÄT SALZBURG

Roboter Interaktionsstudie

Anstrengungsfragebogen

TP Nr.: ________ CT Nr.:________

Mit diesem Fragebogen möchte ich herausfinden, wie anstrengend oder belastend Sie das Bearbeiten des Tasks (Roboter steuern und Bauaufgabe) empfunden haben.
Im ersten Teil des Fragebogens sollen Sie ankreuzen, wie anstrengend Sie das Bearbeiten des Tasks in mehrerer Hinsicht empfunden haben. Kreuzen Sie bitte die für Sie zutreffende Anstrengungsstärke an:

Unter **geistiger Beanspruchung** versteht man das Ausmaß, in dem Denk- und Verstehensbemühungen notwendig waren, um die Aufgabe zu erfüllen (z.B. Denken, Entscheiden, Berechnen, Erinnern, Nachschauen, Suchen,...)
Meine **geistige Beanspruchung war:**

Niedrig Hoch

Mit **körperlicher Beanspruchung** ist gemeint, wie viel körperliche Aktivität nötig war, um die Aufgabe zu erfüllen.
Meine **körperliche Beanspruchung** war:

Niedrig Hoch

Mit **Zeitdruck** ist die Geschwindigkeit gemeint, in der die Aufgabe ausgefüllt werden musste (langsam oder schnell, ausreichend Zeit zum Fertigwerden oder zuwenig Zeit).
Der **Zeitdruck** war:

Niedrig Hoch

Der **Leistungsdruck** bezieht sich darauf, inwieweit Sie denken, die Ziele der Aufgabe erreicht zu haben, wie zufrieden Sie mit Ihrer Leistung waren.
Der **Leistungsdruck** war:

Niedrig Hoch

Die **Mühe** bezieht sich darauf, wie hart man arbeiten musste, um die Ziele der Aufgabe zu erreichen.
Ich habe mir während dem Bearbeiten der Aufgabe **Mühe** gegeben:

Niedrig Hoch

Die **Frustration** bezieht sich darauf, wie unsicher, entmutigt, verwirrt, gestresst oder genervt man sich während der Erfüllung der Aufgabe gefühlt hat.
Meine **Frustration** war:

Niedrig Hoch

DANKE!

ICT&S, Human-Computer Interaction & Usability Unit www.icts.sbg.ac.at

A.10. The Goodspeed Questionnaire about Anthropomorphism in German

ICT&S Center UNIVERSITÄT SALZBURG

Roboter Interaktionsstudie

Fragebogen zum äußeren Erscheinungsbild

Geschlecht: ❑ m ❑ w **Alter:** _______ TP Nr.: _______

Hatte der Roboter ein Gesicht? ❑Ja ❑Nein

Bitte beurteilen Sie Ihren Eindruck des Roboters auf diesen Skalen:

	1	2	3	4	5	
Unecht	☐	☐	☐	☐	☐	Natürlich
Wie eine Maschine	☐	☐	☐	☐	☐	Wie ein Mensch
Hat kein Bewusstsein	☐	☐	☐	☐	☐	Hat ein Bewusstsein
Künstlich	☐	☐	☐	☐	☐	Realistisch
Bewegt sich steif	☐	☐	☐	☐	☐	Bewegt sich flüssig

Gibt es Anmerkungen? Wenn ja, welche?

____

DANKE!

ICT&S, Human-Computer Interaction & Usability Unit www.icts.sbg.ac.at

A.11. Full Paper accepted for the 2013 IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)

"The Harder it Gets": Exploring the Interdependency of Input Modalities and Task Complexity in Human-Robot Collaboration

Gerald Stollnberger,[1] Astrid Weiss[2] and Manfred Tscheligi[3]

***Abstract*— To enhance successful cooperation in human-robot collaboration tasks, many factors have to be considered. We assume, that for specific levels of task complexity, there is always one complementing input modality which increases the corresponding user satisfaction and performance. In order to identify the ideal mix of these elements, we present two experiments in this paper. The first study was in a public space and the second in a controlled laboratory environment. Besides investigating the correlation of task complexity and input modality, we explored, whether the appearance of the robot also has an impact, within the lab environment. We identified strong interdependencies between task complexity and input modalities. Specifically with hard tasks, differences in performance and satisfaction were often highly significant. Additionally we found, that the perceived task complexity was strongly dependent on the cognitive workload, driven by the used input modality, which also emphasized the strong coherency of these factors. Regarding the influence of the appearance of the robot, we found, that the human-like shape increases users' self confidence, to be able to solve a task without help.**

I. INTRODUCTION

Human-Robot Collaboration (HRC) has become more and more important in different contexts, such as in modern homes and hospitals, as assistive systems or in a factory context as assembly, painting, inspection and transportation robots.

From our point of view, it is important for all of these contexts to identify which input modality is most suitable for cooperation in turn-taking tasks, especially concerning different levels of task complexity and different appearances of robots. Therefore, we wanted to investigate the impact and correlation of input devices with different levels of task complexity and different robot appearances. In order to explore our assumption, we built a Lego Mindstorm robot, which can be controlled with three different input devices (a gesture-based input system on a mobile phone, a PC-remote control interface, and a speech recognition system based on Java) and studied these input modalities in two experiments. One study was a field trial in a public context; the second took place in a controlled laboratory setting.

In the following, we want to provide a short summary of related work, our experimental setups, the results of both studies, a discussion part, and finally an outlook about future work.

[1]G. Stollnberger is with the ICT&S Center, University of Salzburg, 5020 Salzburg, Austria gerald.stollnberger at sbg.ac.at

[2]A. Weiss is with the ICT&S Center, University of Salzburg, 5020 Salzburg, Austria astrid.weiss at sbg.ac.at

[3]M. Tscheligi is with the ICT&S Center, University of Salzburg, 5020 Salzburg, Austria manfred.tscheligi at sbg.ac.at

II. RELATED WORK

One of the most important aspects of successful human-robot cooperation is the input modality, which is used to control and interact with the robot.

It is often assumed, that most of the cooperation between humans and robots can be enabled by speech input. However, even with advancing speech processing systems, it is still challenging to make computers understand natural language. For example, Cantrell et al. [1] worked on the issue of teaching natural language to an artificial system. Their system was able to parse a wide variety of spoken utterances, but it still produced errors, which can decrease the performance and user satisfaction rate of the cooperation. Furthermore, Ayres and colleagues [2] also achieved good results in implementing speech commands for not fully embedded systems and robots, using Lego Mindstorms and the Java programming language. They mentioned that further implementation and testing work, especially with users, is required. This was, amongst others, a reason for us to consider simple speech commands to investigate our research assumption. Additionally, the fact that a system which can be controlled by speech, offers important advantages in the interaction, namely a high degree of freedom (both hands are free for other tasks) and a high degree of familiarity with this kind of modality (naturalness of speech). Moreover, the approach of using a small set of speech for commands for collaborating with robots seems to be very promising [1],[2], which motivated us to use a similar small set of commands in order to increase the accuracy.

However, there are other input modalities, apart from speech, to interact with robots such as gesture and keyboards, which can be suitable for specific context situations (e.g., high ambient noise). Rouanet and colleagues [3],[4] compared a keyboard-like and a gesture-based interface to control a system that was driving on two courses, differing in complexity. Results showed that different task complexities had a significant impact on the user performance and satisfaction. Whereas, the input modalities were considered rather equally satisfying and efficient. However, both the keyboard-like and the gesture based interface showed promising results, leading to our decision to further investigate these two types of input modalities additional to speech.

Moreover, the research of Rouanet et al. indicated the importance of taking different task types and complexities into account when investigating and enhancing human-robot collaboration. Thus, in our explorative studies, we wanted to manipulate the task complexity in a similar way as in [3] and we also included findings of task complexity research, how to categorize the difficulty of a cooperation task. In the work of Stork et al. [5], different types of Lego tasks were categorized into different complexity levels.

Another interesting approach is the use of multi-touch interfaces for commanding robots. According to Hayes et al. [6], it is not satisfying to command robots by static input devices, such as a mouse or a keyboard, above all, in situations where the user is on the move, or needs the hands free for a secondary or even for the primary task. They could show, that multi-touch interfaces provide a good usability and a lower overall workload for users controlling the robot. This was another reason for us to compare more static input modalities, like a PC-remote, to more flexible ones, like a gesture-based interface or a speech control system.

All in all, a large amount of research effort has already been undertaken within the research field of HRC to compare input modalities. A lot of research has already been done to analyze the advantages and disadvantages of different modalities, sometimes even considering task-specific differences.

However, what we believe to be missing so far, is a detailed analysis of the correlation between: (1) input modality, (2) task complexity, and (3) robot appearance (functional vs. human-like). Is there an ideal mix of input device and appearance of robots for different levels of task complexity, in terms of the resulting user satisfaction and the overall performance?

To answer this question, we conducted two user studies, (1) one in a public setting and (2) one in a controlled laboratory setting. In order to take the different levels of task complexities into account, we used the findings of task complexity research with regard to assembly tasks by Stork et al. [5] in our second experiment.

Furthermore, we have come to the conclusion, that the impact of different appearances of robots should also not be underestimated, such as in the work of Groom et al. [7], as different tasks could be considered as more suitable for a human-like than a machine-like robot.

III. Experimental Setup

For our experiments, we built two identical robots, driven by The Snatcher from Laurens [8], which were able to drive around and to pick up and transport a box.

A. Preliminary Study in a Public Context

In order to gather an initial insight into the interplay of input modalities and task complexity levels, we conducted a study in a public context during a university event, where participants competed with each other in a race situation. [Fig. 1] [9]

- 24 Participants took part in a Lego Mindstorm Race (10 female, 14 male; age from 11 to 66 years; mean 36.73 years; SD: 14.63)
- The experiment was set up as a 2x2 between-subject experiment.
- Two different input devices; (1) PC-remote control and (2) a gesture-based interface, were explored.
- Two tracks which differed in the number of obstacles (Easy/Hard).
- Performance measures were efficiency (track solution time) and effectiveness (number of collisions).
- User satisfaction measures included perceived task complexity, intuitiveness, satisfaction, and acceptance gathered by questionnaires, driven from the USUS evaluation framework [10], and from a survey used to assess the intuitiveness of the Nabaztag Robot as an ambient interface [11].
- For further analysis four data records were removed because a few participants solved the track without an opponent.

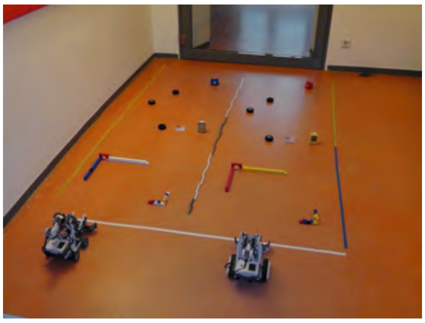

Fig. 1: The Lego Mindstorm race track (Hard) with the two robots

B. Structured Study in a Controlled Laboratory Setting

To enable a deeper investigation (after promising results from the preliminary study) of the correlation of different input modalities, task complexities, and appearances of robots, we conducted a second experiment in a highly structured and controlled laboratory setting. In addition to the input modalities in the preliminary study, we also added speech as third input modality, using Bluetooth for communicating with the robot and Java and Sphinx-4 [12] for parsing the verbal utterances.

In the second study, participants had to command the robot to transport boxes containing the parts for the assembly tasks, using three different input modalities, and solve 3 Lego building tasks differing in complexity. Altogether, 24 participants (11 female, 13 male, aged from 15 to 61 years, mean: 29.46 years; SD: 12.19) took part in this study.

The findings of Stork et al. [5] inspired us regarding the manipulation of the task complexity. Participants had to build a small Lego house, divided into three different building tasks: (1) Firstly the frame of the house, which was, according to [5], the easiest type of building task (class frame, easy), followed by (2) grouping together the frame with the prebuilt door (class group, medium). The last step was to finish the Lego house by building the roof (3), which was the most complex type of building task in our experiment (class roof, hard).

The study was set up as a 3x3x2 mixed experimental design. Task complexity and input modality was tested with a within-subject design and the factor appearance was between-subject. (Half of the participants worked with the functional robot, the others with the same robot with a human-like added feature) [Fig. 2].

All participants had to use all three input devices and to solve all of the construction tasks. In order to avoid outside factors (e.g., learning effects) which could influence our results, we varied the sequence of input devices participants had to use (e.g., speech first, then gesture, then PC remote). After each task, they had to fill in each questionnaire, with exception of the survey about the robot's appearance. This was used only once per participant, since a participant either worked with a functional or with a human-like robot.

For the purpose of investigating the effect of different appearances of robots, we used a 3D printer to produce a head, designed by Neophyte, for our robot. [1] We were interested to see, if this minimalistic human-like cue is already sufficient to impact the user satisfaction with different input modalities, depending on the task difficulty.

As performance measures, we used the track solution time and building solution time to measure efficiency and the number of collisions to measure effectiveness. Additionally, the following user satisfaction measures were gathered by the same questionnaires as in the previous study: perceived task complexity, intuitiveness, satisfaction, and acceptance. Furthermore, we were interested in the cognitive workload of the participants after using each input modality, which was assessed by the German version of the NASA RTLX [13]. About trust, participants have in each input modality, we used a questionnaire following the guidelines of McKnight et al. [14], and for data about the appearance of the robot, we used the German version of the Godspeed questionnaire. [15]

Fig. 2: The functional and the more human-like robot with a 3D printed head following the model of Neophyte.

IV. Results

A. *Results of the Preliminary Study*

We first conducted a manipulation check which revealed that the task complexity was successfully varied. Concerning a Mann-Whitney U test, there were significant differences between the easy and hard track in solution time and also in the number of collisions.

Regarding performance, the PC-remote outperformed the gesture-based interface, which was on the other hand perceived to be more intuitive. The difference was much lower for the easy track, strongly suggesting the relationship between task complexity and input modalities.

Concerning the three scales intuitiveness, acceptance, and satisfaction all of them were perceived better when solving the hard course. This initially sounds to be paradox. However, it also indicates the strong connectivity between task complexity and input modality. In our opinion, this tendency can be explained by the fact that when people manage to solve more challenging tasks, they are more satisfied with themselves and also with the system, when successfully achieving a goal. A more detailed analysis of the results can be found in our previous publication [9].

B. *Results of the Laboratory Study*

We started with a manipulation check to identify, if the modification of the task complexity, regarding the building tasks, was successful. The Kruskal-Wallis test revealed that the distribution was highly significant for the building solution time (H(2) = 36.530, p = 0.000), with a mean rank of 22.22 for class frame (easy), 32.22 for class group (medium), and 58.26 for class roof (hard) [Fig. 3]. In other words, the roof task was the most difficult one, as intended.

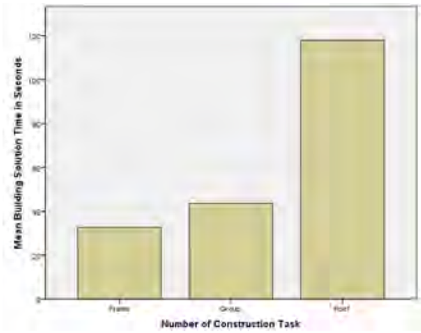

Fig. 3: The mean time needed for accomplishing the building tasks in seconds

Furthermore, the results of the perceived building complexity were significant (H(2) = 9.188, p = 0.010), with a mean rank of 46.34 for class frame, 36.78 for class group, and 30.88 for class roof.

The test also revealed that the distribution of the item "the used input device was fast to learn" offered significant results (H(2) = 8.166, p = 0.017), with a mean rank of 47.44 for class frame, 33.66 for class group, and 32.90 for class roof. It could be interpreted that people perceived the input devices to be easier to learn, as they accomplished an easier

[1] http://www.thingiverse.com/thing:8075

building task. This also emphasizes the relationship between task complexity and used input device.

In addition, the physical demand gathered by the NASA RTLX was significant (H(2) = 6.044, p = 0.49), with a mean rank of 30.90 for class frame, 45.38 for class group, and 37.72 for class roof.

These facts indicate:

- That the building complexity was successfully varied,
- The results of Stork et al. [5], which were used for our manipulation, could be reproduced.

After the manipulation check, we computed the scales of the questionnaires for the items intuitiveness, satisfaction, overall acceptance, trust, and its subcategories reliability and functionality, cognitive workload, and robot appearance. Therefore, we ran a Cronbach's Alpha test to check the internal reliability of all these factors.

For the intuitiveness scale, which consisted of 5 questions, we reached a Cronbach's Alpha of 0.792.

The Alpha value for satisfaction, with 4 questions, was 0.912 and for acceptance, which consisted of 5 items, 0.639.

For the complete trust questionnaire of 7 questions, the value was 0.955, for the subcategories reliability (4 items) 0.948, and functionality (3 items) 0.904.

The cognitive workload, consisting of 6 items, scored an Alpha value of 0.733, and, last but not least, for the appearance questionnaire, the result was 0.745.

For all of our scales, no deletion of items would have improved the Alpha values.

As expected, we were able to reproduce the findings of the preliminary study concerning the performance. The mean number of collisions, when using the PC-remote, was 0.16 (SD: 0.374), for the gesture-based interface 0.24 (SD: 0.723), and for the speech control 2.10 (SD: 1.294). [TABLE I]

TABLE I: Performance measures of the input modalities

	Used Device					
	PC Control		Gesture-based		Speech	
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Track Solution Time in Seconds	55.20	9.10	79.96	19.19	145.65	61.58
Number of Collisions	0.16	0.37	0.24	0.72	2.10	1.29

A Kruskal-Wallis test, moreover, revealed differences in the perception of the control complexity (H(2) = 23.060, p = 0.000), the mean rank was 47.52 for the PC-remote control, followed by 40.50 for the gesture-based interface and 21.76 for the speech control.

Similarly, we could identify interesting differences in user satisfaction , depending on the used input device:

- For the intuitiveness scale (H(2) = 25.238, p = 0.000), the average rank for the PC-remote was 52.74, followed by 38.76 for the gesture-based interface and 22.50 for the speech control. This is a contradiction to the preliminary study [9], where the gesture-based interface was considered more intuitive than the PC-remote.
- For the satisfaction scale (H(2) = 31.947, p = 0.000), the average rank for the PC-remote was 52.66, followed by 42.44 for the gesture-based interface and 18.90 for the speech control.
- For the acceptance scale (H(2) = 16.467, p = 0.000), the average rank for the PC-remote was 48.74, followed by 40.64 for the gesture-based interface and 24.62 for the speech control.
- For the trust scale (H(2) = 45.001, p = 0.000), the average rank for the PC-remote was 55.08, followed by 43.78 for the gesture-based interface and 15.14 for the speech control.
- For the reliability scale (H(2) = 47.850, p = 0.000), the average rank for the PC-remote was 55.90, followed by 43.46 for the gesture-based interface and 14.64 for the speech control.
- For the functionality scale (H(2) = 34.895, p = 0.000), the average rank for the PC-remote was 51.46, followed by 44.66 for the gesture-based interface and 17.88 for the speech control.

The above mentioned scales, used a five point Likert-scale from 1 (totally disagree) to 5 (totally agree). For the cognitive workload, we used a scale between 1 (very low) and 20 (very high) which also revealed a highly significant difference concerning the used input modality.

- For the cognitive workload scale (H(2) = 17.924, p = 0.000), the average rank for the PC-remote was 26.36, followed by 35.44 for the gesture-based interface and 52.10 for the speech control.

Interestingly, even the small change of the appearance of the robot by means of the 3D printed head impacted the results, even though as expected the robot with the head was not considered much more human-like than the functional one.

A Mann-Whitney U test showed, that for the single item "I would not be able to solve a task with the robot without help" more people tended to disagree when using the more human-like robot (z = 2.054, p: 0.040). The mean rank for the robot with head was 42.04, while the average rank for the functional one was 33.62. In other words, participants who collaborated with the more human-like robot were more self-confident in being able to solve a task on their own. Thus, we assume that even a minimalistic human-like appearance increases participants' positive experience while collaborating with a robot.

We further found out that participants, especially after using the speech control, which caused the highest mental workload, perceived the building task more complex, although the building task was independent from the input modality. (People had always to control the robot for getting the required parts and after that to assembly the parts) [Fig.4]

In Fig. 5, it can also be seen that the item "The control device was needlessly complex" was rated nearly the same for the easiest building task (frame), but differed much more for the classes group and roof, which again demonstrates a

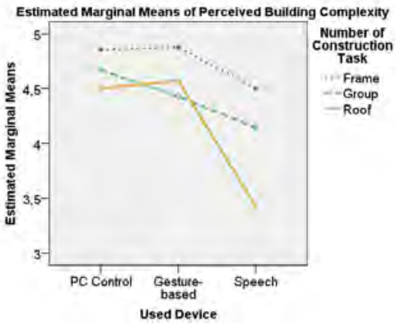

Fig. 4: The building tasks were perceived more complex, when the more complex input modality was used (5 = Easy, 1 = Hard)

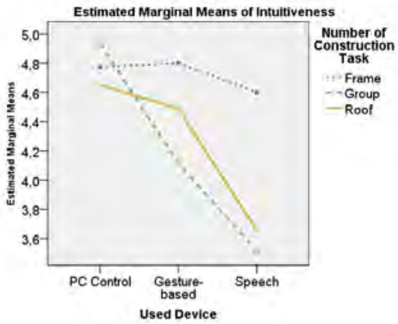

Fig. 6: Intuitiveness rated lower for complex devices and assembly tasks (5 = High, 1 = Low)

strong relationship between building complexity and input modality.

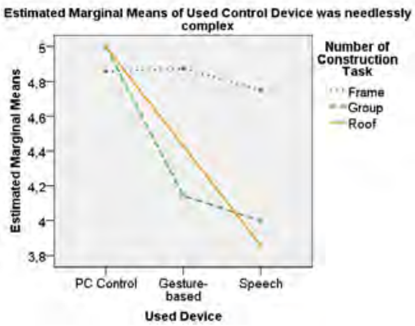

Fig. 5: Perceived control complexity in dependency of the building tasks (5 = Totally disagree, 1 = Totally agree)

In general, regarding the intuitiveness scale, for example, it could be shown that, especially, when the most complex input modality (in our case speech control) and more complex assembly tasks (roof and group) converged, the user satisfaction measures were rated much lower. [Fig. 6] This tendency could be observed for all of our user satisfaction measures and also for the cognitive workload results.

Summarizing, in performance and user satisfaction measures, the PC-remote outperformed the other two input modalities. However, the differences for simpler tasks were much lower than for hard tasks, matching the findings in the preliminary study.

The speech control was rated lowest in all categories in our study, which could be a consequence of the latency as it is often a problem of speech control systems. For example, if the participant wanted to stop the robot, it took a few milliseconds until the robot actually stopped, a fact which people had to take into account. On the other hand, the accuracy of the speech control was satisfying for most of the participants and the disadvantage driven by the latency could be compensated with little experience using the device. For simple tasks, like going forward or backward, it would absolutely suit, according to the comments of many participants.

Furthermore, some people mentioned that although the gesture-based interface did not achieve the best values in terms of performance, it was a pleasant experience to use it, which can indeed be considered as an advantage of this modality. This fact could provide a better long-term motivation and satisfaction for users. In addition, if a person's internal motivation is higher, then they are more receptive to the information [16]. As a consequence, the learning process will be faster and with a higher output.

V. CONCLUSION

In most cases, there is only one possibility to control a robot, but many different tasks to solve in collaboration. Due to the strong interdependency of input modality and task complexity, the design process for planning human-robot collaborative systems should specifically include the tasks which have to be fulfilled. As we have seen, it is not really important for easy tasks, which input modality is used and provided. The differences in performance and also in the user satisfaction rankings were small and not statistically significant. In other words the user was always satisfied when solving easy tasks, and performed them equally well, no matter which input modality was provided. However, the differences in performance and user satisfaction were larger for hard tasks and need to be taken into consideration because of their statistical significance. Our studies revealed that in challenging tasks, users preferred the PC-remote control, because it was considered the most accurate, reliable, and familiar input modality. We are aware of the limitations of our work; we can only make assumptions for the input modalities we studied and compared and that this preliminary work cannot be generalized to all variations of speech, gesture, and point-and-click input.

Therefore, our overall goal is to further explore the ideal input modality for a set of tasks, categorized according to their level of complexity. Additional to complexity, we want to take other factors into account, such as whether or not the robot is physically collocated with the user, ambient noise, light conditions and other factors such as if the robot could get out of sight during the interaction or the user needs one or both hands free for another task, like the necessity of carrying anything. Moreover, it has to be clear, if the user needs to have the possibility of mobility during the interaction. This classification will enable the possibility to provide the ideal complementing input modality, for each type of task.

In our opinion, a typical multimodal interface, which provides many different input possibilities at once, is not the ideal solution for successful human-robot cooperation. We strongly believe that a more adaptive multimodality is needed, which could only be achieved if the tasks have been well classified before.

Finally, we assume, as an initial tendency was found in our experiments, that a humanoid design of robots generates a more positive feeling in the user. Therefore, further investigation of how to produce minimal cues on the robot, to enhance the overall cooperation from the user's point of view is needed. This would be an advantage we should avail ourselves.

All in all, the two studies presented in this paper, proved our assumption about the interdependency of task complexity and input modality in HRC. For us, these studies are a starting point for a series of controlled experiments, to further decode this interdependency and propose adaptive multimodal HRC scenarios.

VI. Future Work

In a next step, we aim to investigate the possibility to reproduce our findings if participants had a longer training phase and got used to all input modalities. Therefore, we plan to give participants the robot and the input modalities for usage and training in their private home for one week, before conducting a controlled experiment in the lab again.

Moreover, we want to gain a deeper insight into the impact of different appearances; therefore, a more human-like robot (e.g., the Nao robot) should be compared to our functional Lego Mindstorm prototype. Similarly, due to the fact that no subject worked with both types of robots, in a next study, a within-subject approach for further investigation is considered.

Finally, we want to explore the concept of an "intelligent" multimodal interface, where people are free to choose which input modality they need in specific situations. By intelligent we understand, that the interface reacts also in accordance to the context and other circumstances like light, noise, and temperature and provides a proper input modality and an according feedback modality (e.g., Visual, haptic, and auditive). The resulting combinations of input modality and feedback mechanism for different task complexities should enable context-specific human-robot cooperation.

Acknowledgment

We greatly acknowledge the financial support by the Federal Ministry of Economy, Family and Youth and the National Foundation for Research, Technology and Development (Christian Doppler Laboratory for Contextual Interfaces).

References

[1] R. Cantrell, M. Scheutz, P. Schermerhorn, and X. Wu, "Robust spoken instruction understanding for hri," in *Proc. of the 5th ACM/IEEE international conf. on Human-robot interaction*, ser. HRI '10. New York, NY, USA: ACM, 2010, pp. 275–282.

[2] T. Ayres and B. Nolan, "Voice activated command and control with java-enabled speech recognition over wifi," in *Proc. of the 3rd international symposium on Principles and practice of programming in Java*, ser. PPPJ '04. Trinity College Dublin, 2004, pp. 114–119.

[3] P. Rouanet, J. Bechu, and P.-Y. Oudeyer, "A comparison of three interfaces using handheld devices to intuitively drive and show objects to a social robot: the impact of underlying metaphors," in *Robot and Human Interactive Communication, RO-MAN 2009. The 18th IEEE International Symposium on*, 27 2009-oct. 2 2009, pp. 1066 –1072.

[4] P. Rouanet, F. Danieau, and P.-Y. Oudeyer, "A robotic game to evaluate interfaces used to show and teach visual objects to a robot in real world condition," in *Proc. of the 6th international conf. on Human-robot interaction*. New York, USA: ACM, 2011, pp. 313–320.

[5] S. Stork, C. Stössel, and A. Schubö, "Optimizing human-machine interaction in manual assembly," in *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*, aug. 2008, pp. 113 –118.

[6] S. T. Hayes, E. R. Hooten, and J. A. Adams, "Multi-touch interaction for tasking robots," in *Proc. of the 5th ACM/IEEE international conf. on Human-robot interaction*. NY, USA: ACM, 2010, pp. 97–98.

[7] V. Groom, L. Takayama, P. Ochi, and C. Nass, "I am my robot: the impact of robot-building and robot form on operators," in *Proc. of the 4th ACM/IEEE international conf. on Human robot interaction*, ser. HRI '09. New York, NY, USA: ACM, 2009, pp. 31–36.

[8] L. Valk, *The LEGO MINDSTORMS NXT 2.0 Discovery Book: A Beginner's Guide to Building and Programming Robots*, 1st ed. San Francisco, CA, USA: No Starch Press, 2010.

[9] G. Stollnberger, A. Weiss, and M. Tscheligi, "Input modality and task complexity: Do they relate?" in *HRI 13: Proc. of the 8th ACM/IEEE International Conf. on Human Robot Interaction*, 2013.

[10] A. Weiss, "Validation of an evaluation framework for human-robot interaction. the impact of usability, social acceptance, user experience, and societal impact on collaboration with humanoid robots," *Unpublished doctoral dissertation, University of Salzburg*, 2010.

[11] T. Mirlacher, R. Buchner, F. Förster, A. Weiss, and M. Tscheligi, "Ambient rabbits likeability of embodied ambient displays," in *Ambient Intelligence*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2009, vol. 5859, pp. 164–173.

[12] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: a flexible open source framework for speech recognition," Mountain View, CA, USA, Tech. Rep., 2004.

[13] J. P. Wölber, "Evaluation zweier computergesttzter lernprogramme in der parodontologie," Dissertation, Albert-Ludwigs-Universitt, 2010.

[14] D. H. Mcknight, M. Carter, J. B. Thatcher, and P. F. Clay, "Trust in a specific technology: An investigation of its components and measures," *ACM Trans. Manage. Inf. Syst.*, July 2011.

[15] C. Bartneck, E. Croft, and D. Kulic, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *Int. Journal of Social Robotics*, 2009.

[16] A. L. Brown, "Motivation to learn and understand: On taking charge of one's own learning," *Cognition and Instruction*, 1988.

A.12. Output of the Data Analysis of the Laboratory Study

```
* All Variables

DESCRIPTIVES VARIABLES=TP_NR Gender Age CTask Device TSolTime CSolTime NumOfColl Distraction
    PerceivedTComplexity PerceivedBComplexity IComplexity IComprehensible ILearn IHLearn IHard SUse
    SatRobot SatDevice SatSelf ASkill AKnowledge AHelp APressure ANever TFailSafeR TSkillF TReliableR
    TFunctionalityF TWorksR TFeaturesF TAbandonR NRTLXMentalDemand NRTLXPhysicalDemand
    NRTLXTimePressure NRTLXPerformPressure NRTLXEffort NRTLXFrustration RobotHead E1 E2 E3 E4 E5
  /STATISTICS=MEAN STDDEV MIN MAX.
```

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Number of Participant	72	1	25	13,00	7,260
Gender	72	1	2	1,44	,500
Age of Participant	69	15	61	29,46	12,192
Number of Construction Task	72	1	3	2,00	,822
Used Device	72	1	3	2,00	,822
Track Solution Time in Seconds	65	42	300	89,89	50,736
Building Solution Time in Seconds	71	17	470	63,33	70,306
Number of Collisions	68	0	5	,74	1,200
IAT Measurement	0				
Perceived Control Complexity	70	1	5	4,18	1,147
Perceived Building Complexity	72	1	5	4,44	,919
Used Control Device was needlessly complex	71	2	5	4,49	,880
Used Control Device was easy comprehensible	72	1	5	4,39	1,126
Used Control Device was fast to learn	72	1	5	4,11	1,110
Used Control Device needs much to learn	72	2	5	4,61	,715
Used Control Device was hard to use	72	1	5	4,04	1,191
Like to use device often	72	1	5	3,55	1,369
Satisfied with rob ots' performance	71	1	5	3,81	1,213
Satisfied with devices' performance	71	1	5	3,64	1,381
Satisfied with own performance	68	1	5	3,93	1,026
Own required skill for commanding robots with device	72	1	5	4,37	1,075
Own required knowledge for commanding robots with device	72	1	5	4,27	1,166
Would not be able to solve a task without help	72	1	5	4,49	,891
Would not be able to solve a task under time pressure	72	1	5	4,16	1,066
Would never be able to solve a task	72	2	5	4,72	,648
The System is very FailSafe	72	1	5	3,72	1,157
The System has the ability to handle what I want	72	2	5	4,04	1,032
The System is very reliable	72	1	5	3,44	1,233
The System has the required functionality	72	1	5	4,24	,984
The System works for me	72	1	5	4,12	1,127
The System provides the Features I need for solving Tasks	72	1	5	4,27	1,004
The System does not abandon me	72	1	5	3,64	1,237

Mental Demand	72	1	17	6,89	4,492
Physical Demand	72	1	9	2,45	1,803
Time Pressure	72	1	14	3,39	2,837
Performance Pressure	72	1	20	8,43	5,985
Effort needed	72	1	20	9,51	5,999
Frustration	72	1	18	4,80	4,662
Existence of a Head	72	0	1	,52	,503
Unecht - Natürlich	24	1	4	2,20	1,080
Wie eine Maschine - Wie ein Mensch	24	1	3	1,40	,645
Hat kein Bewusstsein - Hat ein Bewusstsein	24	1	3	1,28	,614
Künstlich - Realistisch	24	1	4	1,72	1,021
Bewegt sich steif - Bewegt sich flüssig	24	1	4	2,00	,913
Valid N (listwise)	0				

```
* Custom Tables grouped by Number of Construction Task

CTABLES
  /VLABELS VARIABLES=Age TSolTime NumOfColl CSolTime PerceivedTComplexity PerceivedBComplexity
    IComplexity IComprehensible ILearn IHLearn IHard SUse SatRobot SatDevice SatSelf ASkill AKnowledge
    AHelp APressure ANever TFailSafeR TSkillF TReliableR TFunctionalityF TWorksR TFeaturesF TAbandonR
    NRTLXMentalDemand NRTLXPhysicalDemand NRTLXTimePressure NRTLXPerformPressure NRTLXEffort
    NRTLXFrustration RobotHead E1 E2 E3 E4 E5 CTask
    DISPLAY=LABEL
  /TABLE Age [S][MEAN] + TSolTime [S][MEAN] + NumOfColl [S][MEAN] + CSolTime [S][MEAN] +
    PerceivedTComplexity [S][MEAN] + PerceivedBComplexity [S][MEAN] + IComplexity [S][MEAN] +
    IComprehensible [S][MEAN] + ILearn [S][MEAN] + IHLearn [S][MEAN] + IHard [S][MEAN] + SUse [S][MEAN]
    + SatRobot [S][MEAN] + SatDevice [S][MEAN] + SatSelf [S][MEAN] + ASkill [S][MEAN] + AKnowledge
    [S][MEAN] + AHelp [S][MEAN] + APressure [S][MEAN] + ANever [S][MEAN] + TFailSafeR [S][MEAN] +
    TSkillF [S][MEAN] + TReliableR [S][MEAN] + TFunctionalityF [S][MEAN] + TWorksR [S][MEAN] +
    TFeaturesF [S][MEAN] + TAbandonR [S][MEAN] + NRTLXMentalDemand [S][MEAN] + NRTLXPhysicalDemand
    [S][MEAN] + NRTLXTimePressure [S][MEAN] + NRTLXPerformPressure [S][MEAN] + NRTLXEffort [S][MEAN] +
    NRTLXFrustration [S][MEAN] + RobotHead [S][MEAN] + E1 [S][MEAN] + E2 [S][MEAN] + E3 [S][MEAN] + E4
    [S][MEAN] + E5 [S][MEAN] BY CTask [C]
  /CATEGORIES VARIABLES=CTask ORDER=A KEY=VALUE EMPTY=INCLUDE.
```

Custom Tables

	Number of Construction Task		
	Frame	Group	Roof
	Mean	Mean	Mean
Age of Participant	29	29	29
Track Solution Time in Seconds	94	86	91
Number of Collisions	1	1	1
Building Solution Time in Seconds	33	44	118
Perceived Control Complexity	4	4	4
Perceived Building Complexity	5	4	4
Used Control Device was needlessly complex	5	4	4
Used Control Device was easy comprehensible	5	4	4
Used Control Device was fast to learn	5	4	4
Used Control Device needs much to learn	5	4	5
Used Control Device was hard to use	4	4	4
Like to use device often	4	3	4
Satisfied with rob ots' performance	4	4	4
Satisfied with devices' performance	4	4	4
Satisfied with own performance	4	4	4

Own required skill for commanding robots with device	5	4	4
Own required knowledge for commanding robots with device	4	4	4
Would not be able to solve a task without help	4	4	5
Would not be able to solve a task under time pressure	4	4	4
Would never be able to solve a task	5	5	5
The System is very FailSafe	4	4	4
The System has the ability to handle what I want	4	4	4
The System is very reliable	3	4	3
The System has the required functionality	4	4	4
The System works for me	4	4	4
The System provides the Features I need for solving Tasks	4	4	4
The System does not abandon me	4	4	4
Mental Demand	6	8	7
Physical Demand	2	3	2
Time Pressure	2	4	4
Performance Pressure	8	9	9
Effort needed	9	10	10
Frustration	4	6	5
Existence of a Head	1	1	1
Unecht - Natürlich	.	.	2
Wie eine Maschine - Wie ein Mensch	.	.	1
Hat kein Bewusstsein - Hat ein Bewusstsein	.	.	1
Künstlich - Realistisch	.	.	2
Bewegt sich steif - Bewegt sich flüssig	.	.	2

```
* Custom Tables grouped by Used Device

CTABLES
  /VLABELS VARIABLES=Age TSolTime NumOfColl CSolTime PerceivedTComplexity PerceivedBComplexity
    IComplexity IComprehensible ILearn IHLearn IHard SUse SatRobot SatDevice SatSelf ASkill AKnowledge
    AHelp APressure ANever TFailSafeR TSkillF TReliableR TFunctionalityF TWorksR TFeaturesF TAbandonR
    NRTLXMentalDemand NRTLXPhysicalDemand NRTLXTimePressure NRTLXPerformPressure NRTLXEffort
    NRTLXFrustration RobotHead E1 E2 E3 E4 E5 Device
    DISPLAY=LABEL
  /TABLE Age [S][MEAN] + TSolTime [S][MEAN] + NumOfColl [S][MEAN] + CSolTime [S][MEAN] +
    PerceivedTComplexity [S][MEAN] + PerceivedBComplexity [S][MEAN] + IComplexity [S][MEAN] +
    IComprehensible [S][MEAN] + ILearn [S][MEAN] + IHLearn [S][MEAN] + IHard [S][MEAN] + SUse [S][MEAN]
    + SatRobot [S][MEAN] + SatDevice [S][MEAN] + SatSelf [S][MEAN] + ASkill [S][MEAN] + AKnowledge
    [S][MEAN] + AHelp [S][MEAN] + APressure [S][MEAN] + ANever [S][MEAN] + TFailSafeR [S][MEAN] +
    TSkillF [S][MEAN] + TReliableR [S][MEAN] + TFunctionalityF [S][MEAN] + TWorksR [S][MEAN] +
    TFeaturesF [S][MEAN] + TAbandonR [S][MEAN] + NRTLXMentalDemand [S][MEAN] + NRTLXPhysicalDemand
    [S][MEAN] + NRTLXTimePressure [S][MEAN] + NRTLXPerformPressure [S][MEAN] + NRTLXEffort [S][MEAN] +
    NRTLXFrustration [S][MEAN] + RobotHead [S][MEAN] + E1 [S][MEAN] + E2 [S][MEAN] + E3 [S][MEAN] + E4
    [S][MEAN] + E5 [S][MEAN] BY Device
  /CATEGORIES VARIABLES=Device ORDER=A KEY=VALUE EMPTY=INCLUDE.
```

Custom Tables

	Used Device		
	PC Control	Gesture-based	Speech
	Mean	Mean	Mean
Age of Participant	29	29	29
Track Solution Time in Seconds	55	80	146
Number of Collisions	0	0	2

Building Solution Time in Seconds	64	47	78
Perceived Control Complexity	5	4	3
Perceived Building Complexity	5	5	4
Used Control Device was needlessly complex	5	4	4
Used Control Device was easy comprehensible	5	5	4
Used Control Device was fast to learn	5	4	4
Used Control Device needs much to learn	5	5	4
Used Control Device was hard to use	5	4	3
Like to use device often	4	4	3
Satisfied with rob ots' performance	4	4	3
Satisfied with devices' performance	5	4	2
Satisfied with own performance	5	4	3
Own required skill for commanding robots with device	5	4	4
Own required knowledge for commanding robots with device	5	4	4
Would not be able to solve a task without help	5	5	4
Would not be able to solve a task under time pressure	5	4	3
Would never be able to solve a task	5	5	4
The System is very FailSafe	5	4	3
The System has the ability to handle what I want	5	4	3
The System is very reliable	4	4	2
The System has the required functionality	5	5	3
The System works for me	5	5	3
The System provides the Features I need for solving Tasks	5	5	3
The System does not abandon me	5	4	2
Mental Demand	5	6	9
Physical Demand	2	3	3
Time Pressure	2	3	5
Performance Pressure	7	8	10
Effort needed	8	9	12
Frustration	2	4	9
Existence of a Head	1	1	1
Unecht - Natürlich	2	2	2
Wie eine Maschine - Wie ein Mensch	1	1	2
Hat kein Bewusstsein - Hat ein Bewusstsein	1	1	2
Künstlich - Realistisch	2	2	2
Bewegt sich steif - Bewegt sich flüssig	2	2	2

```
* Custom Tables grouped by Gender.
CTABLES
  /VLABELS VARIABLES=TSolTime CSolTime NumOfColl PerceivedTComplexity PerceivedBComplexity
    IComplexity IComprehensible ILearn IHLearn IHard SUse SatRobot SatDevice SatSelf ASkill AKnowledge
    AHelp APressure ANever TFailSafeR TSkillF TReliableR TFunctionalityF TWorksR TFeaturesF TAbandonR
    NRTLXMentalDemand NRTLXPhysicalDemand NRTLXTimePressure NRTLXPerformPressure NRTLXEffort
    NRTLXFrustration E1 E2 E3 E4 E5 Intuitiveness Satisfaction Acceptance Trust Reliability
    Functionality CognitiveWorkload Embodiment Gender
    DISPLAY=LABEL
  /TABLE TSolTime [S][MEAN] + CSolTime [S][MEAN] + NumOfColl [S][MEAN] + PerceivedTComplexity
    [S][MEAN] + PerceivedBComplexity [S][MEAN] + IComplexity [S][MEAN] + IComprehensible [S][MEAN] +
    ILearn [S][MEAN] + IHLearn [S][MEAN] + IHard [S][MEAN] + SUse [S][MEAN] + SatRobot [S][MEAN] +
    SatDevice [S][MEAN] + SatSelf [S][MEAN] + ASkill [S][MEAN] + AKnowledge [S][MEAN] + AHelp [S][MEAN]
    + APressure [S][MEAN] + ANever [S][MEAN] + TFailSafeR [S][MEAN] + TSkillF [S][MEAN] + TReliableR
    [S][MEAN] + TFunctionalityF [S][MEAN] + TWorksR [S][MEAN] + TFeaturesF [S][MEAN] + TAbandonR
    [S][MEAN] + NRTLXMentalDemand [S][MEAN] + NRTLXPhysicalDemand [S][MEAN] + NRTLXTimePressure
    [S][MEAN] + NRTLXPerformPressure [S][MEAN] + NRTLXEffort [S][MEAN] + NRTLXFrustration [S][MEAN] +
    E1 [S][MEAN] + E2 [S][MEAN] + E3 [S][MEAN] + E4 [S][MEAN] + E5 [S][MEAN] BY Gender
  /CATEGORIES VARIABLES=Gender ORDER=A KEY=VALUE EMPTY=INCLUDE.
```

Custom Tables

	Gender	
	Male	Female
	Mean	Mean
Track Solution Time in Seconds	79	103
Building Solution Time in Seconds	51	79
Number of Collisions	1	1
Perceived Control Complexity	4	4
Perceived Building Complexity	4	4
Used Control Device was needlessly complex	4	5
Used Control Device was easy comprehensible	4	5
Used Control Device was fast to learn	4	4
Used Control Device needs much to learn	5	5
Used Control Device was hard to use	4	4
Like to use device often	3	4
Satisfied with rob ots' performance	4	4
Satisfied with devices' performance	4	4
Satisfied with own performance	4	4
Own required skill for commanding robots with device	4	4
Own required knowledge for commanding robots with device	4	4
Would not be able to solve a task without help	5	4
Would not be able to solve a task under time pressure	4	4
Would never be able to solve a task	5	5
The System is very FailSafe	4	4
The System has the ability to handle what I want	4	4
The System is very reliable	3	4
The System has the required functionality	4	4
The System works for me	4	4
The System provides the Features I need for solving Tasks	4	4

The System does not abandon me	4	4
Mental Demand	5	9
Physical Demand	2	3
Time Pressure	3	4
Performance Pressure	8	9
Effort needed	8	12
Frustration	5	5
Unecht - Natürlich	2	3
Wie eine Maschine - Wie ein Mensch	1	2
Hat kein Bewusstsein - Hat ein Bewusstsein	1	1
Künstlich - Realistisch	1	2
Bewegt sich steif - Bewegt sich flüssig	2	2

```
*Cronbach's Alpha Intuitiveness

RELIABILITY
  /VARIABLES=IComplexity IComprehensible ILearn IHLearn IHard
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /SUMMARY=TOTAL.
```

Reliability

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	71	98,7
	Excluded[a]	1	1,3
	Total	72	100,0

Reliability Statistics

Cronbach's Alpha	N of Items
,792	5

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Used Control Device was needlessly complex	17,15	10,539	,531	,766
Used Control Device was easy comprehensible	17,24	9,830	,455	,794
Used Control Device was fast to learn	17,53	8,444	,716	,700
Used Control Device needs much to learn	17,03	10,794	,643	,748
Used Control Device was hard to use	17,59	8,683	,598	,747

```
*Cronbach's Alpha Satisfaction

RELIABILITY
  /VARIABLES=SUse SatRobot SatDevice SatSelf
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
 /SUMMARY=TOTAL.
```

Reliability

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	67	93,3
	Excluded[a]	5	6,7
	Total	72	100,0

Reliability Statistics

Cronbach's Alpha	N of Items
,912	4

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Like to use device often	11,39	10,907	,824	,879
Satisfied with rob ots' performance	11,14	11,892	,820	,879
Satisfied with devices' performance	11,29	10,555	,874	,859
Satisfied with own performance	11,01	13,898	,713	,918

```
*Cronbach's Alpha Acceptance

RELIABILITY
  /VARIABLES=ASkill AKnowledge AHelp APressure ANever
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /SUMMARY=TOTAL.
```

Reliability

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	72	100,0
	Excluded[a]	0	,0
	Total	72	100,0

Reliability Statistics

Cronbach's Alpha	N of Items
,639	5

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Own required skill for commanding robots with device	17,64	6,963	,324	,623
Own required knowledge for commanding robots with device	17,75	6,705	,314	,636
Would not be able to solve a task without help	17,52	7,442	,354	,604
Would not be able to solve a task under time pressure	17,85	5,748	,602	,468
Would never be able to solve a task	17,29	7,886	,454	,582

```
*Cronbach's Alpha Trust Complete

RELIABILITY
  /VARIABLES=TFailSafeR TSkillF TReliableR TFunctionalityF TWorksR TFeaturesF TAbandonR
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /SUMMARY=TOTAL.
```

Reliability

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	72	100,0
	Excluded[a]	0	,0
	Total	72	100,0

Reliability Statistics

Cronbach's Alpha	N of Items
,955	7

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
The System is very FailSafe	23,75	35,543	,800	,951
The System has the ability to handle what I want	23,43	36,410	,839	,948
The System is very reliable	24,03	33,810	,879	,945
The System has the required functionality	23,23	37,664	,769	,954
The System works for me	23,35	34,716	,900	,943
The System provides the Features I need for solving Tasks	23,20	36,757	,834	,949
The System does not abandon me	23,83	33,443	,905	,943

```
*Cronbach's Alpha Reliability (Sub from Trust)

RELIABILITY
  /VARIABLES=TFailSafeR TReliableR TWorksR TAbandonR
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /SUMMARY=TOTAL.
```

Reliability

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	72	100,0
	Excluded[a]	0	,0
	Total	72	100,0

Reliability Statistics

Cronbach's Alpha	N of Items
,948	4

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
The System is very FailSafe	11,20	11,595	,844	,941
The System is very reliable	11,48	10,631	,925	,916
The System works for me	10,80	11,946	,818	,949
The System does not abandon me	11,28	10,664	,915	,919

```
*Cronbach's Alpha Functionality (Sub from Trust)

RELIABILITY
  /VARIABLES=TSkillF TFunctionalityF TFeaturesF
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /SUMMARY=TOTAL.
```

Reliability

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	72	100,0
	Excluded[a]	0	,0
	Total	72	100,0

Reliability Statistics

Cronbach's Alpha	N of Items
,904	3

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
The System has the ability to handle what I want	8,51	3,470	,811	,860
The System has the required functionality	8,31	3,567	,840	,837
The System provides the Features I need for solving Tasks	8,28	3,664	,776	,889

```
*Cronbach's Alpha NASA RTLX

RELIABILITY
  /VARIABLES=NRTLXMentalDemand NRTLXPhysicalDemand NRTLXTimePressure NRTLXPerformPressure
    NRTLXEffort NRTLXFrustration
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /SUMMARY=TOTAL.
```

Reliability

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	72	100,0
	Excluded[a]	0	,0
	Total	72	100,0

Reliability Statistics

Cronbach's Alpha	N of Items
,733	6

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Mental Demand	28,57	219,788	,612	,654
Physical Demand	33,01	295,635	,365	,735
Time Pressure	32,08	267,507	,495	,703
Performance Pressure	27,04	209,039	,443	,717
Effort needed	25,96	188,390	,590	,659
Frustration	30,67	230,631	,488	,690

```
*Cronbach's Alpha Embodiment

RELIABILITY
  /VARIABLES=E1 E2 E3 E4 E5
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /SUMMARY=TOTAL.
```

Reliability

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	24	33,3
	Excluded[a]	48	66,7
	Total	72	100,0

Reliability Statistics

Cronbach's Alpha	N of Items
,745	5

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Unecht - Natürlich	6,40	5,333	,601	,666
Wie eine Maschine - Wie ein Mensch	7,20	6,917	,638	,674
Hat kein Bewusstsein - Hat ein Bewusstsein	7,32	8,893	,063	,815
Künstlich - Realistisch	6,88	5,193	,701	,616
Bewegt sich steif - Bewegt sich flüssig	6,60	6,000	,596	,666

```
*Compute Variable Intuitiveness

COMPUTE Intuitiveness=MEAN(IComplexity,IComprehensible,ILearn,IHLearn,IHard).
EXECUTE.

*Compute Variable Satisfaction

COMPUTE Satisfaction=MEAN(SUse,SatRobot,SatDevice,SatSelf).
EXECUTE.

*Compute Variable Acceptance

COMPUTE Acceptance=MEAN(ASkill,AKnowledge,AHelp,APressure,ANever).
EXECUTE.

*Compute Variable Trust

COMPUTE Trust=MEAN(TFailSafeR,TSkillF,TReliableR,TFunctionalityF,TWorksR,TFeaturesF,TAbandonR).
EXECUTE.

*Compute Variable Reliability (Sub from Trust)

COMPUTE Reliability=MEAN(TFailSafeR,TReliableR,TWorksR,TAbandonR).
EXECUTE.
```

```
*Compute Variable Functionality (Sub from Trust)

COMPUTE Functionality=MEAN(TSkillF,TFunctionalityF,TFeaturesF).
EXECUTE.

*Compute Variable Cognitive Workload

COMPUTE CognitiveWorkload=MEAN(NRTLXMentalDemand,NRTLXPhysicalDemand,NRTLXTimePressure,
    NRTLXPerformPressure,NRTLXEffort,NRTLXFrustration).
EXECUTE.

*Compute Variable Embodiment

COMPUTE Embodiment=MEAN(E1,E2,E3,E4,E5).
EXECUTE.

* Custom Tables of the scales grouped by construction task.
CTABLES
  /VLABELS VARIABLES=Intuitiveness Satisfaction Acceptance Trust Reliability Functionality
    CognitiveWorkload Embodiment CTask
    DISPLAY=LABEL
  /TABLE Intuitiveness [MEAN] + Satisfaction [MEAN] + Acceptance [MEAN] + Trust [MEAN] +
    Reliability [MEAN] + Functionality [MEAN] + CognitiveWorkload [MEAN] + Embodiment [MEAN] BY CTask
  /CATEGORIES VARIABLES=CTask ORDER=A KEY=VALUE EMPTY=INCLUDE.
```

Custom Tables

	Number of Construction Task		
	Frame	Group	Roof
	Mean	Mean	Mean
Intuitiveness	4,61	4,18	4,18
Satisfaction	3,69	3,65	3,79
Acceptance	4,54	4,26	4,42
Trust	3,90	3,90	3,98
Reliability	3,62	3,81	3,76
Functionality	4,27	4,01	4,27
CognitiveWorkload	5,26	6,38	6,09
Embodiment	.	.	1,72

```
* Custom Tables of the scales grouped by used input device.
CTABLES
  /VLABELS VARIABLES=Intuitiveness Satisfaction Acceptance Trust Reliability Functionality
    CognitiveWorkload Embodiment Device
    DISPLAY=LABEL
  /TABLE Intuitiveness [S][MEAN] + Satisfaction [S][MEAN] + Acceptance [S][MEAN] + Trust [S][MEAN]
    + Reliability [S][MEAN] + Functionality [S][MEAN] + CognitiveWorkload [S][MEAN] + Embodiment
    [S][MEAN] BY Device
  /CATEGORIES VARIABLES=Device ORDER=A KEY=VALUE EMPTY=INCLUDE.
```

Custom Tables

	Used Device		
	PC Control	Gesture-based	Speech
	Mean	Mean	Mean
Intuitiveness	4,78	4,42	3,78
Satisfaction	4,44	3,98	2,70
Acceptance	4,70	4,50	4,01
Trust	4,63	4,31	2,83
Reliability	4,58	4,13	2,48
Functionality	4,69	4,55	3,31
CognitiveWorkload	4,34	5,53	7,87
Embodiment	1,78	1,69	1,70

```
* Custom Tables of the scales grouped by existence of head.
CTABLES
  /VLABELS VARIABLES=Intuitiveness Satisfaction Acceptance Trust Reliability Functionality
    CognitiveWorkload Embodiment RobotHead
    DISPLAY=LABEL
  /TABLE Intuitiveness [S][MEAN] + Satisfaction [S][MEAN] + Acceptance [S][MEAN] + Trust [S][MEAN]
    + Reliability [S][MEAN] + Functionality [S][MEAN] + CognitiveWorkload [S][MEAN] + Embodiment
    [S][MEAN] BY RobotHead
  /CATEGORIES VARIABLES=RobotHead ORDER=A KEY=VALUE EMPTY=INCLUDE.
```

Custom Tables

	Existence of a Head	
	No	Yes
	Mean	Mean
Intuitiveness	4,33	4,32
Satisfaction	3,59	3,81
Acceptance	4,39	4,41
Trust	3,71	4,12
Reliability	3,53	3,92
Functionality	3,94	4,40
CognitiveWorkload	5,62	6,18
Embodiment	1,75	1,69

```
* Custom Tables of the scales grouped by gender.
CTABLES
  /VLABELS VARIABLES=Intuitiveness Satisfaction Acceptance Trust Reliability Functionality
    CognitiveWorkload Embodiment RobotHead
    DISPLAY=LABEL
  /TABLE Intuitiveness [S][MEAN] + Satisfaction [S][MEAN] + Acceptance [S][MEAN] + Trust [S][MEAN]
    + Reliability [S][MEAN] + Functionality [S][MEAN] + CognitiveWorkload [S][MEAN] + Embodiment
    [S][MEAN] BY gender
  /CATEGORIES VARIABLES=RobotHead ORDER=A KEY=VALUE EMPTY=INCLUDE.
```

Custom Tables

	Gender	
	Male	Female
	Mean	Mean
Intuitiveness	4,20	4,48
Satisfaction	3,59	3,86
Acceptance	4,51	4,26
Trust	3,94	3,90
Reliability	3,68	3,79
Functionality	4,29	4,05
CognitiveWorkload	5,08	6,97
Embodiment	1,46	2,05

```
*Nonparametric Tests: Independent Samples grouped by Device.
NPTESTS
  /INDEPENDENT TEST (TSolTime CSolTime NumOfColl Distraction PerceivedTComplexity
    PerceivedBComplexity IComplexity IComprehensible ILearn IHLearn IHard SUse SatRobot SatDevice
    SatSelf ASkill AKnowledge AHelp APressure ANever TFailSafeR TSkillF TReliableR TFunctionalityF
    TWorksR TFeaturesF TAbandonR NRTLXMentalDemand NRTLXPhysicalDemand NRTLXTimePressure
    NRTLXPerformPressure NRTLXEffort NRTLXFrustration E1 E2 E3 E4 E5 Intuitiveness Satisfaction
    Acceptance Trust Reliability Functionality CognitiveWorkload Embodiment) GROUP (Device)
  /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
  /CRITERIA ALPHA=0.05  CILEVEL=95.
```

Nonparametric Tests

[DataSet1] C:\Users\Stollnberger\Desktop\FinaleStudie\Auswertung\DatenAuswertung.sav

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Track Solution Time in Seconds is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
2	The distribution of Building Solution Time in Seconds is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.482	Retain the null hypothesis.
3	The distribution of Number of Collisions is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
4	The distribution of IAT Measurement is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.	Unable to compute.
5	The distribution of Perceived Control Complexity is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
6	The distribution of Perceived Building Complexity is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.093	Retain the null hypothesis.
7	The distribution of Used Control Device was needlessly complex is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.005	Reject the null hypothesis.
8	The distribution of Used Control Device was easy comprehensible is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.025	Reject the null hypothesis.
9	The distribution of Used Control Device was fast to learn is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
10	The distribution of Used Control Device needs much to learn is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.002	Reject the null hypothesis.
11	The distribution of Used Control Device was hard to use is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
12	The distribution of Like to use device often is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
13	The distribution of Satisfied with robots' performance is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
14	The distribution of Satisfied with devices' performance is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
15	The distribution of Satisfied with own performance is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
16	The distribution of Own required skill for commanding robots with device is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.328	Retain the null hypothesis.
17	The distribution of Own required knowledge for commanding robots with device is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.005	Reject the null hypothesis.
18	The distribution of Would not be able to solve a task without help is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.189	Retain the null hypothesis.
19	The distribution of Would not be able to solve a task under time pressure is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
20	The distribution of Would never be able to solve a task is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
21	The distribution of The System is very FailSafe is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
22	The distribution of The System has the ability to handle what I want is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
23	The distribution of The System is very reliable is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
24	The distribution of The System has the required functionality is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
25	The distribution of The System works for me is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
26	The distribution of The System provides the Features I need for solving Tasks is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
27	The distribution of The System does not abandon me is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
28	The distribution of Mental Demand is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.001	Reject the null hypothesis.
29	The distribution of Physical Demand is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.035	Reject the null hypothesis.
30	The distribution of Time Pressure is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.024	Reject the null hypothesis.
31	The distribution of Performance Pressure is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.120	Retain the null hypothesis.
32	The distribution of Effort needed is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.050	Retain the null hypothesis.
33	The distribution of Frustration is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
34	The distribution of Unecht - Natürlich is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.425	Retain the null hypothesis.
35	The distribution of Wie eine Maschine - Wie ein Mensch is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.109	Retain the null hypothesis.
36	The distribution of Hat kein Bewusstsein - Hat ein Bewusstsein is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.365	Retain the null hypothesis.
37	The distribution of Künstlich - Realistisch is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.911	Retain the null hypothesis.
38	The distribution of Bewegt sich steif - Bewegt sich flüssig is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.715	Retain the null hypothesis.
39	The distribution of Intuitiveness is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
40	The distribution of Satisfaction is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
41	The distribution of Acceptance is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
42	The distribution of Trust is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
43	The distribution of Reliability is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
44	The distribution of Functionality is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
45	The distribution of CognitiveWorkload is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
46	The distribution of Embodiment is the same across categories of Used Device.	Independent-Samples Kruskal-Wallis Test	.960	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

```
*Nonparametric Tests: Independent Samples grouped by existence of Head

NPTESTS
  /INDEPENDENT TEST (TSolTime CSolTime NumOfColl Distraction PerceivedTComplexity
    PerceivedBComplexity IComplexity IComprehensible ILearn IHLearn IHard SUse SatRobot SatDevice
    SatSelf ASkill AKnowledge AHelp APressure ANever TFailSafeR TSkillF TReliableR TFunctionalityF
    TWorksR TFeaturesF TAbandonR NRTLXMentalDemand NRTLXPhysicalDemand NRTLXTimePressure
    NRTLXPerformPressure NRTLXEffort NRTLXFrustration E1 E2 E3 E4 E5 Intuitiveness Satisfaction
    Acceptance Trust Reliability Functionality CognitiveWorkload Embodiment) GROUP (RobotHead)
    MANN_WHITNEY
  /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
  /CRITERIA ALPHA=0.05  CILEVEL=95.
```

Nonparametric Tests

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Track Solution Time in Seconds is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.570	Retain the null hypothesis.
2	The distribution of Building Solution Time in Seconds is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.500	Retain the null hypothesis.
3	The distribution of Number of Collisions is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.343	Retain the null hypothesis.
4	The distribution of IAT Measurement is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.	Unable to compute.
5	The distribution of Perceived Control Complexity is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.288	Retain the null hypothesis.
6	The distribution of Perceived Building Complexity is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.513	Retain the null hypothesis.
7	The distribution of Used Control Device was needlessly complex is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.882	Retain the null hypothesis.
8	The distribution of Used Control Device was easy comprehensible is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.629	Retain the null hypothesis.
9	The distribution of Used Control Device was fast to learn is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.927	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
10	The distribution of Used Control Device needs much to learn is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.494	Retain the null hypothesis.
11	The distribution of Used Control Device was hard to use is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.845	Retain the null hypothesis.
12	The distribution of Like to use device often is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.454	Retain the null hypothesis.
13	The distribution of Satisfied with robots' performance is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.643	Retain the null hypothesis.
14	The distribution of Satisfied with devices' performance is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.632	Retain the null hypothesis.
15	The distribution of Satisfied with own performance is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.370	Retain the null hypothesis.
16	The distribution of Own required skill for commanding robots with device is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.409	Retain the null hypothesis.
17	The distribution of Own required knowledge for commanding robots with device is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.232	Retain the null hypothesis.
18	The distribution of Would not be able to solve a task without help is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.040	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
19	The distribution of Would not be able to solve a task under time pressure is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.691	Retain the null hypothesis.
20	The distribution of Would never be able to solve a task is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.773	Retain the null hypothesis.
21	The distribution of The System is very FailSafe is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.171	Retain the null hypothesis.
22	The distribution of The System has the ability to handle what I want is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.116	Retain the null hypothesis.
23	The distribution of The System is very reliable is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.465	Retain the null hypothesis.
24	The distribution of The System has the required functionality is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.067	Retain the null hypothesis.
25	The distribution of The System works for me is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.122	Retain the null hypothesis.
26	The distribution of The System provides the Features I need for solving Tasks is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.058	Retain the null hypothesis.
27	The distribution of The System does not abandon me is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.392	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
28	The distribution of Mental Demand is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.187	Retain the null hypothesis.
29	The distribution of Physical Demand is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.081	Retain the null hypothesis.
30	The distribution of Time Pressure is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.156	Retain the null hypothesis.
31	The distribution of Performance Pressure is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.573	Retain the null hypothesis.
32	The distribution of Effort needed is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.068	Retain the null hypothesis.
33	The distribution of Frustration is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.697	Retain the null hypothesis.
34	The distribution of Unecht - Natürlich is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.862	Retain the null hypothesis.
35	The distribution of Wie eine Maschine - Wie ein Mensch is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.370	Retain the null hypothesis.
36	The distribution of Hat kein Bewusstsein - Hat ein Bewusstsein is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.459	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
37	The distribution of Künstlich - Realistisch is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.459	Retain the null hypothesis.
38	The distribution of Bewegt sich steif - Bewegt sich flüssig is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.728	Retain the null hypothesis.
39	The distribution of Intuitiveness is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.858	Retain the null hypothesis.
40	The distribution of Satisfaction is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.384	Retain the null hypothesis.
41	The distribution of Acceptance is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.918	Retain the null hypothesis.
42	The distribution of Trust is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.161	Retain the null hypothesis.
43	The distribution of Reliability is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.234	Retain the null hypothesis.
44	The distribution of Functionality is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.062	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
45	The distribution of CognitiveWorkload is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.319	Retain the null hypothesis.
46	The distribution of Embodiment is the same across categories of Existence of a Head.	Independent-Samples Mann-Whitney U Test	.640	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

```
*Nonparametric Tests: Independent Samples grouped by Number of Construction Task

NPTESTS
  /INDEPENDENT TEST (TSolTime CSolTime NumOfColl Distraction PerceivedTComplexity
    PerceivedBComplexity IComplexity IComprehensible ILearn IHLearn IHard SUse SatRobot SatDevice
    SatSelf ASkill AKnowledge AHelp APressure ANever TFailSafeR TSkillF TReliableR TFunctionalityF
    TWorksR TFeaturesF TAbandonR NRTLXMentalDemand NRTLXPhysicalDemand NRTLXTimePressure
    NRTLXPerformPressure NRTLXEffort NRTLXFrustration E1 E2 E3 E4 E5 Intuitiveness Satisfaction
    Acceptance Trust Reliability Functionality CognitiveWorkload Embodiment) GROUP (CTask)
  /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
  /CRITERIA ALPHA=0.05  CILEVEL=95.
```

Nonparametric Tests

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Track Solution Time in Seconds is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.954	Retain the null hypothesis.
2	The distribution of Building Solution Time in Seconds is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
3	The distribution of Number of Collisions is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.920	Retain the null hypothesis.
4	The distribution of IAT Measurement is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.	Unable to compute.
5	The distribution of Perceived Control Complexity is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.392	Retain the null hypothesis.
6	The distribution of Perceived Building Complexity is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.010	Reject the null hypothesis.
7	The distribution of Used Control Device was needlessly complex is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.443	Retain the null hypothesis.
8	The distribution of Used Control Device was easy comprehensible is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.397	Retain the null hypothesis.
9	The distribution of Used Control Device was fast to learn is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.017	Reject the null hypothesis.
10	The distribution of Used Control Device needs much to learn is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.037	Reject the null hypothesis.
11	The distribution of Used Control Device was hard to use is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.565	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
12	The distribution of Like to use device often is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.780	Retain the null hypothesis.
13	The distribution of Satisfied with robots' performance is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.507	Retain the null hypothesis.
14	The distribution of Satisfied with devices' performance is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.968	Retain the null hypothesis.
15	The distribution of Satisfied with own performance is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.359	Retain the null hypothesis.
16	The distribution of Own required skill for commanding robots with device is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.127	Retain the null hypothesis.
17	The distribution of Own required knowledge for commanding robots with device is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.455	Retain the null hypothesis.
18	The distribution of Would not be able to solve a task without help is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.557	Retain the null hypothesis.
19	The distribution of Would not be able to solve a task under time pressure is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.850	Retain the null hypothesis.
20	The distribution of Would never be able to solve a task is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.524	Retain the null hypothesis.
21	The distribution of The System is very FailSafe is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.803	Retain the null hypothesis.
22	The distribution of The System has the ability to handle what I want is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.803	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
23	The distribution of The System is very reliable is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.816	Retain the null hypothesis.
24	The distribution of The System has the required functionality is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.316	Retain the null hypothesis.
25	The distribution of The System works for me is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.792	Retain the null hypothesis.
26	The distribution of The System provides the Features I need for solving Tasks is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.501	Retain the null hypothesis.
27	The distribution of The System does not abandon me is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.817	Retain the null hypothesis.
28	The distribution of Mental Demand is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.584	Retain the null hypothesis.
29	The distribution of Physical Demand is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.049	Reject the null hypothesis.
30	The distribution of Time Pressure is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.050	Reject the null hypothesis.
31	The distribution of Performance Pressure is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.729	Retain the null hypothesis.
32	The distribution of Effort needed is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	1.000	Retain the null hypothesis.
33	The distribution of Frustration is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.516	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
34	The distribution of Unecht - Natürlich is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.	Unable to compute.
35	The distribution of Wie eine Maschine - Wie ein Mensch is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.	Unable to compute.
36	The distribution of Hat kein Bewusstsein - Hat ein Bewusstsein is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.	Unable to compute.
37	The distribution of Künstlich - Realistisch is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.	Unable to compute.
38	The distribution of Bewegt sich steif - Bewegt sich flüssig is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.	Unable to compute.
39	The distribution of Intuitiveness is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.114	Retain the null hypothesis.
40	The distribution of Satisfaction is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.799	Retain the null hypothesis.
41	The distribution of Acceptance is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.324	Retain the null hypothesis.
42	The distribution of Trust is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.811	Retain the null hypothesis.
43	The distribution of Reliability is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.822	Retain the null hypothesis.
44	The distribution of Functionality is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.604	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
45	The distribution of CognitiveWorkload is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.444	Retain the null hypothesis.
46	The distribution of Embodiment is the same across categories of Number of Construction Task.	Independent-Samples Kruskal-Wallis Test	.	Unable to compute.

Asymptotic significances are displayed. The significance level is .05.

```
*Nonparametric Tests: Independent Samples grouped by Gender

NPTESTS
  /INDEPENDENT TEST (TSolTime CSolTime NumOfColl Distraction PerceivedTComplexity
    PerceivedBComplexity IComplexity IComprehensible ILearn IHLearn IHard SUse SatRobot SatDevice
    SatSelf ASkill AKnowledge AHelp APressure ANever TFailSafeR TSkillF TReliableR TFunctionalityF
    TWorksR TFeaturesF TAbandonR NRTLXMentalDemand NRTLXPhysicalDemand NRTLXTimePressure
    NRTLXPerformPressure NRTLXEffort NRTLXFrustration E1 E2 E3 E4 E5 Intuitiveness Satisfaction
    Acceptance Trust Reliability Functionality CognitiveWorkload Embodiment) GROUP (Gender)
  /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
  /CRITERIA ALPHA=0.05  CILEVEL=95.
```

Nonparametric Tests

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Track Solution Time in Seconds is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.097	Retain the null hypothesis.
2	The distribution of Building Solution Time in Seconds is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.425	Retain the null hypothesis.
3	The distribution of Number of Collisions is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.296	Retain the null hypothesis.
4	The distribution of IAT Measurement is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.	Unable to compute.
5	The distribution of Perceived Control Complexity is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.697	Retain the null hypothesis.
6	The distribution of Perceived Building Complexity is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.535	Retain the null hypothesis.
7	The distribution of Used Control Device was needlessly complex is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.102	Retain the null hypothesis.
8	The distribution of Used Control Device was easy comprehensible is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.014	Reject the null hypothesis.
9	The distribution of Used Control Device was fast to learn is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.172	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
10	The distribution of Used Control Device needs much to learn is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.461	Retain the null hypothesis.
11	The distribution of Used Control Device was hard to use is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.602	Retain the null hypothesis.
12	The distribution of Like to use device often is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.109	Retain the null hypothesis.
13	The distribution of Satisfied with robots' performance is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.352	Retain the null hypothesis.
14	The distribution of Satisfied with devices' performance is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.222	Retain the null hypothesis.
15	The distribution of Satisfied with own performance is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.728	Retain the null hypothesis.
16	The distribution of Own required skill for commanding robots with device is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.824	Retain the null hypothesis.
17	The distribution of Own required knowledge for commanding robots with device is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.137	Retain the null hypothesis.
18	The distribution of Would not be able to solve a task without help is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.036	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
19	The distribution of Would not be able to solve a task under time pressure is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.333	Retain the null hypothesis.
20	The distribution of Would never be able to solve a task is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.604	Retain the null hypothesis.
21	The distribution of The System is very FailSafe is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.332	Retain the null hypothesis.
22	The distribution of The System has the ability to handle what I want is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.901	Retain the null hypothesis.
23	The distribution of The System is very reliable is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.232	Retain the null hypothesis.
24	The distribution of The System has the required functionality is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.111	Retain the null hypothesis.
25	The distribution of The System works for me is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.931	Retain the null hypothesis.
26	The distribution of The System provides the Features I need for solving Tasks is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.571	Retain the null hypothesis.
27	The distribution of The System does not abandon me is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.996	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
28	The distribution of Mental Demand is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
29	The distribution of Physical Demand is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.867	Retain the null hypothesis.
30	The distribution of Time Pressure is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.266	Retain the null hypothesis.
31	The distribution of Performance Pressure is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.162	Retain the null hypothesis.
32	The distribution of Effort needed is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.004	Reject the null hypothesis.
33	The distribution of Frustration is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.621	Retain the null hypothesis.
34	The distribution of Unecht - Natürlich is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.201	Retain the null hypothesis.
35	The distribution of Wie eine Maschine - Wie ein Mensch is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.151	Retain the null hypothesis.
36	The distribution of Hat kein Bewusstsein - Hat ein Bewusstsein is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.099	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
37	The distribution of Künstlich - Realistisch is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.082	Retain the null hypothesis.
38	The distribution of Bewegt sich steif - Bewegt sich flüssig is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.038	Retain the null hypothesis.
39	The distribution of Intuitiveness is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.085	Retain the null hypothesis.
40	The distribution of Satisfaction is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.203	Retain the null hypothesis.
41	The distribution of Acceptance is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.169	Retain the null hypothesis.
42	The distribution of Trust is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.780	Retain the null hypothesis.
43	The distribution of Reliability is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.526	Retain the null hypothesis.
44	The distribution of Functionality is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.673	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
45	The distribution of CognitiveWorkload is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.005	Reject the null hypothesis.
46	The distribution of Embodiment is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.027	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.