HUMAN-ROBOT TRUST

IS MOTION FLUENCY AN EFFECTIVE BEHAVIORAL STYLE FOR REGULATING ROBOT TRUSTWORTHINESS?

AUTHOR
M.E.U. LIGTHART (MIKE)

SUPERVISORS
R. (RIK) VAN DEN BRULE[#]* AND W.F.G. (PIM) HASELAGER*

* DONDERS CENTER FOR COGNITION, RADBOUD UNIVERSITY NIJMEGEN, MONTESSORILAAN 3, 6525 HR NIJMEGEN
# BEHAVIORAL SCIENCE INSTITUTE, RADBOUD UNIVERSITY NIJMEGEN, MONTESSORILAAN 3, 6525 HR NIJMEGEN

CORRESPONDING INFORMATION
MIKE LIGTHART,
E-MAIL: m.ligthart@student.ru.nl
TEL: +31628538529

ABSTRACT

Finding good features to regulate robot trustworthiness will optimize the usage of robots. In previous research done by van den Brule et al. (submitted) motion fluency (smooth robot motions versus trembling robot motions) is studied. After a round of movie experiments a main effect is found, while after a round of Immersive Virtual Environment Technology (IVET) experiments no effect is observed. In this research, I explore the question whether the length of the task contributes to the presence of a fluency effect on the trustworthiness of a robot. More specifically it is investigated whether an effect of motion fluency is present in a short version of a task and disappears when the task is longer. The task in the virtual reality experiment consisted out of two actors: a (human) participant and a virtual robot. Both perform the Van Halen task: an actor has to pick brown balls from a conveyor belt and let other colored ball pass. The goal of the participant is to maximize the sum of their own score and the robot's score. By correcting the robot when it makes a mistake a penalty is prevented. Results showed there is no significant difference between the measured trustworthiness in any of the four conditions. This might be because there is no effect of motion fluency. The meta-study of Hancock et al. (2011) indicates that no effect is very plausible. Also a null effect can be caused by the cognitive load of the task. A different explanation is the influence of interfering factors. An effect of motion fluency might be masked or countered by the recency effect (Desai et al., 2013), the assimilation effect (Herr et al., 1983) or the use of virtual reality (Bainbridge et al., 2008). All possibilities require further research.

TABLE OF CONTENTS

INTRODUCTION

Social robots are not only the future, they are here right now. Meet TWENDY-ONE (see Figure 1), a human cooperative robot who can help both in the household as in hospitals. It is designed to help ageing people with their daily activities. It took about ten years to develop and is scheduled to become commercially available in 2015.

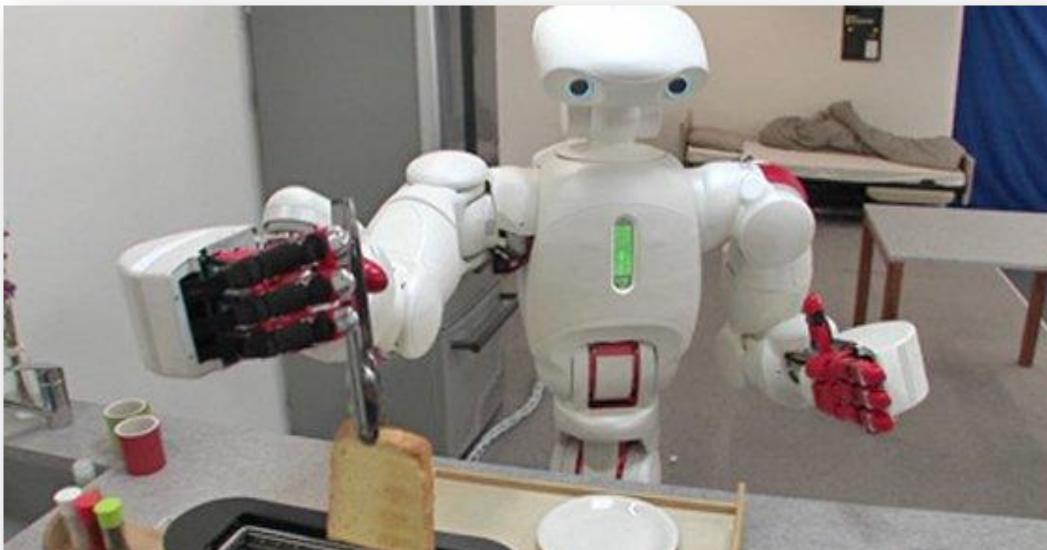

Figure 1: TWENDY-ONE helping in the kitchen. Source: gizmag.com.

The researchers who constructed TWENDY-ONE focused on safety, dependability and dexterity. A special silicon coating is fitted onto its exterior to soften possible collisions. TWENDY-ONE has numerous sensors just under the surface to make a distinction between living and inanimate objects. TWENDY-ONE automatically operates more carefully when dealing with a human than with, for example, a broom which requires less precision. The high amount of degrees of freedom in the hands, arms and torso give TWENDY-ONE the possibility to handle everything in the range from small sensitive objects, like a plastic cup, to a more heavy and large cargo such as humans (Iwata & Sugano, 2009; Quick, 2010).

TWENDY-ONE is not the only example: Bandit is a socially assistive robot that helps caregivers treating their patients. By mimicking movements, making funny sounds, and blowing bubbles he connects with children and the elderly alike. He is specially designed for motivating people to exercise (Fasola & Matarić, 2012; Karlin, 2010). Bandit and TWENDY-ONE are just the tip of the iceberg in the field of social robotics (Fong, Nourbakhsh, & Dautenhahn, 2003).

Social robots are becoming more capable in handling multiple tasks. Due to natural limitations a preference remains for one (group of) task(s). This means, for example, that a cleaning robot is capable of doing laundry unsupervised, but needs some kind of human supervision while doing the dishes. The use of (multipurpose) social robots can be improved by introducing an intuitive way of letting the user know when to check up more often and when not to. Van den Brule, Dotsch, Bijlstra, Wigboldus, & Haselager, (submitted) argue that research on trust between humans and robots can be used to provide an intuitive way of communicating and optimizing the use of robots. This research is an extension of the research done by van den Brule et al. (submitted). Before specifying my research question I will first explain the relevance of human-trust research in more detail, followed by the introduction of a framework. This framework, containing the antecedents of human-robot trust, gives an overview of all elements influencing robot trust.

IMPORTANCE OF HUMAN-ROBOT TRUST RESEARCH

Developments of interactive or social robots can be found in every direction (Oleson, Billings, Kocsis, Chen, & Hancock, 2011). For example, robots are able to improve the performances of human-robot teams by extending human sensory, psychomotor, and cognitive abilities (Riley, Strater, Chappell, Connors, & Endsley, 2010). Two non-optimal patterns are discovered, during both experiments and real experiences from the field with human-robot teams. The first pattern is when a team over-trusts a robot and expects more from the robot than what it actually can perform (Parasuraman & Riley, 1997). This is called *misuse*. The second pattern is when a team does not trust the robot and chooses not to use it, although the use of the robot would have resulted in a better outcome (Desai, Stubbs, Steinfeld, & Yanco, 2009). This is called *disuse.* The mentioned studies indicate that it is very important to know as much as possible about how trust arises and evolves between a human and a robot for a robotic system to be used effectively. (Lee & See, 2004) argue that calibrating trust or matching the human's perception of the robot to reality might be the key for avoiding misuse and disuse. This means that trust influences the success of the interaction between a human and a robot, and plays an important role in the usage of robots in the future (Freedy, DeVisser, Weltman, & Coeyman, 2007).

ANTECEDENTS OF HUMAN-ROBOT TRUST

Human-robot trust is a complicated web of actors and factors that all have some influence on trust. A framework is needed to enforce some structure into this web. (Oleson, Billings, Kocsis, Chen, & Hancock, 2011) propose such a framework which contains all kinds of categories from which trust can be build: the so called antecedents of human-robot trust (see Figure 2). These antecedents are divided in three main categories: human-related, environment related, and robot-related. A meta-analysis using empirical studies resulted in correlation effect sizes, r, (10 studies) and experiment effect sizes, Cohens'd, (11 studies) for all factors (Hancock, Billings, Schaefer, Chen, de Visser, & Parasuraman, 2011). Human related antecedents have both a small correlational (r = .09) and experimental (d =-.02) effect size. Environmental-related antecedents have both a medium correlational (r = .11) and experimental (d =.47) effect size. Robot-related antecedents have overall a high correlational (r = .24) and experimental (d =.67) effect size. Although more studies are needed to be even more conclusive, Hancock et al. (2011) state that "*the current summary of findings emphasizes the importance of focusing on robot-related factors in design and training guidelines for HRI*".

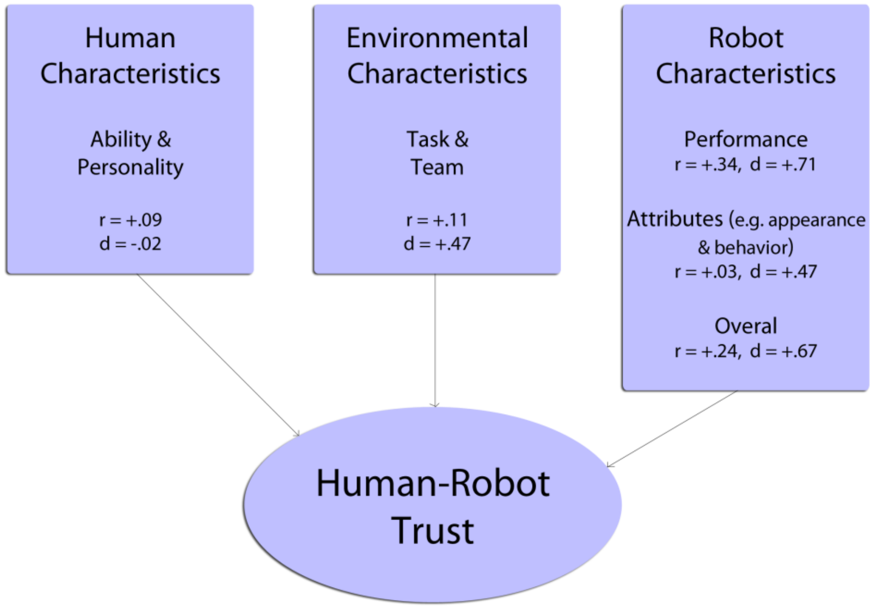

Figure 2: trust framework containing antecedents of human-robot trust divided into three categories: human-related, robot-related or environment related. Included are the effect sizes based on the meta-analysis of (Hancock, Billings, Schaefer, Chen, de Visser, & Parasuraman, 2011). r is the correlational effect size and d is the experimental effect size.

This image was adapted from fig. 1 in Oleson, Billings, Kocsis, Chen, & Hancock (2011).

HUMAN-RELATED & ENVIRONMENT-RELATED FACTORS

Human-related factors are rooted within each individual. Ability-based factors such as prior experiences or expertise gained using the system are included in this category. Experiences with the system lead to forming expectations which, in turn, helps to predict the systems behavior (Muir, 1994). The level of training also influences the development of trust in a system (Freedy, DeVisser, Weltman, & Coeyman, 2007).

Personality-based factors are also part of the human-related factors. Personality-based factors consist of the feelings and the personality of an individual. For example self-confidence (Freedy, DeVisser, Weltman, & Coeyman, 2007) or the ability to trust other humans (Lee & See, 2004) play a role in human-robot trust.

Environment-related factors are built up from team factors and tasks factors. Team factors depend on the composition and interaction between a team. (Adams, Bruyn, Houde, & Angelopoulos, 2003) suggest that teamwork and situation awareness have an effect on trust development. The direction of the task-related effect that influences trust is still unclear. More complex and demanding tasks can have different effects on trust e.g. trust will decrease when there is a rise in task complexity because there is more room for error; an individual relies more on him- or herself in that situation (Adams, Bruyn, Houde, & Angelopoulos, 2003). The opposite is also possible according to (Atoyan, Duquet, & Robert, 2006): because the individual is too preoccupied, over-trusting the system can occur. In this case an increase in trust is observed.

ROBOT-RELATED FACTORS

Robot related factors are intrinsic to the robot. Two subcategories are robot performance and robot attributes. Robot performance includes features that have a direct influence on how humans perceive the performance of the robot. In the experiment by van den Brule et al. (submitted) the number of mistakes by the robot is such a feature. More mistakes, thus a lower perceived performance, mean that there will be less trust invested by a participant (Van den Brule, Dotsch, Bijlstra, Wigboldus, & Haselager, submitted).

Robot performance has a strong effect on trust. Besides the experiment by van den Brule et al. (submitted) it is used by many other researchers (Hancock et al., 2011). Because performance effects are widely acknowledged I will leave it out of my experiment. Finding an effective robot attribute that can be used in regulating trust is far more interesting. The two types of attributes that I will further explore are appearance and behavioral style. Appearance is also a widely studied topic but none of the sources in the meta-analyses of Hancock et al. (2011) studied behavioral style. The research of van den Brule et al. (submitted) covers different behavioral styles. I am going to extend the behavioral style research of van den Brule et al. (submitted). To make the overview complete, appearance-based attributes will be reviewed first, before the behavioral styles are explored.

Appearance

Appearance, anthropomorphism, facial expression and attractiveness are robot attribute-based factors of influence in trust initially as well as gradually during the process of a task (Chang, Doll, M., Frank, & Sanfey, 2010). Chang et al. (2010) used Trust Games to estimate trustworthiness based on the factors mentioned above and the experienced trustworthiness based on the amount of reciprocation. They found a near significant trend where participants tended to invest more money in trustworthy looking robots than in their untrustworthy looking counterparts. They also found, as expected, that high reciprocations gain more investments (and therefore more trust). As a reminder: this is a performance-based attribute. On the first round, so without any experience or knowledge of the performance of the Trust Game partners, participants invested more in the trustworthy looking than in the untrustworthy looking partner. This effect was weak but present until the last round. Chang et al. (2010) conclude that an interaction effect is present between the initial trustworthiness and the amount of money a partner reciprocated after each investment in the amount of trust given to that partner. Meaning that if a partner was perceived more trustworthy in the beginning, more trust was invested in the end if that partner reciprocated more. Furthermore when the reciprocating rate was lower or when the partner was perceived less trustworthy in the beginning less trust was invested.

The next step that Chang et al. (2010) took was to discover how this all related to each other. With the help of reinforced learning models they found that initial trustworthiness influences the way participants use the feedback, but the feedback also influences the expectations. They call it the Dynamic belief model (Chang, Doll, M., Frank, & Sanfey, 2010).

Behavior style and remaining questions

A different area within robot attributes is behavior style. Van den Brule et al. (submitted) try to find easy adaptable behavioral styles for manipulating the trustworthiness of robots in their research. Three different robot-behavioral styles where tested: fluency in movement, looking and hesitating behavior.

Three main steps are taken in the research of van den Brule et al. (submitted). The first is to conduct movie experiments. Participants have a passive role and look at movies of robots performing a task. A questionnaire at the end is used to measure trust. Only the behaviors with significant results continue to the next step. This next step is conducting experiments with Immersive Virtual Environment Technology (IVET). Participants perform a task in a virtual environment. This provides a more realistic experience. Different compared with the movie experiments is that participants now have an active role during the task. The main reason for conducting IVET experiments is the lack of a suitable robot. Even with a suitable robot, it remains a costly operation to test all the behaviors. A preliminary selection of the behaviors based on movie and IVET experiments reduces the amount of resources needed. The final step is doing experiments with real robot using the behaviors that have a significant effect in the IVET experiments. Van den Brule et al. (submitted) did not reach the final step yet.

The results from the movie experiment indicate that only fluency of movement had a significant effect on the trustworthiness of the robot besides the expected strong effect of its performance. The significant interaction effect of performance x motion fluency indicates that the fluency effect is stronger when the robot performed better and is weaker when a robot performed badly. Only the fluency of movements was used in the IVET experiment.

Besides the questionnaire, which only measures an explicit form of trust, looking behavior of the participant was used to measure (implicit) trust in the IVET experiment. The final results of the IVET experiment show that there was indeed, again, a main effect for robot performance on both the explicit and implicit trust measures. But in contradiction with the movie experiments, no relation was found between motion fluency and explicit as well as implicit trustworthiness. Furthermore no relationship between the implicit and explicit trust measures was found either. (Van den Brule, Dotsch, Bijlstra, Wigboldus, & Haselager, submitted) give the following explanation for the absence of a motion fluency effect on the robot’s trustworthiness in their experiment:

> “*One explanation might be that participants were too focused on their own conveyor to connect the motion fluency of the robot to its trustworthiness. Alternatively, the length of the experiment might have influenced the outcome. Participants may have grown accustomed to the motion fluency of the robot. While motion fluency may have had an effect on the trustworthiness of the robot early in the experiment, participants were asked to rate the robot much later, at the end of the experiment.*”

I am interested in the second explanation which directly leads to my research question. Van den Brule et al. (submitted) ask themselves if the length of the experiment influences the outcome and if motion fluency has an effect early in the experiment. They suggest that participants might have grown accustomed to the fluency of the robot. Since the measuring of trustworthiness, done with a questionnaire, can only be done in the end, I propose to do a short and long version of the experiment. Besides the introduction of the duration factor the performance is kept constant. Performance is a strong and always present factor which might overshadow other factors. Keeping performance constant should eliminate this possible effect. I expect that in the short condition, where the participant has not grown accustomed to the behavior of the robot, an effect of motion fluency on trust will be found. In the long condition, where the length is equal to the duration of the experiment done by van den Brule et al. (submitted), no motion fluency is to be expected.

Research question and hypotheses

In the previous section a next step in the robot-human trust research started by van den Brule et al. (submitted) has been proposed. It contained a couple of questions and a proposal for a new experiment together with some expectations. A short summary can be found below.

Following the second explanation on why no motion fluency effect was found in the experiment done by van den Brule et al. (submitted) a research question is formulated:

- Does the length of the Van Halen experiment influence the presence of a fluency effect on the trustworthiness of a robot?

Research question reframed to the following hypothesis:

- Motion fluency (trembling vs. fluent) will have a main effect on trustworthiness only in the short condition of the Van Halen task. In the long condition no motion fluency effect is present.
    - This means that there is an interaction effect between motion fluency and the duration of the Van Halen task.

Direction of motion fluency main effect:

- In the short condition a fluent robot will be valued more trustworthy than a trembling robot.
- In the long condition a fluent robot will not be valued differently in terms of trustworthiness than a trembling robot.

METHOD

The experiment in this thesis was an altered version of the experiment by (Van den Brule, Dotsch, Bijlstra, Wigboldus, & Haselager, submitted) and the used hardware and software were similar. Therefore the methods and materials largely followed Van den Brule et al. (submitted).

PARTICIPANTS

103 participants started the experiment from which 3 could not finish the entire procedure. They experienced dizziness or even nausea. Therefore they discontinued the experiment. A priory was decided to exclude any participant that did not finish the whole experiment. The three participants in question were therefore excluded. Participant #101 – #103 where put in the same condition as the excluded participants. This left 100 participants who fully finished the experiment and were used in the analyses (19 men, 81 women, age: median 20.5, range 18-64). An a priori power analysis[1] for this design determined that 100 participants should provide enough power (π = .80) for a medium effect size ($\eta_p^2 \geq$ .06) for both a main effect of motion fluency as an interaction effect of motion fluency x duration (Cohen, 1988; Wilson VanVoorhis & Morgan, 2007). From 100 participants, 95 participants were recruited from the Radboud University Nijmegen participant pool and 5 signed up after a message on an online social network. As a reward participants received either course credit or a €5 gift card. Participants were randomly assigned to the conditions. There were two experiment leaders (#1 tested 45 participants and #2 tested 55 participants).

IMMERSIVE VIRTUAL ENVIRONMENT

The experiment was held in the Radboud Immersive Virtual Environment Research lab (RIVERlab). The participants wore a Head Mounted Display (HMD) of the type nVisor SX60, which gave them visual access to the virtual world. The HMD generates a stereoscopic 3D view (frame rate: 60 Hz, resolution: 1280x1024 pixels, horizontal field of view: 44° and vertical field of view: 38°). On top of the HMD a tracking sensor was located which tracks the movements of the head via the InterSense IS-900 tracking

[1] Power analysis was performed with G*Power 3.1.7 (http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/)

system at a sampling rate of 300Hz. To manipulate a cursor in the virtual world the participants held a hand tracking sensor in their right hand to track the movements of the hand. The virtual world was generated with WorldViz Vizard 3.0. A standard available virtual room was adapted to suit the needs of the task (see below). The information generated by the tracking system was inserted into Vizard to provide the immersive experience.

TASK

The task incorporated two actors; besides the participant, also a robot was present. Both the participant and the robot performed the Van Halen task. The task of the participants consisted out of two elements that had to be performed at the same time: executing the Van Halen task and correcting mistakes made by the robot that performed its own Van Halen task. The goal of the participant was to maximize the combined score.

The first element of the participant's task was performing the Van Halen task itself. The participant sat in front of a virtual conveyor belt where balls move from right to left. Each ball was assigned a uniform color e.g. blue, red, brown, et cetera. The participant had to pick the brown balls from the conveyor and let the rest pass. The participant could accomplish this by moving his/her right hand forward and therefore move the virtual cursor towards the ball. Touching a ball removed it from the conveyor belt. A scoreboard above the conveyor kept the score of the participant. For each brown ball taken the participants got a point. More points were collected for every non-brown passed. Mistakes, such as taking a different colored ball or letting the brown ball pass, were punished by subtracting a point from the total score. These errors were also displayed on the scoreboard.

The second element of the participant's task was to correct the robot that also performed the Van Halen task. The robot stood in front of a conveyor belt. The robot also had to pick the brown balls and let the rest pass. The same scoring system was used and displayed, on a different scoreboard, behind the robot. The robot was programmed to make mistakes. When the robot made a mistake the participant could prevent the subtraction of points from the robots score by pressing a virtual, big green, button. When the participant mistakenly pressed the button, e.g. when no mistake was made by the robot, 10 points were subtracted from the participants own score while sounding a buzzer. The used robot was a virtual copy of TWENDY-ONE introduced in the introduction (see Figure 3).

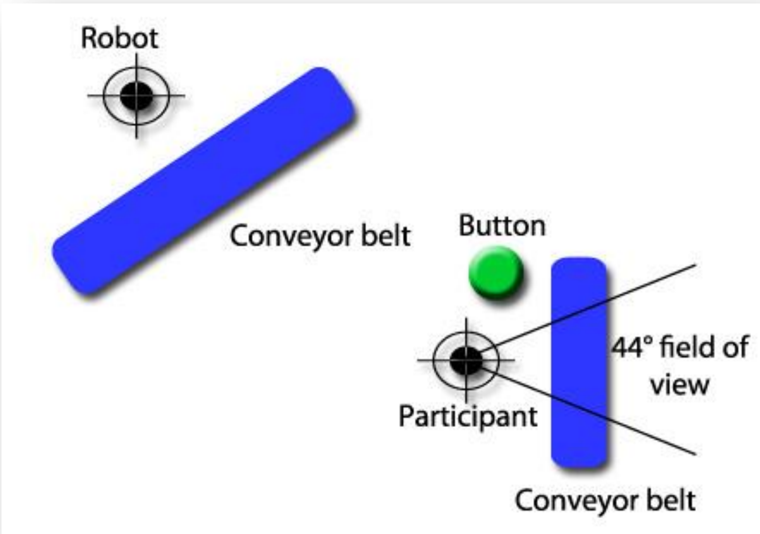

Figure 4: experimental setup – the participant has a horizontal field of view: 44° and vertical field of view: 38°. This means that participant cannot see the whole set up at once. He has to rotate to the left to view the button and even further to the left to view the button and even further to the left to view the robot. To participant has to tilt his head upwards to view the scoreboard behind the conveyor belts.

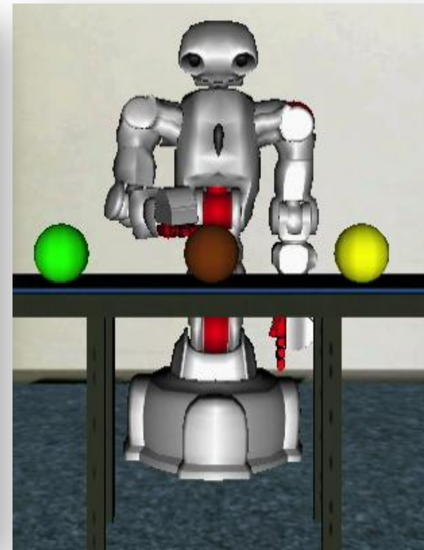

Figure 3: virtual copy of TWENDY-ONE performing the Van Halen Task

The conveyor belt of the robot was located in such a way that it could not be seen in the same view. The participant had to turn left, on the rotating chair, to check up on the robot (see Figure 4).

CONDITIONS

There were two between subject factors with each two levels in the experiment. This resulted in four different conditions (see table 1)

To research if the length of the experiment had influence on trustworthiness a new between-subject factor was introduced: duration (long vs. short). The long duration consisted of three blocks of the dual Van Halen task and the short duration only one block. Motion fluency (fluently vs. trembling) was the second between subject factor. The robot accomplished the task respectively in a fluent or trembling manner.

	Motion Fluency	
Duration	**Fluently**	**Trembling**
Long	Long-Fluently	Long-Trembling
Short	Short-Fluently	Short-Trembling

Table 1: conditions of the experiment. There were two between subject factors with each two levels. Duration (long vs. short) and Motion Fluency (fluently vs. trembling)

PROCEDURE

All participants were escorted from the front desk to the RIVERlab. After reading the instructions and signing the informed consent form the participants were asked to take a seat on the rotating chair in the middle of the room. The virtual environment was loaded and the HMD en sensors were mounted on the participant's head and hand respectively. The participant was introduced to the virtual world and the Van Halen task. A practice round was initiated after pressing the virtual green button. Now the participant could practice its own task: removing brown balls from a conveyor belt. This was followed by the introduction of the robot. The participant was directed to the left by rotating its chair. Behind the second conveyor belt the robot appeared after pressing the green button. A trial version of the Van Halen task of the robot was initiated. The participant was instructed to press the green button whenever the robot makes a mistake. After this round a final practice round was started. This gave the participant the opportunity to try both tasks at the same time. Before the real experiment started the participant was instructed, for the last time, to look at the robot conveyor belt. Where, after pressing the green button, the test robot was replaced by the experiment robot. The experiment robot had exactly the same appearance as the test robot apart from the color. Participants were told that this second robot was the "real" robot with different characteristics as the previous or "practice" robot. From this point on the active condition determined the duration of the experiment (short versus long) and the motion fluency of the new robot (fluently versus trembling). During the real task the experimenter did not instruct the participant any longer but remained present in the lab. When the last ball left the conveyor and the HMD and the hand sensor were removed, the participant was asked to fill in the questionnaire. The experimenter left the room while the participant answered the questions.

QUESTIONNAIRE

In this experiment the dependent variable, trustworthiness, was measured with a questionnaire. Four constructs were used to capture the trustworthiness of the robot: 1) reliability, 2) trust, a 3) positive and a 4) negative valance item. The negative valance item was reversed coded for the analysis. The following questions were used (for original Dutch questionnaire see appendix A.2.):

Reliability: to what extent does the robot appear reliable?

Trust: to what extent would you trust the robot sorting the balls without human supervision?

Positive valence: how positive would you grade the robot?

Negative valence: how negative would you grade the robot?

Responses were recorded on a 7-point Likert scale (1 = not at all and 7 = extremely). These four constructs were taken together in a compound measure of trustworthiness. The Likert questions and the restructuring into a compound measure were also similar to the process of Van den Brule et al. (submitted). Moreover this compound variable was highly reliable (Cronbach's $\alpha = .816$) therefore this combined trustworthiness measure could be used on its own in the analysis. Furthermore two manipulation checks were added to the questionnaire. One assessed if the participant had seen or claimed to have seen the robot tremble and the other assessed the level of performance of the robot. The rest of the questions in the questionnaire were for gathering demographics.

RESULTS

DESIGN

There were two between-subject factors: duration (long versus short) and motion fluency (smooth versus trembling). There was one dependent variable: trustworthiness. The dependent variable was measured with a questionnaire and summarized in a compound measure.

ANALYSIS AND ASSUMPTIONS

Unless otherwise specified all analyses were ANOVA's. In particular the compound trustworthiness measure was analyzed with a 2 (duration: long vs. short) x 2 (motion fluency: smooth vs. trembling) between subject ANOVA. Two tests were executed to check if the trustworthiness measure was normal distributed: the Kolmogorov-Smirnov test and the Shapiro-Wilk test. Variables in each group were normally distributed (see appendix - Table 2). The histograms showed a slightly different picture (see appendix - Figure 6). They indicated a leptokurtic distribution for half of the groups: the short-smooth condition was bimodal and the long-smooth condition was multimodal. Furthermore the short-trembling condition was fairly skewed. Only the long-trembling matched all the characteristics of a normal distribution. The possible violations of normality posed no problem since the F-statistic is fairly robust against violations of normality when group sizes are equal (Lunney, 1970; Donaldson, 1968). Beforehand manipulation, experimenter and acquaintance checks were done to check for unwanted effects (see next section).

CHECKS

Checks were done to investigate some confounding factors. Manipulation checks were performed to see if participants noticed the trembling behavior in the right condition. Experimenter checks were performed to investigate if the experimenter who conducted the experiment had any influenced on any factor. For similar reasons the participants knowing one or both of the experimenters were inspected. Performance checks were done to see if the participants perceived a difference in performance between the conditions. Since performance was not manipulated no effect was expected.

MANIPULATION CHECKS

The manipulation check for the fluency behavior indicated that the participants noticed the motion fluency manipulation. Participants perceived a trembling robot more trembling ($M_{preceived\_trembling}$ = 4.00, *SD* = 1.78) than a robot with fluent behavior ($M_{preceived\_trembling}$ = 1.88 , *SD* = 1.06): $F(1,96) = 55.08$, $p < .0001$ and $\eta_p^2 = .365$. Although no differences were expected between the long ($M_{preceived\_trembling}$ = 2.56, *SD* = 1.67) and short ($M_{preceived\_trembling}$ = 3.32, *SD* = 1.88) condition, a small duration effect was found in the fluency manipulation check: $F(1,96) = 7.08$, $p = .009$ and $\eta_p^2 = .069$. No fluency x duration interaction effect was found $F(1,96) = .490$, $p = .486$.

EXPERIMENTER AND ACQUAINTANCE CHECKS

To make sure that the experimenter who guided the experiment had no effect on the trustworthiness measure an experimenter check was done. A three-way ANOVA with trustworthiness as dependent variable and motion fluency, duration and experimenter as between-subject factors confirmed the absence of any influence of the experimenter: all *F*'s < 1.0, *p*'s > .48

Furthermore a one-way ANOVA was conducted to compare trustworthiness when participants respectively knew or did not know one of the experimenters. There was a not a significant difference in the scores for knowing the experimenter (*M* = 3.15 , *SD* = .99) and not knowing the experimenter (*M* = 3.49, *SD* = 1.01) conditions; $F(1,98) = .561$, $p = .456$.

PERFORMANCE CHECKS

A performance check was conducted to investigate if participants noticed a difference in performance. The performance measure was the dependent variable where motion fluency and duration were the between subject factors. Since the performance of the robot is kept constant no difference was expected. This was confirmed by a two-way ANOVA: all *F*'s < 1.46, *p* > .23.

TRUSTWORTHINESS

The two-way ANOVA of the trustworthiness measure showed that there was no difference between trembling and fluent behavior in any condition: all *F's* < 1, *p*'s > .608 (see Figure 5).

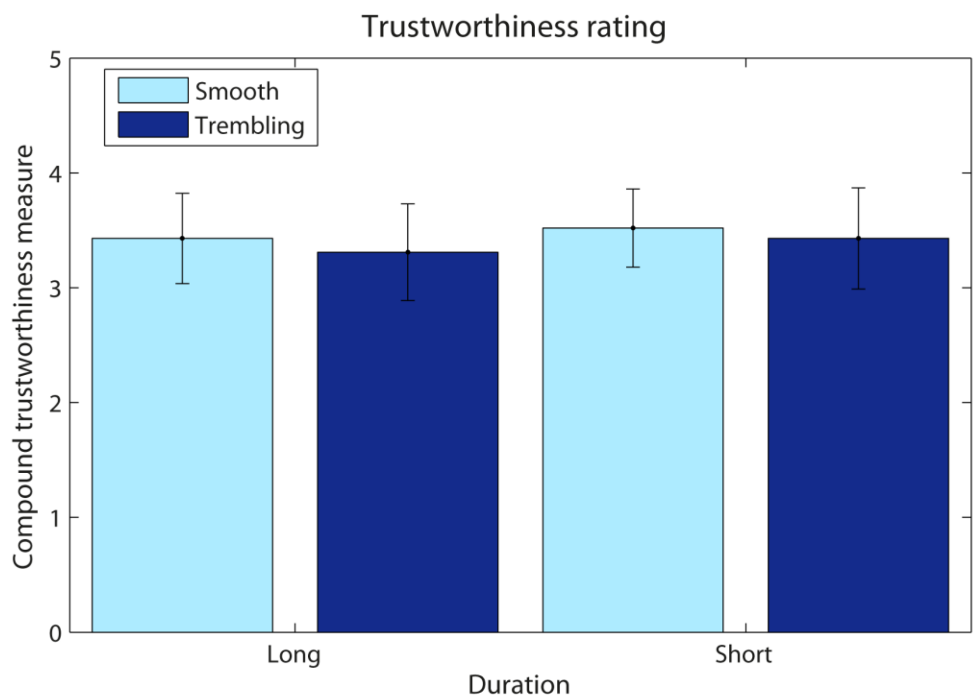

Figure 5: the means and 95%-confidence intervals shown in a bar plot. An ANOVA showed no significant difference between any of the conditions.

DISCUSSION

The results from the experiment showed there is no significant difference between the measured trustworthiness in any of the four conditions. Manipulation checks confirm the absence of the possibility of the participants not noticing the trembling of a robot. Experimenter and acquaintance checks confirm the absence of the influence of the difference in experimenter or participants knowing one of the experimenters. The performance check confirms the assumption that performance is not an interfering factor. There are two possible main explanations for the insignificant difference between the conditions:

1) There is no effect of motion fluency on trust in both the short and long conditions of the Van Halen task.
2) There is a motion fluency effect on trust but the experimental design does not allow measuring the effect.

In the following sections both explanations will be discusses in more detail.

NO EFFECT OF MOTION FLUENCY

The found evidence indicates that an effect of motion fluency is not present in any condition. I will discuss two possible explanations that make this conclusion more plausible.

Let's look back at the framework of Hancock et al. (2011). Behavioral styles, such as motion fluency, are a subclass of robot attributes. The effect of robot attributes on trust is medium-small[2] compared to other trust related factors e.g. robot performance or environmental trust factors. Given a power in this study of $\pi = .80$ and 100 participants it is reasonable to state that the effect of motion fluency on trust is too small or is even non-existing (Hancock, Billings, Schaefer, Chen, de Visser, & Parasuraman, 2011). However behavioral styles are not directly part of the meta-study of Hancock et al. (2011) which means that more research is required.

A different approach is to look at the differences between the video experiments of van den Brule et al. (submitted) and the IVET experiment described in this thesis. In the video experiments a significant

[2] The correlational effect sizes are small and the experimental effect sizes are medium (Cohen, 1988)

main effect is found of motion fluency on trust. A big difference, as pointed out in the introduction, is the activeness of the task of the participant. In the movie experiments participants have a passive role. They make their judgment on robot trustworthiness after seeing a video of a robot performing a task. In the IVET experiment participants have an active role. They make their judgment not only after seeing a robot perform a task but also after simultaneously performing their own task. This results in a difference in cognitive load. Participants are far less strained in the video task. During the movie experiments participants have more resources to actively deliberate the possible influences of the trembling of the robot. During the IVET experiment participants are too busy, in other words cognitively occupied, to actively deliberate about the trembling of the robot even though participants perceived the robot as more trembling in the corresponding conditions. In other words, although they do notice the robot's trembling, this may not become part of their cognitive evaluation concerning the robot's trustworthiness. This would imply that processing information concerning motion fluency, in relation with trust evaluation, is not an automated process. If this is the case motion fluency can never have an effect on trust in settings where people are cognitively occupied.

INTERFERING FACTORS

The absence of a significant difference between all the conditions does not necessarily mean an absence of an effect of motion fluency on trust. In this section I will discuss several possible interfering factors that might hide or counter a motion fluency effect. This implies that if we change the design of the experiment, in such a way that we prevent any influence of the discussed interfering factors, an effect of motion fluency can be found.

The first interfering factor is the recency effect. Desai et al. (2013) argue that measuring trust post-run masks real-time changes in trust. This results in a reduction of factors like inertia which leads to biases like the recency effect. This means that when the questionnaire is presented to the participants they recall their most recent experiences with the robot. If the hypothesis is correct and there is an effect of motion fluency in the short condition we are too late to measure it, because at this point in the experiment the possible effect of motion fluency is already gone. Desai et al. (2013) research finds early drops in reliability of traditional post-run survey approaches. Furthermore they argue that predicting real-time trust from user behavior is more feasible than trying to predict post-run survey trust (Desai, Kaniarasu, Medvev, Steinfeld, & Yanco, 2013).

The second interfering factor is the assimilation effect. Let's again look back at the framework introduced by Hancock et al. (2011). Experience or training with a robot will help predict the future behavior of that robot (Muir, 1994). Training can reduce initial biases and strengthen the development of trust by giving participants information about the robot capabilities and behaviors (Freedy, DeVisser, Weltman, & Coeyman, 2007). In general this is a good thing. Experience and training strengthens the knowledge a human has about a robot and therefore reduces disuse and misuse (Parasuraman & Riley, 1997). In relation to the experiment it might have a negative effect. Participants might not make a big enough distinction between the first robot, purely meant for introducing the tasks, and the second, experimental robot. Besides the difference in color both robots are equal in appearance. It might not be enough to let the participant observe the replacement of the first robot by the second robot while stating that they are different (the exact procedure is described in the method section). Participants assimilate their experiences with the first robot and project it to the behavior of the second robot. Since the first robot is constant over all the conditions no difference will be found in the questionnaire. This is called the assimilation effect (Herr, Sherman, & Fazio, 1983). A primacy effect might result in the same outcome. Participants use, as opposed to the recency effect, only their first experience for evaluating the robot. Since the first experience was with the first robot the same consequences as the assimilation effect apply resulting in using the experiences of the first robot for evaluating the robot (Desai, Kaniarasu, Medvev, Steinfeld, & Yanco, 2013).

A final interfering factor might be the use of virtual reality. Bainbridge et al. (2008) suggests that co-located robots, robots that are actually there, may gain more trust then robots in a virtual environment. Glitches caused by the use of virtual reality, e.g. fluctuations in frame rate that might interfere with the perceiving of the trembling behavior, result in a different experience than was meant in the experimental design. In other words participants might attribute the behavior of the robot to virtual reality. These problems will not occur with real robots. Although different challenges have to be faced, experiments with real robot can lead to reliable trust measurements (Bainbridge, Hart, Kim, & Scassellati, 2008).

SUMMARY AND FURTHER RESEARCH

The results from the experiment showed there is no significant difference between the measured trustworthiness in any of the four conditions. In the previous section two main conclusions were discussed. Since there are no significant results proving either of the conclusions, we can only provide possible explanations for the lack of significant results. Each of the possible explanations needs further research:

1) There is no effect of motion fluency on trust in both the short and long conditions of the Van Halen task.
    a. The framework of human-robot trust antecedents suggests a null effect (Hancock, Billings, Schaefer, Chen, de Visser, & Parasuraman, 2011). Furthermore given a power of $\pi = .80$ and 100 participants it is reasonable to state that the effect of motion fluency on trust is too small or is even non-existing. Since the meta-analysis does not actually include behavioral styles, let alone motion fluency, as a trust antecedent, more experiments are needed to provide more effect sizes.
    b. The relatively high cognitive load experienced by the participants during the experiment prevents them to actively take the motions of the robot into account. If motion fluency in relation with trust is not an automated process it would explain that an effect was found in the movie experiments done by van den Brule et al. (submitted) and no effect was found in the experiment described in this thesis. The relation between cognitive load and motion fluency needs to be researched further. For example by repeating the experiment

after removing the Van Halen task for a selection of the participants. Leaving the part where the participant just has to correct the robot would make the task less active.

The proposed new experimental set up makes room for additional research. For example a study to investigate if reaction time might be a (different) measure for trust. It seems plausible that a participant would react slower if the robot is perceived as more trustworthy. It might be that the participant would invest less attention if he or she trusts the robot more. Attention has an effect on reaction time (Welford, 1980).

The further research on cognitive load and the investigation of the relation between reaction time and trust can be studied separately.

2) There is a motion fluency effect on trust but the experimental design does not allow measuring the effect.

   a. The recency effect results in participant making their judgment based on their most recent experiences with the robot. Measurements are done post run. At this time, even in the short condition, the effect of motion fluency might already have faded away. To prevent this from happening real time trust measurements can be done during the experiments. Desai et al. (2013) suggest prompting short trust questions during a task e.g. every 25 seconds. This results in a trust curve that shows the amount of trust over time. A final trust value is then calculated by, as Desai et al. (2013) coined it, area under the trust curve (AUTC) (Desai, Kaniarasu, Medvev, Steinfeld, & Yanco, 2013).

   b. The assimilation effect or similar effects such as the primacy effect result in participants using their experiences of the introduction (first) robot and fit them on their expectations of the experiment (second) robot (Freedy, DeVisser, Weltman, & Coeyman, 2007; Herr, Sherman, & Fazio, 1983). Judgments via the questionnaire in the end are based on, or at least strongly influenced by, the introduction robot. Since the behavior of the introduction robot is constant over all conditions it might explain the lack of difference between the conditions. Making the contrast between both robots bigger would remove the assimilation effect (Herr, Sherman, & Fazio, 1983). Removing the introduction block entirely would make the primacy effect an advantage. In this case the participant can only base their evaluation on experiences with the experimental robot. Since the effect of motion fluency on trust supposedly is only present in the beginning a primacy effect would heighten the change of finding it. A downside is that explaining the task to the participants gets harder. These improvements can be used for a possible next version of the experiment.

c. Participants attribute (parts of) the behavior of the robot to virtual reality. Glitches during the experiments might have interfered the participants' evaluation of the robot. In other words no differences between the conditions are found due to the use of virtual reality. The use of real robots might lead to finding a significant effect of motion fluency on trust (Bainbridge, Hart, Kim, & Scassellati, 2008).

Although no effect of motion fluency on robot trustworthiness was found still a lot of food for thought and further (follow-up) experiments is generated. With the framework provided by Hancock et al. (2011) as basis and with the help of both recent findings, such as a new trust measuring technique (Desai et al. 2013), and widely accepted concepts, such as primacy and recency effects, I believe that research on human-robot trust will make big steps in the near future.

REFERENCES

Adams, B. D., Bruyn, L. E., Houde, S., & Angelopoulos, P. (2003). Trust in automated systems literature review. *Defence research and development* .

Atoyan, H., Duquet, J., & Robert, J. M. (2006). Trust in new decision aid systems. *Franscophone d'Interaction Homme-Machine*, (pp. 115 - 122). New York.

Bainbridge, W., Hart, J., Kim, E., & Scassellati, B. (2008). The effect of presence on human-robot interaction . *The 17th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 701 - 706). Munich, Germany: IEEE.

Chang, L., Doll, B., M., v. '., Frank, M., & Sanfey, A. (2010). Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology , 61*, 87-105.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* New Jersey: Lawrence Erlbaum Associates.

Desai, M., Kaniarasu, P., Medvev, M., Steinfeld, A., & Yanco, H. (2013). Impact of Robot Failures and Feedback on Real-Time Trust. *HRI 2013 Proeceedings (IEEE)* , 251 -258.

Desai, M., Stubbs, K., Steinfeld, A., & Yanco, H. (2009). Creating trustworthy robots: Lessons and inpiration from automated systems. *Artificial Intelligence and Simulation of Behavior: New Frontiers in Human-Robot Interaction.* Edinburgh, Schotland.

Donaldson, T. (1968). Robustness of the F-test to errors of both kinds and the correlation between the numerator and denominator of the F-ratio. *American Statistical Journal* , 660-676.

Fasola, J., & Matarić, M. J. (2012). Using Socially Assistive Human–Robot Interaction to Motivate Physical Exercise for Older Adults. *Proceedings of the IEEE , 100* (8), 2512 - 2526.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems* , 143–166.

Freedy, A., DeVisser, E., Weltman, G., & Coeyman, N. (2007). Measurement of trust in human-robot collaboration. *International Symposium on Collaborative Technologies and Systems* (pp. 106-114). Orlando, Florida, United States of America: IEEE.

Hancock, P., Billings, D., Schaefer, K., Chen, J., de Visser, E., & Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors , 53*, 517-527.

Herr, P. M., Sherman, S. J., & Fazio, R. H. (1983). On the Consequences of Priming: Assimilation and Contrast Effects. *Journal of Experimental Social Psychology* , 323-340.

Iwata, H., & Sugano, S. (2009). Design of Human Symbiotic Robot TWENDY-ONE. *IEEE International Conference on Robotics and Automation* , 580–586.

Karlin, S. (2010, January 29). *Caregiver Robots*. Retrieved June 5, 2013, from IEEE Spectrum: http://spectrum.ieee.org/robotics/humanoids/caregiver-robots

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate relience. *Human Factors* , 50-80.

Lunney, G. H. (1970). Using analysis of variance with a dichotomous dependent variable: an emprical study. *Journal of education measurement* , 263 - 269.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonom* , 1905 - 1922.

Oleson, K. E., Billings, D. R., Kocsis, V., Chen, J. Y., & Hancock, P. (2011). Antecedents of Trust in Human-Robot Collaborations. *International Multi-Disciplinaire Conference on Cognitive Methods in Situation Awereness and Decision Support* (pp. 175 - 178). Miami Beach, Florida, USA: IEEE.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors* , 230-253.

Quick, D. (2010, March 11). *Twendy-One ready to lend a robotic helping hand to the elderly.* Retrieved June 08, 2013, from Gizmag: http://www.gizmag.com/twendy-one-robot-elderly/14496/

Riley, J. M., Strater, L. .., Chappell, S. L., Connors, E. S., & Endsley, M. R. (2010). Situation awereness in human-robot interaction: Challenges and user interface requirements. *Human-Robot Interaction in Future Military Operations* , 173 - 191.

Van den Brule, R., Dotsch, R., Bijlstra, G., Wigboldus, D., & Haselager, W. (submitted). Human-Robot Trust - A Multi-Method Approach.

Welford, A. T. (1980). Choice reaction time: Basic concepts. *Reaction Times* , 73-128.

Wilson VanVoorhis, C. R., & Morgan, B. L. (2007). Understanding Power and Rules of Thumb for Determining Sample Sizes. *Tutorials in Quantitative Methods for Psychology* , 43-50.

APPENDIX

A.1. CHECKS OF NORMALITY

Condition	Kolmogorov-Smirnov p-value	Shapiro-Wilk p-value	Skew Stat(std. dev)	Curtosis Stat(std. dev)
Short-smooth	0.200	0.251	.352 (.464)	-.828 (.902)
Short-trembling	0.166	0.106	.917(.464)	.647(902)
Long-smooth	0.122	0.096	.080(.464)	-1.095(.902)
Long-trembling	0.200	0.837	-.021(464)	-.470(.902)

Table 2: Test for normality that both indicated that the groups belong to a normal distribution

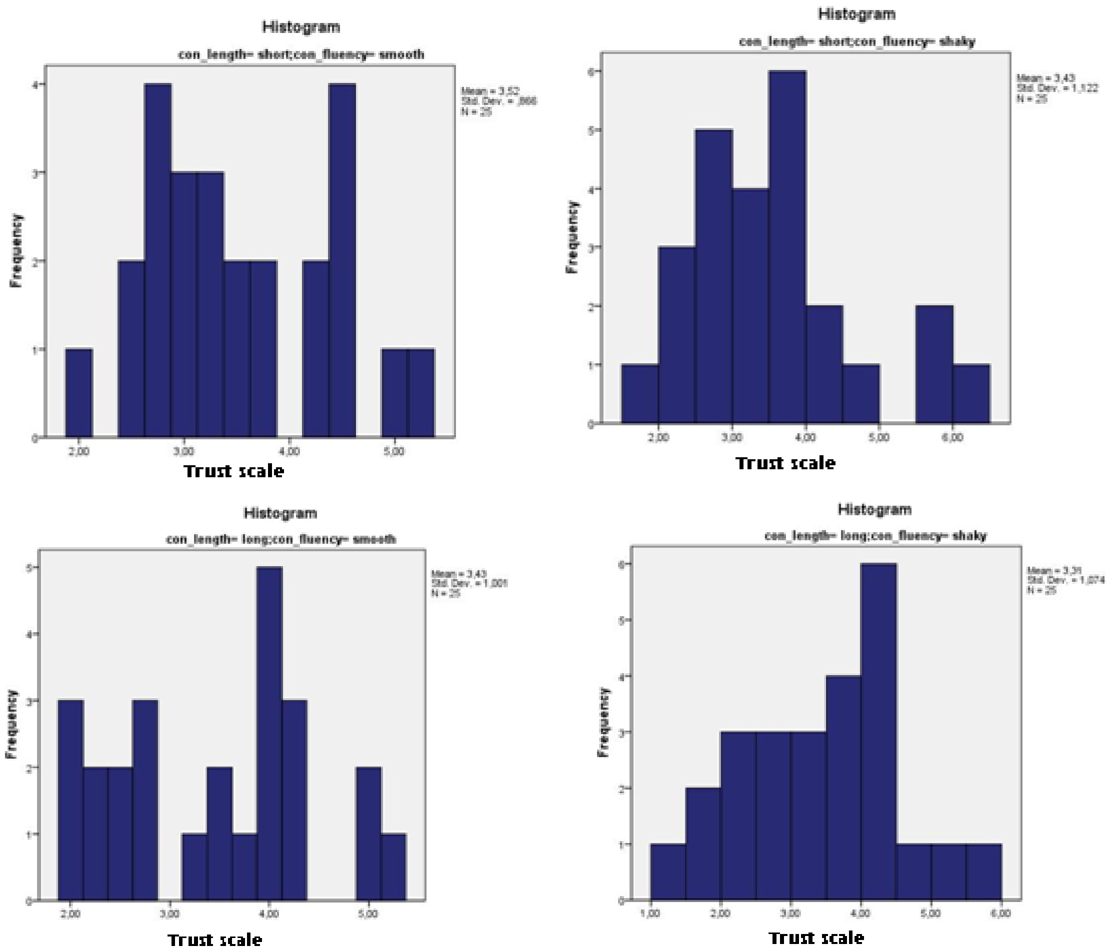

Figure 6: histograms of trustworthiness compound measure plotted per group

A.2. QUESTIONNAIRE (DUTCH)

Questions regarding the experiment

1. In hoeverre kwam de robot betrouwbaar op je over?

   [1 = "in het geheel niet"; 7= "in het geheel wel"]

2. In hoeverre zou je de robot toevertrouwen om ballen te sorteren zonder menselijk toezicht?

   [1 = "in het geheel niet"; 7= "in het geheel wel"]

3. Hoe positief zou je de robot willen beoordelen?

   [1 = "helemaal niet positief"; 4 = "neutraal"; 7= "heel erg positief"]

4. Hoe negatief zou je de robot willen beoordelen?

   [1 = "helemaal niet positief"; 4 = "neutraal"; 7= "heel erg positief"]

5. In hoeverre vond je de robot trillen?

   [1 = "in het geheel niet"; 7= "in het geheel wel"]

6. Geef aan hoe goed de robot de taak heeft volbracht.

   [1 = "helemaal niet goed"; 7= "helemaal goed"]

Demographics

7. Wat is je leeftijd?
8. Wat is je geslacht?
9. Ben je rechts-, links- of tweehandig?
10. Wat studeer je?
11. Is Nederlands je moedertaal?
12. Hoeveel uur speel je gemiddeld per week computergames?
13. Hou je van science fiction (boeken / films / computergames)?
14. Hoe vaak doe je mee met experimenten in het BSI-lab?
15. Heb je eerder meegedaan in een Virtual Reality onderzoek?
16. Hoe goed heb je tijdens het experiment je best gedaan?
17. Hoe leuk vond je het experiment?
18. Heb je eerder aan een versie van dit experiment (van dezelfde onderzoeker) meegedaan?
19. Heb je een strategie gebruikt om de taak voor jezelf makkelijker te maken?
20. Waar gaat dit onderzoek volgens jou over?
21. Heb je opmerkingen over dit onderzoek?