Looks good, sounds nice

Intonation and bodily appearance in robot-mediated communicative treatment for children with autism

MA Thesis C. L. van Straten (3659704)

Date of submission: June 3rd, 2016

Research Master Artificial Intelligence
Study track Cognitive Processing
Graduate School of Natural Sciences
Utrecht University

1st supervisor: Dr. Aoju Chen (UU)
2nd supervisor: Dr. Emilia Barakova (TU/e)
3rd supervisor: Iris Smeekens (Karakter)
2nd examiner: Dr. Chris Janssen (UU)

Universiteit Utrecht

TU/e Technische Universiteit Eindhoven University of Technology

kinder- en jeugdpsychiatrie karakter

Abstract

Children with high-functioning autism (HFA) encounter difficulties with respect to social interaction and communication. Former research has shown that robot-mediated therapy is effective in the behavioral treatment of this disorder. However it has also been shown that not all robots are equally appropriate to be used for this purpose, due to patients' limited processing capacities and selective attention patterns. The current study investigated whether intonation in a robot's voice and humanization of its bodily appearance influence the outcomes of pivotal response treatment (PRT) sessions with children with HFA. The children (age range: 4-8) played a puzzle game with a robot, which required communication. Intonation (normal vs. monotonous) was varied on a within-subject basis, whereas appearance manipulation via human clothing was varied between subjects. Performance was assessed by two types of measurements: required prompting of communicative acts and spontaneous communicative initiatives towards the robot. Additionally, affective states received scoring derived from the individual assessment of demonstrated interest, happiness, and appropriate behavior. No effects on performance measurements were found. The influence of intonation on affective states was nonsignificant, although both general and dimension-specific affect scores tended to be higher in monotonous conditions. Appearance had a significant effect on general affect and interest scores, that were higher in 'human appearance' conditions. Finally, an interaction effect showed happiness scores to be significantly higher when intonation and appearance were either both machinelike (mechanical appearance, monotonous intonation) or both humanlike (humanlike appearance, normal intonation). Together these findings suggest that intonation and bodily appearance indeed influence affective treatment outcomes, whereas their influence on performance could not be confirmed.

Keywords: children - high-functioning autism - pivotal response treatment - robot-mediated therapy - socially assistive robotics - intonation - appearance - NAO

Acknowledgements

The report laying in front of you constitutes the outcome of huge a learning experience. Over the past months that I have been working on this project, my initial ideas developed into a full grown scientific research that after two years of studying will conclude my master's education. More importantly even, it concludes six years of being a student at Utrecht University. Looking back this time period reflects an exploration of my major interests, and combining the disciplines of linguistics, artificial intelligence and cognitive science, the research project reported in this paper comprises all of them in a way I could only hope for up until November last year.

Now it is about time to thank the people that have enabled me to actually reach this finish line. First and foremost I would like to express my sincere gratitude to Dr. Aoju Chen, who as my first supervisor has offered generous guidance from the day I told her about my abundant number of - more and certainly also less qualitative - research ideas to the very end of this learning process. Beyond all of my expectations, she helped me to develop and conduct this study by sharing her expertise and time as well as by providing mental coaching whenever I panicked or looked alarmingly pale. The second person I would like to thank is Iris Smeekens, who welcomed me in the Picasso project and provided daily support during my time at Karakter Nijmegen. Initially introducing me to NAO, and ultimately introducing her clients to my experiment, she has been indispensable in the realization of this project. And then of course, I owe a big "thank you" to Dr. Emilia Barakova who provided me with the wonderful opportunity to join the Picasso project in the first place. Would she not have accepted my internship request and facilitated this study, who knows what paper you would currently be reading instead.

I would also like to thank those on the sidelines. Such as David, for supporting me in many many ways whose specification would take the word count of this report to a whole new level. And my parents, who always showed great interest in the progression of my project and throughout my studies answered each "I can't do it" with a "yes you can". Finally, thanks to family and friends I discussed my ideas with and whose differing points of view provided me with new insights along the ride. I hope you will have a good read.

Table of contents

1 Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that is characterized by a range of deficits with respect to social interaction and communication. Moreover, individuals with ASD have a restricted number of interests and activities in which they like to engage. The experiment reported here focuses on children with high-functioning autism (HFA). High-functioning, as opposed to low-functioning, stands for the absence of mental retardation and resulting problems regarding the development of cognitive skills (American Psychiatric Association, 2013). Skills that children with HFA generally do encounter problems with are, among others, sharing information and feelings, requesting information, responding to others in their surroundings and cooperating with them (Macintosh & Dissanayake, 2006). Whereas HFA and Asperger Syndrome (AS) are sometimes treated as one and the same disorder, AS distinguishes itself from HFA by the presence of a delay in language development (American Psychiatric Association). Since not all former studies distinguish between different forms of ASD in their participant samples, both studies concerning children with HFA and studies concerning children with ASD in general will be reviewed. Studies that specifically examined children with low-functioning autism (LFA) or AS will not be considered.

Within the field of Socially Assistive Robotics (SAR), researchers investigate behavioral training methods for children with ASD that make use of robot-mediated social interaction. Multiple successful outcomes have been achieved. For instance, children with ASD have been shown to benefit from therapy sessions in which they are allowed to freely communicate with a robot. Such sessions were demonstrated to provoke proactive social behavior in these children, leading to an increase in the number of verbal utterances they produced (Feil-Seifer & Matarić, 2009). They moreover have been found to be more responsive to feedback that is provided through technology, which resulted in quicker acquisition of target skills than similar feedback they received from human therapists (Charlop-Christy, Le & Freeman, 2000). The success of robot-mediated treatment for children with ASD may be due to the fact that technological objects are intrinsically appealing to these children. Their great interest in this topic provides them with a high level of motivation to interact with the robots they are presented with during therapy sessions (Scassellati, 2007; Diehl, Schmitt, Villano & Crowell, 2012). Furthermore, robot behavior has its limitations and is therefore far more predictable than the complex behavior that human beings exhibit. This has been shown to be another attractive property to children with ASD (Flannery & Horner, 1994). Lastly, when interacting with a robot, these children have the feeling that they are able to control the situation - which forms one of their strong desires (Dautenhahn & Werry, 2004; Frith, 1979).

Although robots have been proven to be a welcome supplement to traditional treatment methods, not every robot seems to be equally appropriate to be used to this end. That is, research has shown that children with HFA perform better during interaction experiments with doll-like robots that do not possess a humanlike appearance than during identical experimental sessions in which the same robots do possess such an appearance. An explanation for this difference in performance is that the minimal number of features of the non-humanlike-looking robot makes its appearance far less complex to process (Robins, Dautenhahn & Dubowski, 2006).

This is an important advantage for children with ASD, whose senses are easily overstimulated (Johnson & Myers, 2007). Nonetheless, while the authors ascribe the preferability of the robot's minimal appearance mainly to the absence of facial features, other studies suggest that bodily appearance should be taken into account as well. It was for instance demonstrated by Billard, Robins, Nadel and Dautenhahn (2007) that children with ASD react positively to a doll-like robot that possesses a mechanical looking body but a humanlike head. While Robins et al. reported a tendency among participants to avoid looking at the robot's face when it contained humanlike features, participants in the study by Billard et al. showed such avoidance behavior to a lesser extent. Moreover, it has been confirmed that the robot used during the latter research provokes comparable responses from children with ASD as does a non-humanoid, completely mechanical looking robot (Billard, 2003). Given the additional findings that face processing primarily forms a problem for children with ASD in case of faces that are moving at a natural speed (Gepner, Deruelle & Grynfeltt, 2001), as well as that these children tend to pay increased attention to mechanical body parts of robots (Kozima & Nakagawa, 2006), it seems that the usability of various robots more strongly depends on the presence of mechanical features than on the absence of human ones. This is in accordance with the recommendations made by Cabibihan, Javed and Ang (2013), who suggested that the appearance of robots to be used in ASD treatment programs should ideally contain both human and mechanical elements.

Whereas attention has been paid to the different ways in which children with ASD react to robots with various appearances, to the best knowledge of the author no research has yet investigated the effects of a robot's vocal features on the successfulness of its interaction with these children. Strikingly, studies such as the above-mentioned pay no attention to differences between various robot voices and hence seem to assume their irrelevance. An ample amount of research however has addressed abnormalities exhibited by the auditory processing capacities of children with ASD. For instance it has been reported that as compared to a typically developing (TD) control group, children with HFA demonstrate superior pitch discrimination capacities (O'Riordan & Passetti, 2006). This could partly be explained by the finding that a selection of individuals with HFA possesses absolute pitch (Miller, 1999). Nevertheless, there seems to be an additional explanation that is worth our consideration. That is, another study that reported the same superior performance regarding pitch discrimination interestingly also found that, while the performance of control subjects was impaired when the stimuli to be discriminated consisted of speech fragments instead of musical segments, children with HFA performed equally well across conditions. The authors suggest that this difference in performance is due to an increased focus on perceptual elements of speech in children with HFA, whose discriminatory abilities are thereby unaffected by the semantic content of the speech stimuli that forms a distraction to the controls (Jarvinen-Pasley & Heaton, 2007). A follow-up research confirmed this ignorance of semantic features, showing that improved perceptual abilities go hand-in-hand with impaired comprehension task performance (Jarvinen-Pasley, Wallace, Ramus, Happé & Heaton, 2008). These results are consistent with the idea that children with ASD are only able to respond to a limited number of simultaneously presented cues, showing overly selective attention to only some of the features of a presented stimulus (Lovaas, Koegel & Schreibman, 1979). Additionally, it was previously demonstrated that this overselectivity indeed not only appears between semantic and acoustic aspects of auditory input, but also between its various perceptual aspects. Namely, children with ASD were found to selectively attend pitch-related elements of auditory input over its other perceptual components such as rhythm and stress (Reynolds, Newsom & Lovaas, 1974).

Brain imaging research on adults with HFA has come up with a possible explanation for this deviant focus on perceptual over semantic elements of speech. That is, a failure in these individuals was found to exist with respect to the inhibition of prosody-related brain areas (Hesling et al., 2010). Moreover it was demonstrated that functional connectivity between cortical areas relevant to language processing is decreased in adults with HFA (Just, Cherkassky, Keller & Minshew, 2004). Finally, yet another brain imaging study found that in adolescents with HFA, semantic comprehension based on perceptual disambiguation elicited general instead of prosody-specific brain activation (Eigsti, Schuh, Mencl, Shultz & Paul, 2012). Together these outcomes show that individuals with HFA perform well on tasks that merely involve perception, but encounter difficulties when it comes to the cognitive integration of various types of linguistic information during performance on high-level tasks.

The current study aims to fill the gap in our understanding of the effects of intonation, a key vocal feature, and bodily appearance of robots on treatment outcomes for children with HFA. Intonation is here defined as variations in pitch that stretch over units larger than segments (e.g. vowels and consonants) and convey meaning at the utterance-level. The reported experiment was designed to test the hypothesis that intonation and bodily appearance of robots influence treatment outcomes of robot-mediated therapy sessions for children with HFA. This hypothesis comes with the following three predictions.

First, based on above-mentioned findings on deviant attention patterns and integration difficulties, it was predicted that using a robot with monotonous intonation (i.e. pitch variation is removed but temporal variation is maintained) will lead to better treatment outcomes than using a robot with normal intonation. Better outcomes are taken to be reflected in improved interaction task performance and more positive affective states in children during treatment sessions. It is expected that when pitch variation is removed from the speech signal, perceptual distraction will decrease and attentional focus will shift towards the semantic content of the message. Therefore participants will presumably become aware sooner of the interactive act they should engage in, such that performance will improve. Furthermore, the absence of pitch-related information leads to a situation in which less individual features of the speech signal will have to be integrated during processing, which poses less cognitive pressure on the children. This might affect both performance and affective states. A monotonously sounding voice moreover emphasizes a robot's technological nature to which children with HFA naturally feel attracted, which might further improve the latter. While one could argue that intonation normally functions as one of the aspects of prosody that aid listeners to foresee upcoming information, hence making a discourse more predictable (Snedeker & Trueswell, 2004), this is not expected to be the case for children with HFA since using perceptual information in this way requires high-level information integration. Moreover, prosodic features other than pitch variations, such as rhythmic properties, remain present in the signal. Removal of these features is assumed to be unnecessary, based on the aforementioned sensitivity to pitch patterns as well as the overselectivity of attention towards pitch-related information over other low-level aspects of an auditory signal.

The second prediction addresses the robot's looks, stating that humanizing a robot's bodily appearance will result in better treatment outcomes (as measured by interaction task performance and affective states) compared to when the same robot is being presented to the children in its regular appearance. While additional specification and motivation regarding the exact manner of manipulation will be provided in section 3.2, for now it is worth noting that changing the robot's bodily appearance instead of facial appearance mirrors the situation tested by Billard et al., who humanized a robot's facial appearance while leaving its body looking mechanical. The expectation regarding improvement of performance is motivated by the aforementioned finding by Kozima et al. that children with ASD tend to pay more attention to the mechanical body parts of a robot than to the interaction task they are bound to perform with it. Once this distraction will be decreased, it is expected that attention will shift towards the ongoing conversation such that communicative performance will be improved. Affective states respectively are expected to enhance based on the following line of research. It has been found that TD children and children with HFA very comparably classify robots. When judging from their appearance based on a picture, both groups generally perceive the robot to be used in the current experiment (NAO, see section 3.2) as some kind of toy (Peca, Simut, Pintea, Costescu & Vanderborght, 2014). However, when TD children engage in real life interaction with this robot, they tend to classify it as something hybrid between a peer and a toy (Nalin, Bergamini, Giusti, Baroni & Sanna, 2011). As their picture-based judgements were highly comparable to those of TD children, the same presumably counts for children with HFA. Their perception of the robot as a peerlike entity is furthermore expected to be strengthened by the explicit humanization of its bodily appearance. Since mechanical aspects remain present, this allows for an approach of the vividness of a human interaction partner while the appealing technological, predictable and controllable aspects of the robot remain evident. This situation might lead to a higher level of engagement and, as a consequence, to more positive affective states.

Finally, the third prediction states that optimal treatment outcomes will be obtained when intonation and bodily appearance are congruent. Such congruence is reached in two situations: when both emphasize the robot's mechanical nature, and when both express its humanlikeness. This implies that ultimately, the robot should either be presented in its regular mechanical appearance with a monotonously sounding voice, or be presented in its humanlike appearance with a voice that contains normal intonation. Once more, better outcomes are understood as improved interaction task performance and enhanced affective states. The motivation of this last prediction partly lies in the aforementioned finding that children with ASD are attracted to predictability (Flannery et al.). Moreover, a brain imaging study has previously found that individuals with ASD show increased audio-visual congruence effects as compared to a TD control group. In this study, participants watched animal pictures while simultaneously being presented with animal sounds that either did or did not match the depicted animals. The increased sensitivity in participants with ASD to audio-visual congruence was assigned to enhanced perceptual functioning (Russo, 2007). While a match between intonation and bodily appearance in the current experiment may be less evident than the matches made by Russo, it is expected that individuals with HFA will be sensitive to this type of congruence as well.

Note that the consideration of both performance and affective states adds to the validity of the present research. That is, in order for results to be scientifically relevant they should be applicable in the field of ASD treatment. It is therefore important to know how participants appreciate the various situations, since a very effective method that has a negative effect on the child's emotional well-being will probably not be practiced. Conversely, adjustments that do not influence participants' performance but do increase their level of enjoyment would be welcome variations to existing treatment programs. The experiment reported here was embedded in a larger research project. In what follows, relevant background information regarding this project and the ASD treatment method it implements will be provided. Thereafter, the experimental set-up will be discussed in detail. Following a section in which obtained results will be reported a general discussion will subsequently conclude this report, while making several suggestions for future research.

2 Background

2.1 The Picasso project

The current study was conducted as part of the Picasso project (short for 'PRT-based Intervention with robots for training Children with AutiSm in SOcial and communicative skills'). This research project hence relied on a cooperation between Utrecht University and non-UU partners, including Karakter, a Dutch centre for child and adolescent psychiatry, the department of Industrial Design at the Technical University Eindhoven (TU/e), and the department of Cognitive Neuroscience at the Radboud University Nijmegen Medical Centre (Radboud UMC). The main objective of the Picasso project is to compare various treatment methods for children with HFA. Previous research has demonstrated that both the method of pivotal response treatment (PRT) (more on this in section 2.2), and the incorporation of a robot in treatment programs are effective means to train social communicative abilities in children with ASD (Barakova, Bajracharya, Willemsen, Lourens & Huskens, 2014; Huskens, Palmen, Van der Werff, Lourens & Barakova, 2015; Huskens, Verschuur, Gillessen, Didden & Barakova, 2013). In the Picasso project, three groups of children with HFA each receive a different type of treatment. One group only receives usual care, consisting of psychoeducation and medical management, the second group is provided with PRT guided by a human therapist in addition to that usual care, and the third group receives robot-mediated PRT in addition to usual care. Both objective performance improvement, qualitative reports of children and their caretakers, and biological measurements such as cortisol levels in salivary samples are taken into consideration, to ultimately provide an answer as to which method would be optimal in the treatment of HFA. All treatment methods are tested in a naturalistic treatment setting to ensure that the effects to be found can equally be expected to arise in the outside world. During the PRT sessions, children play games with either a robot (NAO, section 3.2) or their therapist. Various games (e.g. card games, puzzle games, LEGO games and board games) are currently being developed, each game coming in multiple themes between which children may choose. This allows for the adaptation of treatment sessions to their personal interests, which provokes motivation and enjoyment. Other than stimulating a natural and motivating style of interaction, the games provide some structure and predictability for the children by means of their rules. Each child will receive five months of therapy during which it will be assessed whether generalization of skills acquired during the sessions appears in his or her everyday environment. Three months after the last Picasso session, a follow-up study will be conducted to verify whether treatment effects have translated themselves into long lasting benefits for the children in question.

2.2 Pivotal response treatment

From the moment they are born, children direct themselves towards other human beings in their environment; they are intrinsically motivated to establish contact. A lack of such motivation during childhood has been suggested to explain many problems encountered by individuals with ASD. That is, in the absence of motivation a decrease in (or even complete absence of) communicative initiatives performed by children suffering from this disorder manifests itself. This in turn results in a shortage of naturally occurring social learning opportunities (Chevallier, Kohls, Troiani, Brodkin & Schultz, 2012).

During PRT, such opportunities are therefore actively established by the use of techniques finding their origins in the field of behavioral therapy. This treatment method addresses four central (pivotal) skills: motivation to interact, initiation of communication, responsivity to multiple presented cues and self-management. When a child makes progress with respect to one of these core areas this should automatically lead to improvements of other abilities as well, such as making eye contact and sharing enjoyment (Koegel, Koegel & McNerney, 2001). Learning opportunities are established using the A-B-C format (Koegel & LaZebnik, 2004), in which the capitals stand for Antecedent, Behavior, and Consequence. The idea is that the therapist first follows the child's interest, than captures his or her attention and subsequently establishes a learning opportunity while providing help to guide the child in the desired behavioral direction. This sequence of actions constitutes the antecedent. The child's response respectively forms the behavior component, on which the consequence component depends. In case the child does not show the intended communicative behavior, the child can be encouraged to do so by means of various prompts. In the Picasso project, these are open question prompts ("What could you ask/tell me now?"), finish prompts (e.g. "May I…") and repetition prompts (e.g. "May I have the puzzle pieces?"). They are provided during learning opportunities in the aforementioned order of increased explicitness. Once the child has responded appropriately, or has at least seriously attempted to do so, direct acknowledgement follows in the form of a natural reward. It is of great importance that the obtained reward is indeed a natural one, which as Williams, Koegel and Egel (1981) showed leads to quicker skill learning. In their study, children with ASD learned faster how to open a box when correct attempts were rewarded by a piece of candy located inside the box, compared to when the candy was provided by the therapist. The naturalness of PRT furthermore adds to its value by allowing for easier generalization of behavioral skills learned through therapy sessions to situations occurring in the outside world

(Koegel & Koegel, 2006). The PRT sessions conducted during the Picasso project mainly address communicative abilities, which can be specified as the abilities to communicate intentions, request help or information, protest, maintain a conversation, and place remarks.

3 Method

The participants in the current study all belonged to the group of children that receives robot-mediated PRT in the Picasso project. Their Picasso treatment sessions were adapted in content such that these constituted the trials of the reported experiment. However, to allow for optimal progress of the children in the Picasso project itself, sufficient interaction task variation between treatment sessions was of crucial importance. Hence only a limited number of a child's Picasso treatment sessions could be dedicated to the current experiment, whose trials had to be identical. Therefore it was decided to adopt a mixed design, with intonation as a within-subject factor and appearance as a between-subject factor. Participants were divided into two experimental groups. Group 1 encountered the robot in two conditions: one in which its voice contained normal intonation and the robot was presented in its usual mechanical appearance, and one in which it again appeared as usual but its voice contained monotonous intonation. Group 2 was presented with both intonation conditions as well, however this group encountered the robot in its humanlike appearance (see section 3.2). Hence there were four conditions in total, whose order has been counterbalanced between participants.

3.1 Participants

Each experimental group consisted of four children who have been clinically diagnosed with ASD according to either the fourth or fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM) (American Psychiatric Association), or the International Classification of Diseases and Related Health Problems (World Health Organization, 1993). All diagnoses were additionally confirmed by the Autism Diagnostic Observation Schedule (De Bildt, Greaves-Lord, & De Jonge, 2013). None of the participants (6 male) had intellectual disabilities, reflected by IQ scores over 70 ($M = 98.13$, $SD = 11.93$) as measured by either the Wechsler Intelligence Scale for Children (WISC) (Kort et al., 2005), the Wechsler Preschool and Primary Scale of Intelligence (WPPSI) (Hendriksen & Hurks, 2009), or Mullen Scales of Early Learning (MSEL) (Mullen, 1995). All fell within the chronological age range of 4 to 8 years old ($M = 6.70$, $SD = 1.52$), which adds to the scientific relevance of the current research. Namely, as early communicative treatment proves crucial for reaching optimal results later in life, the effectiveness of different forms of treatment should be tested during early childhood (Kasari, 2002). Parental consent was given prior to the start of the experiment. During their participation, the possible medical dose that participants received was fixed. All were native speakers of Dutch with no history of audiovisual or additional neurological deficits, and all had previous experience with NAO (see section 3.2).

Detailed information regarding individual participants as well as participant groups has been provided in table 1 and 2 below. The two participant groups were matched for chronological age as well as verbal IQ (VIQ) scores, such that influences of language development were ruled out. The between-group difference in mean VIQ score has been checked for significance. As the data proved to be normally distributed for both participant groups, $W(4) = .899$, $p>.05$ and $W(4) = .828$, $p>.05$ respectively, and their variances proved to be equal, $F(1,6) = .395$, $p>.05$, an independent samples t-test was performed that showed the difference to be nonsignificant, $t(6) = .360$, $p>.05$. While the number of participants per group is usually desired to be larger, the majority of research that has been conducted regarding the use of robots in ASD treatment methods as well as those regarding PRT for children with ASD equally tested no more than three to four participants (Diehl, Schmitt, Villano & Crowell, 2012; Scassellati, Admoni & Matarić, 2012; Verschuur, Didden, Lang, Sigafoos & Huskens, 2014). Whereas this certainly forms a methodological limitation, due to strict ethical regulations and a scarcity of families that were willing to let their child participate in the Picasso project as well as to commit to the accompanying time investment, it is one that will have to be tolerated.

Table 1: Participant details

Participant ID	Experimental group	Gender	Chronological age	VIQ score
110001	1	M	4:04	91
110003	1	M	8:02	101
110005	2	M	7:09	90
110014	1	M	6:05	114
110015	1	F	7:11	93
110019	2	M	8:02	93
110026	2	F	5:02	85
110030	2	M	5:07	118

Table 2: Experimental group details

	Mean chronological age	Mean VIQ score
Group 1	6:8	99.75
Group 2	6:8	96.50

3.2 Apparatus

The robot that was used during the current experiment is the humanoid robot NAO (V4, fig. 1), equipped with the NaoQi 1.14.1 software development kit (Aldebaran Robotics). NAO is 58 cm tall and has 25 degrees of freedom that allow for flexible, humanlike movements. It has simple features, making it appropriate for usage with children with ASD. The robot's speech is produced via text-to-speech synthesis, realized by a TTS-module exclusively available on the NAO robot and based on a speech synthesizer provided by the Acapela Group. Integrated speakers allow sound to be coming from exactly the right direction. The experiment was programmed and executed on an ASUS Zenbook UX31A (OS Windows 8.1) using Tino's Visual Programming En-

Figure 1: NAO (Aldebaran)

vironment (TiViPE) 2.1.2 (Lourens & Barakova, 2011), which has been proved to work properly during earlier research (Huskens et al., 2013). Trials were videotaped by use of a Sony Handycam HDR-CX240E video camera, mounted on an adjustable tripod stand.

It is known that children with autism are reluctant to change, a characteristic that was first described by Kanner (1943) as their ‘need for sameness’. Therefore using a familiar voice (i.e. female voice ‘Jasmijn’, used in the Picasso project) in normal intonation conditions as opposed to an unfamiliar voice in monotonous conditions could have been of influence on the results. Hence it was decided to use an unfamiliar female voice in normal intonation conditions, as a more radical change to a male voice was prohibited by the protocol of the Picasso project. Most engaged therapists being female, this minimizes the difference between robot-mediated and therapist-guided PRT sessions. One alternative Dutch female voice could be selected for NAO, however this voice sounded less natural and was less intelligible than voice Jasmijn. Therefore, voice Jasmijn was manipulated by changing its mean pitch from 166 Hz to 232 Hz. This alteration made it sound notably different than in previous sessions, without rendering it unnatural or less intelligible. A positive side effect of the manipulation was that the increased pitch made the voice more child-directed than it originally was. The voice used in monotonous conditions was subsequently created as follows. First, all of NAO’s utterances (using the adjusted version of voice Jasmijn) were recorded with an H1 Handy Recorder (Zoom) positioned on an adjustable tripod stand. The resulting recording was manipulated with Praat 6.0.12 (Broersma & Weenink, 2015). All pitch points were removed except for one that equaled the average pitch of the recording (i.e. 232 Hz), hence indirectly setting the pitch range to 0. Since playing the original recording via the robot resulted in a severe loss of volume, its volume was amplified three times to match the volume of the text-to-speech output in normal intonation conditions. The recording was then cut into one-sentence utterances using Audacity 2.0.6, that were uploaded onto the robot and invoked in the game scripts such that all text-to-speech commands were replaced. Note that the difference between the voices used across intonation conditions was purely reflected by intonation, while other prosodic features such as rhythmic properties were preserved.

NAO’s bodily appearance was humanized by covering his[1] body parts in human clothing. To this end NAO wore a long sleeved T-shirt and a pair of trousers, that entirely covered his torso and limbs. The colors of both garments were neutral ones that could equally well please boys and girls. While the trousers were originally white, they were painted orange to avoid transparency. A picture of NAO’s humanlike appearance has been included below (fig. 2). The fact that his bodily instead of facial appearance has been manipulated during this study adds to its scientific relevance, due to the position of humanoid robots with respect to the Uncanny Valley (Mori, 2012, fig. 3). The Uncanny Valley illustrates the range of robots that appear very humanlike but are not realistic enough to be indistinguishable from humans. These imperfections may provoke feelings of uncanniness in those who interact with such robots. Research has

[1] The reader might at this point signal a gender discrepancy between the use of male pronouns and the female voice that the robot was equiped with. Motivation for making reference to NAO as if it was a boy can be found in the fact that participants in both the pilot sessions and the experiment itself indicated to perceive him as such. Where adults noticed the robot’s female voice and hence tended to classify NAO as a girl, this did not count for the children. As it was decided to follow their natural perception, NAO was referred to by use of male pronouns during the pilot and experimental sessions.

confirmed that, from the perspective of TD children, NAO locates himself on Mori's curve right before the Valley (Woods, Dautenhahn & Schultz, 2004). That is, NAO's appearance differs just enough from that of human beings to provoke optimal affinity instead of uncanny feelings. Moreover, it has been found that adults with ASD are more sensible to the negative effects of the Uncanny Valley than controls (Destephe, Zecca, Hashimoto & Takanishi, 2014). Since as was previously mentioned, robot perception in TD children and children with ASD is very alike, and taking into consideration this sensibility in adults with ASD, a more radical humanization by manipulation of the robot's facial appearance would be undesirable.

Figure 2: Humanlike appearance

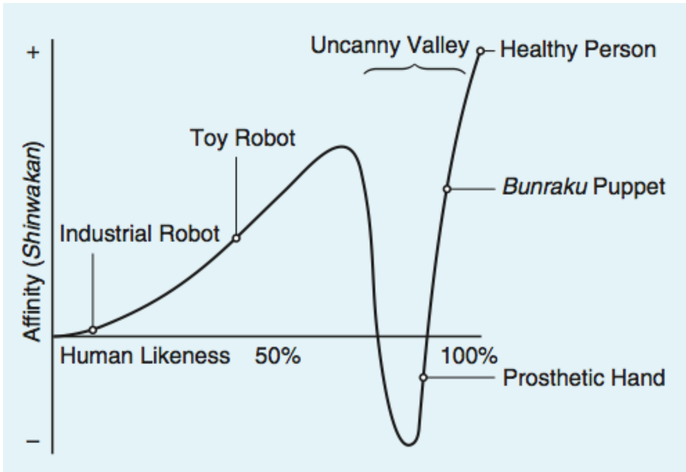


Figure 3: Graphic representation Uncanny Valley (Mori, 2012)

3.3 Experimental task

The experimental task consisted of a puzzle game that was played with NAO and that has also been used during regular treatment sessions. During the game, participants completed three animal-shaped puzzles of their choice. That is to say they were presented with three boxes that each contained three animal puzzles, and they could choose one box. Crucial to the game was the fact that the child could not complete it on its own - to this end it needed to cooperate with NAO. For instance, at the beginning all the puzzle boxes were contained in a larger box that the child could not open by itself. In order to access the puzzles, interaction with NAO had to take place. When access to the box had been gained, each smaller box appeared to contain only one piece of each puzzle. NAO kept the rest of the pieces as well as the pins to attach them, until the child would ask for them correctly. Once a puzzle had been completed the robot would make the sound of the animal in question, but only if the child succeeded to engage in the right interaction. Throughout the game additional learning opportunities were provided with respect to not only functional, but also purely social interaction such as showing interest and maintaining a conversation. Appendix 1 contains commented pictures that illustrate the game's various stages, as a further clarification of its contents.

As the game had already been played by other Picasso participants, there existed awareness of some malfunctions and room for improvement. For instance, therapists previously had great difficulty controlling NAO to perform the right actions at the right time when responding to a child during a session. Therefore the program to be executed was adjusted to allow for easier controlling, using preprogrammed prompts and an adapted interface. Second, children often did not properly hear NAO's utterances. To make sure therapist intervention could be limited to a minimum, the possibility was implemented to make NAO repeat his utterances by adding this option to the interface. The intelligibility of the utterances themselves was also improved, by making use of special characters in their text-to-speech files that enhance pronunciation and improve prosodic features such as rhythm and stress. Other fixed malfunctions were of varying nature. For instance, during previous sessions NAO always told the child "The puzzle pieces are in the red bag", whereas there were multiple colored bags to which NAO should refer depending on which puzzle the child decided to make. To be able to react adequately and quick on the child's choice, key-press selection menus were implemented via which the therapist could easily control NAO's utterances at such points of variation. Furthermore, NAO's voice (Jasmijn) as well as its speed, pitch, and volume parameters were previously only specified at the beginning of the game script. Due to an unknown error, they would sometimes switch over utterances later in the game. Therefore, it was decided to separately specify all voice parameters for each individual text-to-speech utterance.

Finishing the game and hence completing a trial resulted in a total child-robot interaction time of 15 minutes on average. Participants completed one trial in each intonation condition (normal and monotonous) while appearance was kept constant. Both trials were separated in time by a one-week break in order to stick with the existing treatment schedule. Due to personal circumstances of one participant, an exception was made as a two-week break was allowed for. In both trials the game was identical, however each time a different box of animal puzzles could be selected by the child in order to keep the sessions engaging and maintain the important motivational aspect of personal contribution. The level of the interaction task (difficulty 1 to 7) was selected according the needs of each individual participant, such that their developmental training sequence would not be disturbed as well as to ensure comparability of performance between subjects. The contents of the game did not differ between levels. Adjustments only occurred in the formulation of NAO's utterances, and the way in which participants were required to respond. Lastly, it is worth noting that while a single trial per condition is very little, time limitations as well as treatment requirements prohibited further testing.

3.4 Measurements

In order to assess both differences in performance and variations in affective states of participants across conditions, four types of measurements were taken into account. The first one constituted the number of prompts a participant received, expressed as a percentage of the total number of available prompts during a trial. This measurement reflects the required amount of stimulation for participants to engage in the envisioned interaction with NAO during learning opportunities. Three prompts were available during each opportunity: an open question prompt, a finish prompt and a repetition prompt. They were always presented in this order. If a reaction was given without a need for prompting, the number of prompts received during that opportunity was set to zero. In case no reaction was given even after all three prompts had been provided, the trainer stepped in. Such intervention was counted as an extra prompt, hence a maximum of four prompts could be received. Responses to learning opportunities that differed from the envisioned reactions were considered to be successful only when they fitted the conversational context and formed an appropriate attempt or alternative to the envisioned reaction. If the predictions regarding performance were correct, less prompting should be needed in the presumably preferable conditions, leading to lower average percentages of required prompts. Note that no distinction was made between situations in which the trainer intervened and either did or did not receive a reaction from the child. This decision was made since participants' reactions to the trainer were under no direct influence of any of the robot's features. Percentages were calculated instead of means, as the experimental task was presented to participants at various levels that did not all contain the same number of learning opportunities.

A second type of measurements that assessed performance was the number of spontaneous communicative initiatives by a child during a session. This number should be higher in assumed to be preferable conditions. A distinction was made between verbal and nonverbal initiatives. Furthermore, it was taken into account whether an initiative was directed towards the robot or not. Initiatives could be of varying nature, including questions, remarks, imitations, communicative gestures and touching the robot, but always had to be situation appropriate. This category of measurements was added to the experimental design after conducting a pilot study, as will be described in section 3.6. Note that Tapus et al. (2012) equally distinguished between self-initiated and prompted verbal interaction, and that touch and imitation are two communicative acts that were also taken into account by Robins et al.

Participants' affective states were assessed by use of the Affect Rating Scale that has been developed by Koegel and Egel (1979) to score a child's affective state during a PRT session. The scale runs from 0 to 5 and can be employed to appreciate affective states on three dimensions: demonstrated interest, happiness and appropriate behavior. It implicitly also addresses the extent to which a child is sharing enjoyment and establishing joint attention, which are two tendencies that are weaker in children with HFA as compared to TD children (American Psychiatric Association; Mundy, Gwaltney & Henderson, 2010). For the dimension of interest (towards the game as a whole, i.e. not solely towards the robot), a more detailed description of observations that could lead to a certain score was provided by Koegel, Singh and Koegel (2010). This description was used and taken as a guideline for scoring the other two dimensions, such that less was left to personal interpretation. During analysis, trials were divided into one-minute segments that each obtained separate scores on the three dimensions in question. By averaging the scores for the individual segments, four mean scores were ultimately obtained: separate interest, happiness, and behavior scores, plus a general affect score that averaged over dimensions.

Finally, a questionnaire was used that assessed participants' appreciation of the robot as well as their affective states right before and after the start of the game. The questionnaire has been developed and used during the Picasso project and contains three questions: (1) How happy are you right now? (before the game), (2) How happy are you right now? (after the game), and (3) How do you like the robot? (after the game). Participants answered the questions on a 5-point Likert scale, in which verbal answer options are complemented by colored illustrations to clarify their meanings (see appendix 2). After the child picked an answer on the scale, the therapist asked for a motivation of its choice. Both the questionnaire and the Affect Rating Scale should show more positive outcomes in the presumed to be optimal conditions in case the earlier listed predictions are borne out.

While previous studies regarding appearance have measured aspects like gaze direction and duration (Billard et al.; Robins et al.; Shamsuddin et al., 2012), no such measurements were included in the current study. This decision was made since, for exact and objective gaze measurements to be obtained, an eye-tracker should have been used that precisely follows the movements of participants' pupils. The eye-tracker should have been positioned right in front of the child, at the eye level. However, this was prohibited by the experimental set-up in which the robot already occupied this position (see fig. 4 and 5 below). The use of eye-tracking equipment would furthermore have required participants to sit relatively still at a fixed distance from the eye-tracker. One can hardly ask this from children in the current age rage without using a head-fixed set-up, that would disturb the naturalistic treatment setting of the trials. Moreover, many children with ASD suffer from abnormalities in the somatosensory system. These can lead to extreme oversensitivity in response to tactile stimuli (Güçlü, Tanidir, Mukaddes & Ünal, 2007), which would have made the head-fixed set-up extremely unpleasant. Fortunately, previous research using NAO for interaction experiments with children with ASD has found a correlation between gaze frequency and duration towards the robot's face on the one hand, and smiling frequency and duration on the other (Tapus et al.). Since facial expressions were addressed by means of the Affect Rating Scale, gaze measurements might not have been crucial in the first place.

Figure 4: Experimental set-up, start of trial

Figure 5: Experimental set-up, halfway through trial

3.5 Pretests intelligibility

To ensure the absence of qualitative differences between the voices used in normal and monotonous intonation conditions, their intelligibility has been pretested. First, five subjects that were unfamiliar with NAO's speech (all female, mean age 41.8, no history of hearing deficits) scored the intelligibility of ten utterances and their audio recordings in order to check for quality loss related to the recording procedure. None of the subjects had any interest in the outcomes of the pretest. The order of conditions (text-to-speech versus audio recordings) was counterbalanced between participants and volume was matched between conditions. The scoring form contained one scoring table on each side, such that participants were not able to compare or revise their previous answers once the scoring of the secondly encountered condition began. Scoring happened on a 5-point Likert scale running from 1 (completely unintelligible) to 5 (completely intelligible). Participants were asked to place an X in the cell of the table corresponding to their judgement. It was stressed that only intelligibility should be taken into account, such that comprehension or preferences of any other kind would not influence their responses. Utterances in the text-to-speech condition received a mean score of 4.08 (SD = .29) on the intelligibility scale, whereas audio recordings were assigned a score of 4.02 (SD = .50) on average. Normality tests showed that data in both the speech condition, $W(5)$ = .942, p>.05, and the recordings condition, $W(5)$ = .888, p>.05, were normally distributed. A paired samples t-test subsequently showed that intelligibility scores across conditions did not differ significantly, $t(4)$ = 0.408, p>.05.

Five different subjects meeting identical requirements (4 female, mean age 47.2) scored the intelligibility of the same audio recordings as well as their monotonous counterparts following the same procedure. This time the original audio recordings received a mean score of 3.88 (SD = .63), whereas their monotonously sounding equals received a score of 3.64 (SD = .76) on average. Normality tests showed the monotonous condition to be normally distributed, $W(5)$ = .835, p>.05. However the same was not true for the original recordings condition, $W(5)$ = .763, p<.04. Hence the Wilcoxon signed-rank test was performed, that showed the difference in intelligibility between conditions to be nonsignificant, z = -.408, p>.05. The lower overall scoring as compared to the first pretest might have been due to background noise, since the second test was conducted in a notably more noisy room. All results were considered satisfactory, hence allowing for the usage of both voices in the actual experiment.

3.6 Pilot sessions

A pilot study has been conducted with two TD children (one female and one male, respective chronological ages 3:5 and 6:3). Both participants completed one trial on a level that a therapist determined to be age appropriate. During these trials, NAO's voice contained normal intonation. Since this presumably would be the most beneficial condition for TD children, the usage of the voice in question allowed for optimal testing of the contents of the experimental task. Moreover, as the monotonously sounding voice had been pretested for intelligibility, there was no need to include it in the pilot study. The participants were introduced to the experimental task after a short period of playful interaction with the therapist, in order to create a relaxed setting. Since the participants in the actual experiment had previous knowledge of the expectations during a PRT session with NAO, explicit instructions were given to those participating in the pilot study. This way, the experimental set-up was approached as much as possible. The therapist was provided with a manual containing an exact description of the instructions to be given, which can be consulted in appendix 3. Both sessions were videotaped, and investigation of the footage delivered valuable information summarized below.

First, participants sometimes only attended NAO's utterances after encouragement by the therapist or caretaker that attended the sessions. Since intervention by either of these actors should be limited as much as possible, it was decided to equip NAO with the possibility to trigger the child's attention himself. The trigger consisted of the utterance "Listen closely", that could be selected by the therapist preceding any of NAO's utterances she wanted the child to focus on. Second, it appeared that participants often spontaneously initiated (non)verbal interaction when this wasn't required by the game scenario. They for instance imitated NAO's movements, copied his utterances, tried to hand or show him objects and both gave verbal responses and asked questions at other time points than the envisioned ones (see fig. 6 and 7 below). Because spontaneous interaction is one of the goal behaviors that PRT sessions aim to provoke, these initiatives constitute clear signs regarding the effectiveness of the treatment. Therefore, as was already mentioned in section 3.4, a type of measurements that assessed spontaneous communication was added to the experimental design. Third, the robot sometimes offered help while the children did not need him to. Hence an interface option was provided that could skip a help offer to immediately continue with the next step of the game. This allowed for more flexibility in response to the child's actions. Since the interface at this point contained more key-press options than was originally the case, it was adjusted and clarified to allow for response selections to be made as quick as possible. Lastly it is worth noting that the video footage of the pilot sessions confirmed the relevance of the Affect Rating Scale in relation to the current experiment. Example observations that might lead one to assign a certain score could in many cases be matched to actually observed phenomena.

Figure 6: Showing an object (pilot session)

Figure 7: Handing an object (pilot session)

3.7 Procedure

The experiment was conducted in a closed, quiet room in order to minimize possible distractions. Other than the child and the therapist, a caretaker was present in the room. Both the therapist and caretaker were seated at a certain distance from the child, to encourage participants to focus their attention on the robot. It was explained to the caretaker that the child should interact exclusively with NAO. The therapist manual provided specific information regarding the preparation and guidance of each session (see appendix 3).

Before the start of the experimental task, participants answered the first question of the questionnaire (i.e. "How happy are you right now?"). To justify the differences in voice and appearance not only between conditions but also as compared to previous sessions in which the children had encountered NAO, the therapist provided participants with the following explanation. Participants who encountered the robot in its regular appearance were told during the first experimental trial that they would play with a friend of NAO today. This robot was called NEO, and it was pointed out that NAO and NEO look exactly alike and are capable of the same actions. The only difference was, without further specification, explained to be in their voices. During the second trial the child would be reminded that last week, it played with a friend of NAO. This time, another friend (NIO) was said to be visiting. Again, the only difference between NIO and the others was explained to be his voice. For participants who encountered the robot in its humanlike appearance, the story was minimally adjusted. This time it was pointed out during the first session that NEO looks just like NAO, except for his clothing. The rest of the story was kept identical. The names NAO, NEO and NIO were adjusted to the names that participants assigned to the robot during their first Picasso session. The order of conditions was of no influence on the content of the story.

The therapist was instructed to answer any questions that participants might pose using standard answers described in the manual. Unforeseen questions were shortly answered, the therapist's replies being as neutral in form and content as possible. This way, variation between trials was minimized. While normally attention would not specifically be drawn towards aspects of the baseline condition (i.e. the voice used in normal intonation conditions), it was explicitly decided to do so during the current experiment. This decision was made since children with ASD are not good at handling change. Literature discussed in the introduction suggested that the difference in voice between normal intonation conditions and previous robot-mediated therapy sessions would be clearly noticed by these children. Hence not informing them on this change might have caused negative reactions. Besides, mentioning the presence of a change in monotonous conditions but not in normal intonation conditions might have caused participants to prefer the latter simply because less emphasis was put on change. Note that no attention was drawn towards the exact nature of the difference between the voices used across conditions. Hence the experimental manipulation of intonation was not revealed.

During the experimental task, therapist intervention was reduced to a minimum. This was among others accomplished by providing prompts, repetitions and attention triggers exclusively via the robot. After finishing the game, participants answered the two remaining questionnaire items (i.e. "How happy are you right now?" and "How do you like the robot?").

3.8 Data analysis & reduction

Following the recommendation of Fisher, Piazza and Roane (2011), 25% of the data has been scored by a second rater in order to check for interrater reliability. This was accomplished by calculating the intraclass correlation coefficient (ICC). The 25% consisted of four trials of which each represented another experimental condition, that took place at various time points during the execution of the experiment. Within conditions and time clusters, the trials were randomly selected. The ICCs for required prompting, spontaneous initiatives, and affect scores respectively came down to 1, 0.75 and 0.79. For the prompt and initiative measurements, the single measures ICC is reported as exact agreement over individual scores was relevant. For affect scores on the other hand the average measures ICC is reported, since the mean of all scores obtained was the unit of analysis here. The ICCs were interpreted following the guidelines by Cicchetti (1994), that take scores less than .40 to reflect poor agreement, scores between 0.40 and 0.59 to indicate fair agreement, scores between .60 and .74 to prove good agreement, and scores of .75 and up to reflect excellent agreement. Hence excellent agreement has been reached for all types of measurements.

After data collection was completed, learning opportunities that proved to be superfluous during a specific trial were excluded from analysis (such as when the child should have asked for help, but was capable of performing the action on its own). This way, such superfluency was prevented from influencing the results. It was furthermore decided to only take verbal and nonverbal initiatives directed towards the robot into account. Their numbers were expressed as percentages of respectively the total number of verbal and nonverbal initiatives during a trial. This decision was made since it proved to be hard to control for influences on initiatives that were not directed towards the robot, such that differences in corresponding numbers could not with certainty be addressed to the experimental manipulations. Finally, the introspective questions of the questionnaire proved to be too difficult for many participants to answer and/or motivate. As obtained answers were often unreliable, the questionnaire was excluded from analysis. Nonetheless, attention will be paid to its outcomes in the discussion section of this report.

4 Results

Data were analyzed by means of the linear mixed effects model (Gelman & Hill, 2007), using SPSS Statistics V22.0 (IBM). Intonation and appearance were fixed factors, whereas participant identity was a random factor that corrected for irrelevant interpersonal differences. All effects were taken to be significant at $p<.05$. Note that graphs do not contain error bars, as SPSS does not produce reliable ones in case of non-normally distributed data (Field, 2009). Moreover, their depiction can be misleading when variance is large.

Three significant effects were found. First, there was a main effect of appearance on general affect scores, $F(1,6) = 6.588$, $p = .043$. Obtained scores were significantly higher in humanlike as compared to mechanical appearance conditions (fig. 8).

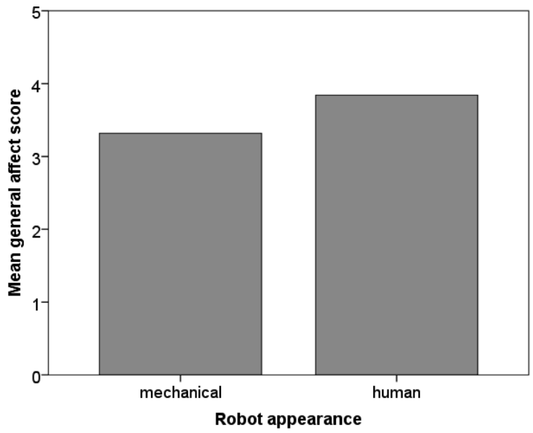


Figure 8: Significant main effect of appearance on general affect scores

Second, another main effect of appearance was found on interest scores, $F(1,6) = 8.446$, $p = .027$. That is, in humanlike appearance conditions significantly higher interest scores were obtained as compared to mechanical appearance conditions (fig. 9).

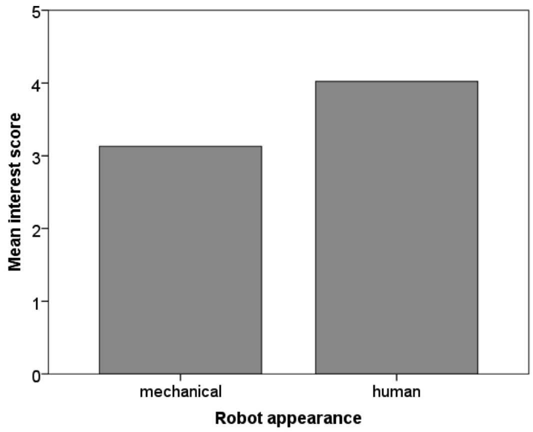


Figure 9: Significant main effect of appearance on interest scores

Lastly, there was a significant interaction effect of intonation and appearance on happiness scores, $F(1,6) = 9.936$, $p = .020$. Participants obtained higher happiness scores (1) when a mechanical appearance was combined with monotonous intonation, and (2) when a humanlike appearance was combined with normal intonation (fig. 10).

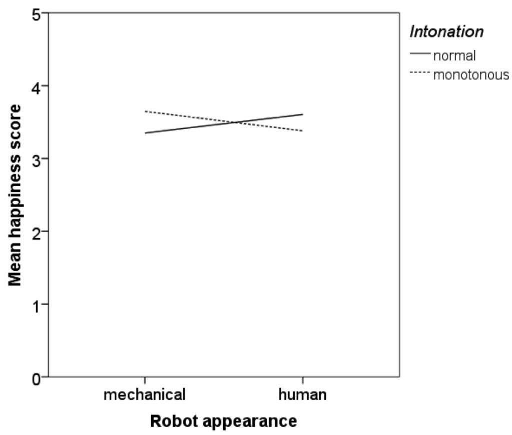


Figure 10: Significant interaction effect on happiness scores

No significant effects were found of either intonation, appearance or their interaction on the average percentage of required prompts. The same holds for the measurements of verbal and nonverbal spontaneous initiatives. Intonation neither had a significant effect on affect scores, whether general or per individual dimension, and appearance effects failed to reach significance for the dimensions of happiness and behavior. Finally, interaction effects were absent for general affect scores as well as interest and behavior scores. Appendix 4 contains SPSS output tables for all measurements, in which their corresponding p-values can be consulted.

5 Discussion

The hypothesis as formulated in the introduction stated that intonation and bodily appearance of robots influence treatment outcomes of robot-mediated therapy sessions for children with HFA. The expected influences were subsequently specified in three accompanying predictions. First, using a robot with monotonous intonation was predicted to result in better treatment outcomes than using a robot with normal intonation. Second, humanizing a robot's bodily appearance was predicted to lead to superior treatment outcomes as compared to presenting the same robot in its regular, mechanical appearance. Third, it was predicted that congruence of intonation and appearance (both human- or machinelike) would lead to optimal treatment outcomes. In all cases, treatment outcomes were taken to be reflected in improved interaction task performance and enhanced affective states.

The first prediction, concerning the influence of intonation, was not borne out. There were no effects of intonation on either of the performance measurements. The absence of effects on the average percentage of required prompts might be due to a floor effect, as the mean across conditions only came down to 8.55%. This floor effect could in turn be a consequence of the previous experience that participants had with PRT therapy sessions. During their previous practice, they have learned what kind of reactions are expected and have become more capable of giving these reactions by themselves. Therefore prompting might only have been of marginal need to these children, which could have prevented experimental manipulations from further lowering the number of prompts they received. On the other hand, the absence of effects on verbal as well as nonverbal spontaneous initiatives is likely to be due to large variances in the data (that sometimes even exceeded the mean), and hence to the small sample size that was mentioned earlier. Effects of intonation on affective states were not found either. This time the experimental design might form an explanation for the absence of significant findings, as will be discussed further on. Nevertheless, there existed an overall tendency among affect measurements towards more positive outcomes in monotonous conditions. Whereas differences were small and nonsignificant, both general affect scores and separate interest, happiness and behavior scores were higher when the robot's voice contained monotonous intonation (table 3).

Table 3: Mean affect scores in normal and monotonous intonation conditions

	Normal intonation (M)	Monotonous intonation (M)
General affect	3.53	3.63
Interest	3.49	3.67
Happiness	3.48	3.51
Behavior	3.63	3.71

The second prediction, addressing influences of bodily appearance, was partially shown to be correct. No matter the possible above-mentioned explanations, appearance had no significant effect on the performance measurements. However, this factor did have a clear effect on participants' affective states. Namely, participants obtained higher general affect scores as well as higher interest scores when presented with the robot in its humanlike appearance. Apparently the positive influence of the humanized appearance outperformed the 'need for sameness' of children with ASD, that would normally lead to a negative influence of changed features. Reluctance to change might nevertheless explain differences between mean scores for all significant findings to be rather small. That is, negative effects of habituation to the robot's mechanical appearance might have leveled out the positive effects of its humanization. The same could count for differences between mean affect scores reported in table 3. As the voice used in normal intonation conditions resembled the voice children were familiar with to a higher extent than did the voice used in monotonous conditions, habituation effects might have leveled out positive effects of the latter.

Finally, the reported experiment has not been able to show congruence of intonation and appearance to result in optimal interaction task performance. As for affective measurements however, this third prediction receives support from an interaction effect on happiness scores. Whereas participants that were presented with the robot in its mechanical appearance obtained higher happiness scores in case its voice contained monotonous intonation, participants that encountered the robot in its humanlike appearance were more happy when its voice contained normal intonation. For the dimension of happiness, congruence hence indeed leads to optimal treatment outcomes. However, only one interaction effect was found across affective measurements. This finding therefore merely indicates the possibility that congruence of intonation and appearance might equally improve affective states as a whole.

While the experimental results do not strongly support the predictions, discussed findings and observations do suggest that both intonation and appearance influence participants' affective states. On the contrary, the current study was not able to confirm their influence on interaction task performance. The fact that appearance did have more outspoken effects than intonation could possibly be related to the level on which both factors operated. Appearance was varied on a between-subject basis, and therefore only changed as compared to regular treatment sessions for participants who encountered the robot in its humanlike appearance. Intonation on the other hand constituted a within-subject factor that presented each participant with two unknown robot voices, hence also establishing change between trials. This might have resulted in a larger influence of participants' need for sameness on intonation-related outcomes, which could explain the absence of significant findings for this factor.

Reluctance to change surfaces most clearly when it comes to conscious preferences. While the questionnaire was excluded from analysis as most participants were unable to provide motivated and coherent answers, responses that did meet these criteria generally reflected a negative attitude towards the experimental manipulations. More specifically, in contrast to the outcomes of the Affect Rating Scale, half of the children reported to prefer playing with the robot as they knew it from earlier therapy sessions following at least one of the trials they engaged in. Three of the four participants that indicated such a preference belonged to experimental group 2, that experienced both changes in intonation and appearance. This supports the idea that influences of need for sameness increased with the experienced amount of change, as was earlier suggested to explain the absence of significant intonation effects. Nonetheless, only one time a negative attitude towards the changed features was accompanied by a negative robot appreciation score. Moreover, there were equally answers that reflected a positive attitude towards the humanlike appearance and/or monotonous voice, sometimes coming from the same participants who reported to prefer the 'familiar robot'. The questionnaire data hence also seem to confirm that while participants did like the changed features, their conscious reluctance to change negatively influenced affective measurements, as was suggested during the discussion of the second prediction.

Regarding suggestions for future research, it is worth noting that many methodological limitations of the current study were due to its embeddedness in the Picasso project. Therefore, researchers engaged in this project have indicated to be interested in performing an independent follow-up study (p.c. M. van Dongen, W. Staal & I. Oosterling). This could lead to firmer conclusions, that surpass the exploratory outcomes reported here. First of all, a bigger participant sample and increased number of trials per participant would decrease variance and increase statistical power. Second, it would be desirable to engage participants without previous experience with the robot such that habituation effects would be ruled out. Effects that were found during the current study could then possibly be magnified. Additionally, testing participants who have no (or in each case less) experience with PRT therapy sessions would cancel the floor effect encountered for the average percentage of required prompts. Varying both factors on a between-subject basis might furthermore allow for effects of intonation to surface more clearly. A suggestion for future research that does not directly relate to the limitations of the present would be to conduct a longitudinal study. Robot-mediated PRT could be provided to various groups of children, that each receive treatment using robots with different voices and appearances. This way it could be assessed whether certain feature manipulations and their various combinations could enhance treatment outcomes on the long term, as well as fasten generalization of obtained communicative skills in the child's everyday environment.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: Author.

Barakova, E. I., Bajracharya, P., Willemsen, M., Lourens, T., & Huskens, B. (2014). Long-term LEGO therapy with humanoid robot for children with ASD. *Expert Systems, 32*(6), 698–709.

Bildt, A. de, Greaves-Lord, K., & Jonge, M. de. (2013). *ADOS-2: Autisme Diagnostisch Observatie Schema. Handleiding.* Amsterdam: Hogrefe.

Billard, A. (2003). Robota: Clever toy and educational tool. *Robotics and Autonomous Systems, 42*, 259–269.

Billard, A., Robins, B., Nadel, J., & Dautenhahn, K. (2007). Building Robota, a mini-humanoid robot for the rehabilitation of children with autism. *Assistive Technology, 19*, 37–49.

Boersma, P., & Weenink, D. (2015). *Praat: Doing phonetics by computer* [Computer program]. Version 6.0.05, retrieved 12 November 2015 from http://www.praat.org/.

Cabibihan, J. J., Javed, H., & Ang, M. (2013). Why robots? A survey on the roles and benefits of social robots in the therapy of children with autism. *International Journal of Social Robotics, 5*(4), 593–618.

Charlop-Christy, M. H., Le, L., & Freeman, K. A. (2000). A comparison of video modeling with in-vivo modeling for teaching children with autism. *Journal of Autism and Developmental Disorders, 30*(6), 537–552.

Chevallier, C., Kohls, G., Troiani, V., Brodkin, E. S., & Schultz, R. T. (2012). The social motivation theory of autism. *Trends in Cognitive Science, 16*(4), 231–239.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290.

Dautenhahn, K., & Werry, I. (2004). Towards interactive robots in autism therapy: Background, motivation and challenges. *Pragmatics & Cognition, 12*(1), 1–35.

Destephe, M., Zecca, M., Hashimoto, K., & Takanishi, A. (2014, December). Uncanny Valley, robot and autism: Perception of the uncanniness in an emotional gait. In *International Conference on Robotics & Biomimetics* (pp. 1152–1157). IEEE.

Diehl, J. J., Schmitt, L. M., Villano, M., & Crowell, C. R. (2012). The clinical use of robots for individuals with autism spectrum disorders: A critical review. *Research in Autism Spectrum Disorders, 6*, 249–262.

Eigsti, I-M., Schuh, J., Mencl, E., Schultz, R. T., & Paul, R. (2012). The neural underpinnings of prosody in autism. *Child Neuropsychology, 18*(6), 600–617.

Feil-Seifer, D., & Matarić, M. J. (2009). Towards socially assistive robotics for augmenting interventions for children with autism spectrum disorders. In O. Khatib, V. Kumar & J. Pappas (Eds.), *Experimental robotics* (pp. 201–210). Berlin: Springer.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: SAGE.

Fisher, W. W., Piazza, C. C., & Roane, H. S. (Eds.). (2011). *Handbook of applied behavior analysis.* New York: The Guilford Press.

Flannery, K. B., & Horner, R. H. (1994). The relationship between predictability and problem behavior for students with severe disabilities. *Journal of Behavioral Education, 4*, 157–176.

Frith, U. (1970). Studies in pattern detection in normal and autistic children: II. Reproduction and production of color sequences. *Journal of Experimental Child Psychology, 10*(1), 120–135.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambrigde University Press.

Gepner, B., Deruelle, C., & Grynfeltt, S. (2001). Motion and emotion: A novel approach to the study of face processing by young autistic children. *Journal of Autism and Developmental Disorders, 31*(1), 37–45.

Güçlü, B., Tanidir, C., Mukaddes, N. M., & Ünal, F. (2007). Tactile sensitivity of normal and autistic children. *Somatosensory and Motor Research, 24*(1/2), 21–33.

Hendriksen, J., & Hurks, P. (2009). *Wechsler Preschool and Primary Scale of Intelligence - third edition - Nederlandse bewerking (WPPSI-III-NL).* Amsterdam: Pearson.

Hesling, I., Dilharreguy, B., Peppe, S., Amirault, M., Bouvard, M., & Allard, M. (2010). The integration of prosodic speech in high functioning autism: A preliminary fMRI study. *PLOS ONE, 5*(7). doi:10.1371/journal.pone.0011571

Huskens, B., Palmen, A., Werff, M. van der, Lourens, T., & Barakova, E. I. (2015). Improving collaborative play between children with autism spectrum disorders and their siblings: The effectiveness of a robot-mediated intervention based on Lego therapy. *Journal of Autism and Developmental Disorders, 45*, 3746–3755.

Huskens, B., Verschuur, R., Gillessen, J., Didden, R., & Barakova, E. I. (2013). Promoting question-asking in school-aged children with autism spectrum disorders: Effectiveness of a robot intervention compared to a human-trainer intervention. *Journal of Developmental Neurorehabilitation, 16*(5), 345–356.

Jarvinen-Pasley, A., & Heaton, P. (2007). Evidence for reduced domain-specificity in auditory processing in autism. *Developmental Science, 10*(6), 786–793.

Jarvinen-Pasley, A., Wallace, G. L., Ramus, F., Happe, F., & Heaton, P. (2008). Enhanced perceptual processing of speech in autism. *Developmental Science, 11*(1), 109–121.

Johnson, C. P., & Myers, S. M. (2007). Identification and evaluation of children with autism spectrum disorders. *Pediatrics, 120*(5), 183–215.

Just, M. A., Cherkassky, V. L., Keller, T. A., & Minshew, N. J. (2004). Cortical activation and synchronization during sentence comprehension in high-functioning autism: Evidence of underconnectivity. *Brain, 127*(8), 1811–1821.

Kanner, L. (1943). Autistic disturbances of affective contact. *The Nervous Child, 2*, 217–250.

Kasari, C. (2002). Assessing change in early intervention programs for children with autism. *Journal of Autism and Developmental Disorders, 32*(5), 447–461.

Koegel, R. L., & Egel, A. L. (1979). Motivating autistic children. *Journal of Abnormal Psychology, 88*(4), 418–426.

Koegel, R. L., & Koegel, L. K. (2006). *Pivotal response treatment for autism: Communication, social, and academic development.* Baltimore, MD: Brooks.

Koegel, R. L., Koegel, L. K., & McNerney, E. K. (2001). Pivotal areas in intervention for autism. *Journal of Clinical Child Psychology, 30*(1), 19–32.

Koegel, L. K., & LaZebnik, C. (2014). *Overcoming autism: Finding the answers, strategies, and hope that can transform a child's life.* New York: Penguin.

Koegel, L. K., Singh, A. K., & Koegel, R. L. (2010). Improving motivation for academics in children with autism. *Journal of Autism and Developmental Disorders, 40*, 1057–1066.

Kort, W., Schittekatte, M., Dekker, P. H., Verhaeghe, P., Compaan, E. L., Bosmans, M., & Vermeir, G. (2005). *Wechsler Intelligence Scale for Children - third edition - Nederlandse bewerking (WISC-III-NL). Handleiding en verantwoording.* Amsterdam: Harcourt Test.

Kozima, H., & Nakagawa, C. (2006). Interactive robots as facilitators of children's social development. In A. Lazinica (Ed.), *Mobile robots: Towards new applications* (pp. 169–186). Vienna: I-Tech.

Lourens, T., & Barakova, E. I. (2011). User-friendly robot environment for creation of social scenarios. In J. M. Ferrández, J. R. Álvarez, F. de la Paz & F. J. Toledo (Eds.), *Foundations on natural and artificial computation* (pp. 212–221). Berlin: Springer.

Lovaas, O. I., Koegel, R. L., & Schreibman, L. (1979). Stimulus overselectivity in autism: A review of research. *Psychological Bulletin, 86*(6), 1236–1254.

Macintosh, K., & Dissanayake, C. (2006). Social skills and problem behaviours in school aged children with high functioning autism and Asperger's Disorder. *Journal of Autism and Developmental Disorders, 36*(8), 1065–1076.

Miller, L. K. (1999). The savant syndrome: Intellectual impairment and exceptional skill. *Psychological Bulletin, 125*, 31–46.

Mori, M. (2012). The uncanny valley. (K. F. MacDorman & N. Kageki, Trans.). *Robotics & Automation Magazine, 19*(2), 98–100.

Mullen, E. M. (1995). *Mullen scales of early learning: AGS edition.* Circle Pines, MN: American Guidance Service.

Mundy, P., Gwaltney, M., & Henderson, H. (2010). Self-referenced processing, neurodevelopment and joint attention in autism. *Autism, 14*(5), 408–429.

Nalin, M., Bergamini, L., Giusti, A., Baroni, I., & Sanna, A. (2011, March). Children's perception of a robotic companion in a mildly constrained setting. In *Proceedings of the Human-Robot Interaction Conference* (pp. 26–29). ACM.

O'Riordan, M., & Passetti, F. (2006). Discrimination in autism within different sensory modalities. *Journal of Autism and Developmental Disorders, 36*, 665–675.

Peca, A., Simut, R., Pintea, S., Costescu, C., & Vanderborght, B. (2014). How do typically developing children and children with autism perceive different social robots? *Computers in Human Behavior, 41*, 268–277.

Reynolds, B. S., Newsom, C. D., & Lovaas, O. I. (1974). Auditory overselectivity in autistic children. *Journal of Abnormal Child Psychology, 2*(4), 253–263.

Robins, B., Dautenhahn, K., & Dubowski, J. (2006). Does appearance matter in the interaction of children with autism with a humanoid robot? *Interaction Studies, 7*(3), 479–512.

Russo, N. (2007). *Barking frogs and chirping frogs: A behavioral and brain EEG study of multisensory matching among persons with autism spectrum disorder* (Doctoral dissertation). Retrieved from http://digitool.library.mcgill.ca/webclient/StreamGate?folder_id=0&dvs=1463492302636 127.

Scassellati, B. (2007). How social robots will help us diagnose, treat, and understand autism. *Robotics Research, 28*, 552–563.

Scassellati, B., Admoni, H., & Matarić, M. (2012). Robots for use in autism research. *Annual Review of Biomedical Engineering, 14*, 275–294.

Shamsuddin, S., Yussof, H., Ismail, L. I., Mohamed, S., Hanapiah, F. A., & Zahari, N. I. (2012). Initial response in HRI: A case study on evaluation of child with autism spectrum disorders interacting with a humanoid robot NAO. *Procedia Engineering, 41*, 1448–1455.

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical biases and referential scenes in child and adult sentence processing. *Cognitive Psychology, 49*, 238–299.

Tapus, A., Peca, A., Aly, A., Pop, C., Jisa, L., Pintea, S., ... David, D. (2012). Children with autism social engagement in interaction with Nao, an imitative robot - A series of single case experiments. *Interaction Studies, 13*(3), 315–347.

Verschuur, R., Didden, R., Lang, R., Sigafoos, J., & Huskens, B. (2014). Pivotal response treatment for children with autism spectrum disorders: A systematic review. *Review Journal of Autism and Developmental Disorders, 1*(1), 34–61.

Williams, J. A., Koegel, R. L., & Egel, A. L. (1981). Response-reinforcer relationships and improved learning in autistic children. *Journal of Applied Behavior Analysis, 14*, 53–60.

Woods, S., Dautenhahn, K., & Schultz, J. (2004, September). The design space of robots: Investigating children's views. In *International Workshop on Robot and Human Interactive Communication* (pp. 47–52). IEEE.

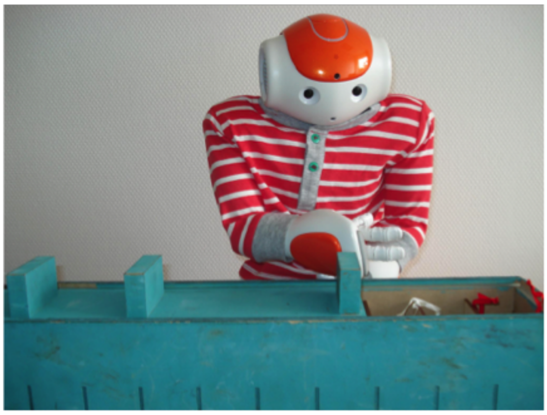
World Health Organization. (1993). *The ICD-10 classification of mental and behavioural disorders.* Geneva: Author.

Appendix 1: Game illustrations


At the start of the game the child sat in front of NAO, a big brown box located at his side. After appropriate child-robot interaction had taken place (e.g. greeting, noticing that the box is tied up, asking for scissors), the child gained access to the box. In the box were three smaller boxes, each containing one puzzle piece of each animal puzzle depicted on its cover.

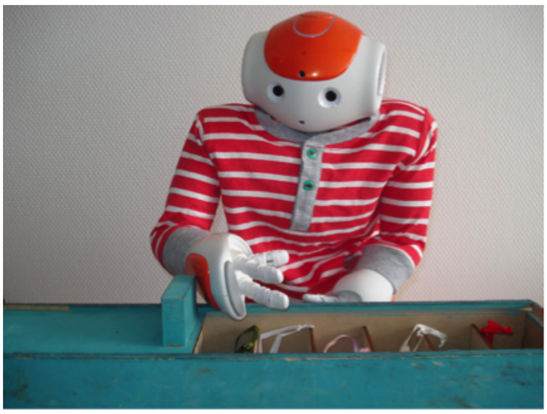

The child took the puzzle boxes and the brown box was removed by the trainer. NAO then told the child to pick one of the puzzle boxes. The trainer removed the two remaining ones.


Then the trainer put a blue sliding box into position. In the blue box were nine colored bags, that contained the missing pieces of all animal puzzles. For some game levels, the bags did also contain pins that should be used to attach the various pieces. For others, additional questions had to be posed by the child in order to obtain the pins. In this case, they were provided after appropriate child-robot interaction had taken place.



When starting the first puzzle, the child noticed that the provided materials were insufficient. Following appropriate interaction (i.e. asking for the remaining pieces), the robot slided the cover of the blue box such that three colored bags would be in sight, and told the child which bag contained the right puzzle pieces.

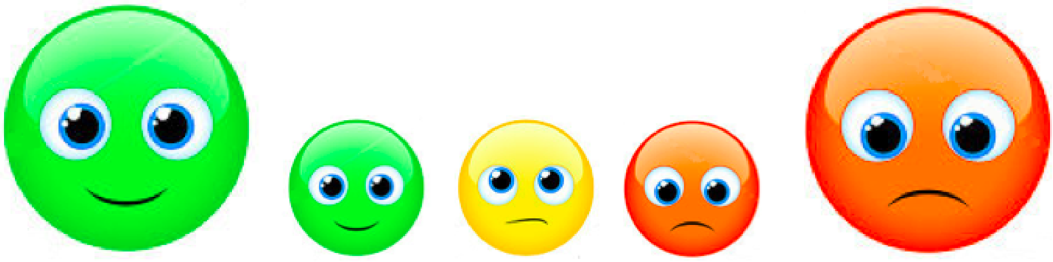
The child then took the correct bag from the blue box, hence gaining access to the remaining materials. With these materials, it finished the first puzzle. During puzzle completion, functional as well as social learning opportunities were provided (e.g. asking for help, showing interest). Afterwards, the robot would make the sound of the animal in question when the child asked for it correctly.



This procedure repeated itself two more times, such that the remaining two animal puzzles could be completed by the child as well. After all three puzzles had been put together, additional learning opportunities among others involved social interaction regarding the puzzles and tidying up. Afterwards, the robot and child greeted each other and the game came to an end.

Appendix 2: Questionnaire items

Hoe blij ben je nu?



☐	☐	☐	☐	☐
heel blij	blij	een beetje blij	niet blij	helemaal niet blij

Waarom: ....

Hoe leuk vind je de robot?

☐	☐	☐	☐	☐
heel leuk	leuk	een beetje leuk	niet leuk	helemaal niet leuk

Waarom: ....

Appendix 3: Therapist manual

Inhoudsopgave

1. Informatieoverzicht experimenten

Groep 1 (4 participanten) *
Conditie 1: intonatie + normaal uiterlijk
Conditie 2: monotoon + normaal uiterlijk

Groep 2 (4 participanten) *
Conditie 1: intonatie + kleding
Conditie 2: monotoon + kleding

* Bij elke participant worden twee sessies afgenomen: een in conditie 1, en een in conditie 2

Tijdens het experiment gebruiken we, afhankelijk van het PRT-behandelniveau van de participant, de hieronder aangegeven spelscenario's. De scenario's waarvan de naam eindigt op 'into' worden gebruikt in de intonatie condities, scenario's die eindigen op 'mono' in de monotone condities.

- PG_1_02122015into / PG_1_02122015mono
- PG_2_02122015into / PG_2_02122015mono
- PG_3_02122015into / PG_3_02122015mono
- PG_4_02122015into / PG_4_02122015mono
- PG_5_02122015into / PG_5_02122015mono
- PG_6_02122015into / PG_6_02122015mono
- PG_7_02122015into / PG_7_02122015mono

Alle experimentsessies worden gefilmd.

Voorafgaand en na afloop van het spel worden 3 vragen met het kind ingevuld:

- Hoe blij ben je nu? (voorafgaand)
- Hoe blij ben je nu? (na afloop)
- Hoe leuk vind je de robot? (na afloop)

2. Voorbereiding van experimentsessies

Let op!
In aanvulling op wat in deze sectie volgt moet NAO in voorbereiding op alle sessies van groep 2 zijn kleding aankrijgen. Het shirt gaat het gemakkelijkst aan door de knoopjes los te maken, dan eerst NAO's armen door de mouwen te steken en vervolgens het shirt over zijn hoofd te trekken. Maak de knoopjes daarna weer dicht. Om de broek aan te trekken leg je NAO op zijn rug en maak je zijn benen en voeten zo recht mogelijk. Trek het koord aan om zijn middel en leg er een strik/knoopje in. Zet NAO klaar in zijn positie, en zorg ervoor dat de broekspijpen zijn benen/voeten geheel bedekken. Dit doe je door te zorgen dat de stof aan de achterkant niet klem zit tussen zijn opgevouwen benen, en vervolgens de broekspijpen helemaal over zijn benen/voeten te trekken (de broek rekt mee). Zie onderstaande foto.

Op de volgende pagina's wordt per behandelniveau uitgelegd wat er hoe moet worden klaargezet.

Level 1

Bruine doos
Bevat drie genummerde plastic bakjes, met in elk bakje drie stukken van dierenpuzzels. Op elk bakje dienen de plaatjes te worden geplakt van de dieren die erin zitten. Keuze uit de volgende stukken (belangrijk i.v.m. kiesselectie uitspraken NAO):

- Poten van de hond
- Lijf van de uil
- Lijf van de pinguïn
- Vleugel van de vlinder
- Hoofd van het varken
- Hoofd van de bij
- Lijf van de kip
- Hoofd van de olifant
- Lijf van de koe
- Lijf van het schaap
- Lijf van het hert
- Vleugel van de vleermuis
- Lijf van het nijlpaard

Blauwe doos
Positie 1: rood zakje, bevat resterende stukjes van dier 1 uit bakje 1 + pinnetjes
Positie 4: geel zakje, bevat resterende stukjes van dier 2 uit bakje 1 + pinnetjes
Positie 7: paars zakje, bevat resterende stukjes van dier 3 uit bakje 1 + pinnetjes

Positie 2: groen zakje, bevat resterende stukjes van dier 1 uit bakje 2 + pinnetjes
Positie 5: roze zakje, bevat resterende stukjes van dier 2 uit bakje 2 + pinnetjes
Positie 8: bruin zakje, bevat resterende stukjes van dier 3 uit bakje 2 + pinnetjes

Positie 3: lichtblauw zakje, bevat resterende stukjes van dier 1 uit bakje 3 + pinnetjes
Positie 6: wit zakje, bevat resterende stukjes van dier 2 uit bakje 3 + pinnetjes
Positie 9: donkerblauw zakje, bevat resterende stukjes van dier 3 uit bakje 3 + pinnetjes

Trainer
De trainer houdt niets achter.

Level 2

Bruine doos
Bevat drie genummerde plastic bakjes, met in elk bakje drie stukken van dierenpuzzels. Op elk bakje dienen de plaatjes te worden geplakt van de dieren die erin zitten.

Blauwe doos
Positie 1: rood zakje, bevat resterende stukjes van dier 1 uit bakje 1
Positie 4: geel zakje, bevat resterende stukjes van dier 2 uit bakje 1
Positie 7: paars zakje, bevat resterende stukjes van dier 3 uit bakje 1

Positie 2: groen zakje, bevat resterende stukjes van dier 1 uit bakje 2

Positie 5: roze zakje, bevat resterende stukjes van dier 2 uit bakje 2
Positie 8: bruin zakje, bevat resterende stukjes van dier 3 uit bakje 2

Positie 3: lichtblauw zakje, bevat resterende stukjes van dier 1 uit bakje 3
Positie 6: wit zakje, bevat resterende stukjes van dier 2 uit bakje 3
Positie 9: donkerblauw zakje, bevat resterende stukjes van dier 3 uit bakje 3

Trainer
De trainer houdt alle pinnetjes achter. De hoeveelheid pinnetjes hangt af van:

- De keuze van het kind uit bakje 1, 2 of 3
- De dieren die het desbetreffende bakje bevat

De trainer geeft, wanneer het kind om staafjes vraagt, steeds niet in een keer alle staafjes die het kind nodig heeft om het dier waarmee het bezig is af te maken.

Level 3

Bruine doos = dichtgemaakt met lint
Bevat drie genummerde plastic bakjes, met in elk bakje drie stukken van dierenpuzzels. Op elk bakje dienen de plaatjes te worden geplakt van de dieren die erin zitten.

Blauwe doos
Positie 1: rood zakje, bevat resterende stukjes van dier 1 uit bakje 1
Positie 4: geel zakje, bevat resterende stukjes van dier 2 uit bakje 1
Positie 7: paars zakje, bevat resterende stukjes van dier 3 uit bakje 1

Positie 2: groen zakje, bevat resterende stukjes van dier 1 uit bakje 2
Positie 5: roze zakje, bevat resterende stukjes van dier 2 uit bakje 2
Positie 8: bruin zakje, bevat resterende stukjes van dier 3 uit bakje 2

Positie 3: lichtblauw zakje, bevat resterende stukjes van dier 1 uit bakje 3
Positie 6: wit zakje, bevat resterende stukjes van dier 2 uit bakje 3
Positie 9: donkerblauw zakje, bevat resterende stukjes van dier 3 uit bakje 3

Trainer
De trainer houdt de volgende spullen achter:

- Alle pinnetjes *
- Een schaar

* De hoeveelheid pinnetjes hangt af van:

- De keuze van het kind uit bakje 1, 2 of 3
- De dieren die het desbetreffende bakje bevat

De trainer geeft, wanneer het kind om staafjes vraagt, steeds meteen alle staafjes die nodig zijn om het dier af te maken waarmee het kind bezig is.

Level 4

Bruine doos = dichtgemaakt met lint
Bevat drie genummerde plastic bakjes, met in elk bakje drie stukken van dierenpuzzels. Op elk bakje dienen de plaatjes te worden geplakt van de dieren die erin zitten.

Blauwe doos
Positie 1: rood zakje, bevat resterende stukjes van dier 1 uit bakje 1 + pinnetjes
Positie 4: geel zakje, bevat resterende stukjes van dier 2 uit bakje 1 + pinnetjes
Positie 7: paars zakje, bevat resterende stukjes van dier 3 uit bakje 1

Positie 2: groen zakje, bevat resterende stukjes van dier 1 uit bakje 2 + pinnetjes
Positie 5: roze zakje, bevat resterende stukjes van dier 2 uit bakje 2 + pinnetjes
Positie 8: bruin zakje, bevat resterende stukjes van dier 3 uit bakje 2

Positie 3: lichtblauw zakje, bevat resterende stukjes van dier 1 uit bakje 3 + pinnetjes
Positie 6: wit zakje, bevat resterende stukjes van dier 2 uit bakje 3 + pinnetjes
Positie 9: donkerblauw zakje, bevat resterende stukjes van dier 3 uit bakje 3

Trainer
De trainer houdt de volgende spullen achter:
- De pinnetjes van het dier in het paarse, bruine of donkerblauwe zakje *
- Een schaar

* De hoeveelheid pinnetjes hangt af van:
- De keuze van het kind uit bakje 1, 2 of 3
- Het dier in het paarse, bruine of donkerblauwe zakje

De trainer geeft, wanneer het kind om staafjes voor het desbetreffende dier vraagt, meteen alle staafjes die nodig zijn om het dier af te maken.

Level 5

Bruine doos = dichtgemaakt met lint
Bevat drie genummerde plastic bakjes, met in elk bakje drie stukken van dierenpuzzels. Op elk bakje dienen de plaatjes te worden geplakt van de dieren die erin zitten.

Blauwe doos
Positie 1: rood zakje, bevat resterende stukjes van dier 1 uit bakje 1
Positie 4: geel zakje, bevat resterende stukjes van dier 2 uit bakje 1

Positie 7: paars zakje, bevat resterende stukjes van dier 3 uit bakje 1

Positie 2: groen zakje, bevat resterende stukjes van dier 1 uit bakje 2
Positie 5: roze zakje, bevat resterende stukjes van dier 2 uit bakje 2
Positie 8: bruin zakje, bevat resterende stukjes van dier 3 uit bakje 2

Positie 3: lichtblauw zakje, bevat resterende stukjes van dier 1 uit bakje 3
Positie 6: wit zakje, bevat resterende stukjes van dier 2 uit bakje 3
Positie 9: donkerblauw zakje, bevat resterende stukjes van dier 3 uit bakje 3

Trainer
De trainer houdt de volgende spullen achter:
- Alle pinnetjes *
- Een schaar

* De hoeveelheid pinnetjes hangt af van:
- De keuze van het kind uit bakje 1, 2 of 3
- De dieren die het desbetreffende bakje bevat

De trainer geeft, wanneer het kind om staafjes vraagt, steeds meteen alle staafjes die nodig zijn om het dier af te maken waarmee het kind bezig is.

Level 6

Bruine doos = dichtgemaakt met lint
Bevat drie genummerde plastic bakjes, met in elk bakje drie stukken van dierenpuzzels. Op elk bakje dienen de plaatjes te worden geplakt van de dieren die erin zitten.

Blauwe doos
Positie 1: rood zakje, bevat resterende stukjes van dier 1 uit bakje 1 + pinnetjes
Positie 4: geel zakje, bevat resterende stukjes van dier 2 uit bakje 1 + pinnetjes
Positie 7: paars zakje, bevat resterende stukjes van dier 3 uit bakje 1 + pinnetjes

Positie 2: groen zakje, bevat resterende stukjes van dier 1 uit bakje 2 + pinnetjes
Positie 5: roze zakje, bevat resterende stukjes van dier 2 uit bakje 2 + pinnetjes
Positie 8: bruin zakje, bevat resterende stukjes van dier 3 uit bakje 2 + pinnetjes

Positie 3: lichtblauw zakje, bevat resterende stukjes van dier 1 uit bakje 3 + pinnetjes
Positie 6: wit zakje, bevat resterende stukjes van dier 2 uit bakje 3 + pinnetjes
Positie 9: donkerblauw zakje, bevat resterende stukjes van dier 3 uit bakje 3 + pinnetjes

Trainer
De trainer houdt een schaar achter.

Level 7

Bruine doos = dichtgemaakt met lint
Bevat drie genummerde plastic bakjes, met in elk bakje drie stukken van dierenpuzzels. Op elk bakje dienen de plaatjes te worden geplakt van de dieren die erin zitten.

Blauwe doos
Positie 1: rood zakje, bevat resterende stukjes van dier 1 uit bakje 1 + pinnetjes
Positie 4: geel zakje, bevat resterende stukjes van dier 2 uit bakje 1 + pinnetjes
Positie 7: paars zakje, bevat resterende stukjes van dier 3 uit bakje 1 + pinnetjes

Positie 2: groen zakje, bevat resterende stukjes van dier 1 uit bakje 2 + pinnetjes
Positie 5: roze zakje, bevat resterende stukjes van dier 2 uit bakje 2 + pinnetjes
Positie 8: bruin zakje, bevat resterende stukjes van dier 3 uit bakje 2 + pinnetjes

Positie 3: lichtblauw zakje, bevat resterende stukjes van dier 1 uit bakje 3 + pinnetjes
Positie 6: wit zakje, bevat resterende stukjes van dier 2 uit bakje 3 + pinnetjes
Positie 9: donkerblauw zakje, bevat resterende stukjes van dier 3 uit bakje 3 + pinnetjes

Trainer
De trainer houdt een schaar achter.

3. Tijdens de experimentsessies: Instructies voor de trainer

Algemene instructies

- Vertel voorafgaand aan de sessie het volgende aan de ouder/verzorger die aanwezig is in de ruimte: "De sessie van vandaag maakt deel uit van een experiment waarin we onderzoeken hoe we de robot kunnen verbeteren. Het is dus belangrijk dat de kinderen zoveel mogelijk alleen met NAO communiceren. Daarom wil ik u vragen om tijdens de sessie niet in interactie te treden met [naam kind], en hem/haar niet te helpen tijdens het spelletje - ook niet wanneer hij/zij een passieve houding aanneemt. Wanneer [naam kind] zich direct tot u richt, is het belangrijk dat uw inbreng minimaal blijft. Als hij/zij u bijvoorbeeld een vraag stelt over het spel of aangeeft niet te weten hoe het verder moet, kunt u antwoorden dat NAO hier vast en zeker antwoord op weet of bij kan helpen. Wanneer uw inbreng echter direct nodig is voor de voortgang van het spel, zoals wanneer [naam kind] aan u aangeeft geen zin te hebben om verder te spelen, kunt u hem/haar op uw manier aansporen om toch verder te gaan."
- Vraag de ouder/verzorger om (wanneer deze normaal vlakbij het kind zit of het kind op schoot neemt tijdens de sessie) wat verder weg van het kind te gaan zitten dan normaal, zodat het kind gestimuleerd wordt om zich op NAO te concentreren.
- Ga zelf niet direct schuin achter NAO zitten tijdens het experiment, maar neem wat meer afstand van de robot. Wanneer de trainer niet direct in het blikveld van het kind te zit,

wordt het kind opnieuw gestimuleerd zich op de robot te concentreren.

- Zorg ervoor dat interventie door de trainer tot een minimum wordt beperkt. Wanneer een kind NAO niet goed verstaat of niet heeft opgelet wanneer NAO iets zei, gebruik dan altijd eerst *maximaal twee keer de herhaal-optie (r)* alvorens wanneer dat nodig blijkt de uitspraak zelf te herhalen.
- Gebruik altijd de voorgeprogrammeerde prompts, zodat NAO degene is die het kind aanspoort tot interactie en niet de trainer. Ook deze prompts kunnen indien nodig herhaald worden, door nogmaals dezelfde selectie te maken in het key-press menu. *Herhaal de prompts maximaal een keer met behulp van het key-press menu.* Wanneer het kind ook na deze herhaling aangeeft NAO niet verstaan te hebben, herhaal dan zelf letterlijk de prompt in kwestie. Wanneer het kind niet reageert op een prompt of de herhaling daarvan, en er geen aanwijzing is dat het de prompt niet heeft verstaan, wordt de volgende prompt geselecteerd.
- Tijdens de experimentsessies zal geen gebruik worden gemaakt van de WoOZ-methode (waarbij de trainer NAO antwoord laat geven op onverwachte vragen of gedragingen van het kind door middel van ter plekke getypte text-to-speech). Dit zorgt ervoor dat de verschillende sessies zoveel mogelijk identiek aan elkaar zullen zijn. Wanneer het kind een vraag stelt of een antwoord geeft waarop NAO niet kan reageren, los dit dan op door zelf een reactie te geven richting het kind. Zorg er echter voor dat deze reactie zo neutraal mogelijk is, zowel qua inhoud als vorm, om de invloed van de trainer op de sessies tot een minimum te beperken. Voorbeelden van reacties die door de trainer kunnen worden toegepast:
  - "Volgens mij bedoelt NAO iets anders" (bij vragen/antwoorden anders dan voorzien)
  - "Daar kunnen we later over praten, maar nu maken we eerst het spelletje af" (bij vragen/antwoorden die onderwerpen aansnijden die momenteel irrelevant zijn)
  - "Goed dat je dat zegt" (voor momenten waarop NAO zelf niet kan belonen)
- Houdt, wanneer het kind een bakje met puzzels uit de bruine doos heeft gekozen, het deksel van dat bakje bij de hand/in zicht. Hierop staat een nummer, dat overeenkomt met het nummer van de te selecteren kleur in het key-press menu voor de kleurselectie van de zakjes in de blauwe doos. Voorbeeldscenario: het kind kiest bakje 2. Vervolgens komt op het einde van het spel het laatste key-press menu voor kleurselectie, dat er als volgt uitziet:
  - 1 = paars voor bakje 1
  - 2 = bruin voor bakje 2
  - 3 = donkerblauw voor bakje 3

  De trainer weet niet meer welk bakje het kind had gekozen, maar werpt een blik op het deksel waarop staat '2'. De trainer selecteert dan optie 2, d.w.z. het bruine zakje.
- Wees zoveel mogelijk consistent in de manier waarop met de kinderen gecommuniceerd wordt, om de kans op beïnvloeding van de resultaten te minimaliseren.

Instructies voor pilotsessies

Tijdens de pilotsessies is het belangrijk om de kinderen voorafgaand aan het spel te vertellen wat de bedoeling is. Zij hebben immers nog nooit eerder PRT gehad, en zijn dus onbekend met bijbehorende leerdoelen. Geef hen daarom van tevoren de volgende uitleg:

“Straks ga je een spelletje spelen samen met dit robotje. Die heet NAO, leuk ziet hij eruit hè? NAO weet precies wat de bedoeling is, en kan je tijdens het spelletje overal mee helpen. Maar robots zijn niet zo slim als kinderen, dus soms moet jij hem een beetje helpen. Hij vind het leuk als je dingen aan hem vertelt, en hoopt dat je veel vragen aan hem zult stellen wanneer je iets wilt weten of wanneer je wilt dat hij iets doet. Soms is hij eventjes stil. Dan hoopt hij dat jij misschien iets tegen hem zult gaan zeggen. Ik ben er alleen om NAO te helpen bij dingen die hij nog niet zo goed kan of weet, maar doe niet mee met het spelletje. Ben je klaar om te beginnen?”

Wanneer een kind ergens op reageert tijdens deze uitleg, kun je daar kort op reageren. Bijvoorbeeld:

Kind: “Hij ziet er gek uit.”
Trainer: “Dat klopt wel, maar hij weet [precies wat… zie hierboven]”

Kind: “Mag ik alles vragen?”
Trainer: “Ja, maar NAO weet alleen antwoord op vragen die over het spelletje gaan.”

Kind: “Waar moet ik bij helpen?”
Trainer: “Dat merk je straks vanzelf.”

Counterbalancing en inleidend verhaal

De experimentele condities zullen binnen de groepen ‘counterbalanced’ worden aangeboden. Concreet betekent dit het volgende voor de volgorde waarin participanten de condities zullen tegenkomen:

Groep x	participant 1	conditie 1 → conditie 2
	participant 2	conditie 2 → conditie 1
	participant 3	conditie 1 → conditie 2
	participant 4	conditie 2 → conditie 1

Aan beide participantgroepen zal, aan het begin van iedere sessie, moeten worden uitgelegd waarom NAO’s stem anders klinkt (hoger dan normaal in condities 1 en monotoon in condities 2). Aan groep 2 moet daarnaast worden verantwoordt dat NAO kleding draagt, terwijl dit normaal niet zo is. Hieronder wordt beschreven hoe dit aan het begin van iedere experimentsessie zal worden gedaan. De uitleg staat vast per groep en sessienummer, ongeacht met welke conditie de participant begint.

Groep 1, experimentsessie 1

"Vandaag ga je met een vriendje van NAO spelen. Dit robotje heet NEO. Gelukkig lijken NAO en NEO precies op elkaar: ze weten precies dezelfde dingen, en kunnen ook nog eens precies hetzelfde. Het maakt daarom eigenlijk niks uit met welk robotje je speelt. Het enige verschil is dat NEO's stem een beetje anders klinkt, dat zul je straks wel horen. Best grappig hoor, al die verschillende geluidjes die robots maken. Zullen we beginnen?"

Groep 1, experimentsessie 2

"Weet je nog dat je de vorige keer met robotje NEO speelde in plaats van met NAO? Vandaag is er weer een ander vriendje van NAO op bezoek: die heet NIO. Gelukkig is ook dit robotje precies hetzelfde als alle andere robotjes die je al kent. Alleen zijn stem klinkt weer een beetje anders dan je bent gewend van de andere robotjes, maar dat is bij mensen ook zo. Ja toch? Want papa praat ook anders dan mama. Eigenlijk wel grappig, al die verschillende stemmen. We gaan gauw beginnen, dan hoor je vanzelf hoe NIO's stem klinkt. Goed?"

Groep 2, experimentsessie 1

"Vandaag ga je met een vriendje van NAO spelen. Dit robotje heet NEO. Gelukkig lijken NAO en NEO precies op elkaar: ze weten precies dezelfde dingen, en kunnen ook nog eens precies hetzelfde. Het maakt daarom eigenlijk niks uit met welk robotje je speelt. NEO ziet er ook nog eens precies uit als NAO, alleen heeft 'ie kleertjes aan net als jij. En z'n stem klinkt een beetje anders dan die van NAO, dat zul je straks wel horen. Best grappig hoor, al die verschillende geluidjes die robots maken. Zullen we beginnen?"

Groep 2, experimentsessie 2

"Weet je nog dat je de vorige keer met robotje NEO speelde in plaats van met NAO? Vandaag is er weer een ander vriendje van NAO op bezoek: die heet NIO, en draagt ook kleertjes. Gelukkig is ook dit robotje precies hetzelfde als alle andere robotjes die je al kent. Alleen zijn stem klinkt weer een beetje anders dan je bent gewend van de andere robotjes, maar dat is bij mensen ook zo. Ja toch? Want papa praat ook anders dan mama. Eigenlijk wel grappig, al die verschillende stemmen. We gaan gauw beginnen, dan hoor je vanzelf hoe NIO's stem klinkt. Goed?"

Wanneer een kind tijdens de uitleg reageert, kun je daar kort op ingaan. Zorg er opnieuw voor dat gegeven reacties zoveel mogelijk overeenkomen tussen sessies en participanten. Bijvoorbeeld:

Kind: "Ik wil met NAO spelen."
Trainer: "Nee joh, met NEO/NIO wordt het net zo leuk."

Kind: "Waarom is NAO er niet?"
Trainer: "Omdat andere robots soms ook willen spelen."

Kind: "Dat is NAO."
Trainer: "Luister maar, dan hoor je het verschil."

Appendix 4: Type III tests of fixed effects (SPSS)

Average percentage of required prompts during communicative learning opportunities

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	6	14,808	,008
voice	1	6	,531	,494
appearance	1	6	,220	,656
voice * appearance	1	6	,803	,405

a. Dependent Variable: prompts_used.

Spontaneous verbal initiatives towards robot

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	6	37,537	,001
voice	1	6	,341	,580
appearance	1	6	1,055	,344
voice * appearance	1	6	,318	,593

a. Dependent Variable: verbal_robot.

Spontaneous nonverbal initiatives towards robot

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	6	7,527	,034
voice	1	6	,096	,767
appearance	1	6	1,342	,291
voice * appearance	1	6	1,699	,240

a. Dependent Variable: nonverbal_robot.

General affect scores

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	6	1231,073	,000
voice	1	6	1,124	,330
appearance	1	6	6,588	,043
voice * appearance	1	6	1,124	,330

a. Dependent Variable: av_affect.

Interest scores

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	6	540,250	,000
voice	1	6	2,487	,166
appearance	1	6	8,446	,027
voice * appearance	1	6	,177	,689

a. Dependent Variable: av_interest.

Happiness scores

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	6	1534,564	,000
voice	1	6,000	,188	,679
appearance	1	6	,001	,975
voice * appearance	1	6,000	9,936	,020

a. Dependent Variable: av_happiness.

Behavior scores

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	6	307,023	,000
voice	1	6,000	,275	,619
appearance	1	6	2,724	,150
voice * appearance	1	6,000	,017	,901

a. Dependent Variable: av_behavior.