Modelling the user's skill and performance with the use of a Bayesian rating system

Motivating children to play games with a social robot by providing the optimal challenge

Bob R. Schadenberg

December 2012

Master Thesis
Human-Machine Communication
Dept of Artificial Intelligence,
University of Groningen, The Netherlands

Internal supervisor:
Dr. Fokie Cnossen (Artificial Intelligence, University of Groningen)

External supervisor:
Prof. Dr. Mark Neerincx (Perceptual and Cognitive Systems, TNO)

university of groningen

TNO innovation for life

Abstract

Playing (educational) games with a social robot can provide a user with entertainment and a way of learning that is encouraging and engaging. For the robot to be effective, interaction with the robot should keep the child motivated to play the games with the robot for a longer period of time, when the initial novelty has worn off. One aspect that can affect the motivation of the user is the difficulty of the game. A game should be challenging, while at the same time the user should be confident to meet the challenge. We designed a user modelling system that adapts the difficulty of a game to the user's skill, in order to provide users with the optimal challenge. To this end, we used a Bayesian rating system to estimate the user's skill and performance. In the experiment, we used our user modelling system to test if users who are optimally challenged are more intrinsically motivated to play games with the robot, than users that are not optimally challenged. Furthermore, we evaluated whether the Bayesian rating system could be used to detect a loss of motivation to play the current game with the robot, by relating the expected performance to the actual performance. 22 children participated in the experiment, aged between 10 and 12 years old. Due to not having enough data, we were not able to achieve the measurement precision that is required to make reliable estimations of the probability of a participant answering an item correctly. Because the participants were not optimally challenged, we cannot answer whether the participants were more intrinsically motivated to play the games. Also, there were not enough events where there was a large discrepancy between the expected performance and the actual performance to conclude if and how reliable the detection of a loss of motivation to play the current game with the robot was. We discuss several improvements that can be made to the user-adaptive system.

Acknowledgments

I would like to thank the following people who have helped me to come to this thesis. First of all, I would like to thank Mark Neerincx for his advice and support, and for giving me the opportunity to do research at TNO. It is an amazing place to do research. And I would like to thank Fokie Cnossen for her support and feedback. Next, I would like to thank Bert Biermans for his assistance with the NAO. His help has been invaluable. I would like to thank the elementary school 'Griftschool' in Woudenberg, for allowing us to conduct the experiment. And last, I would like to thank my family and friends for always supporting me.

Table of Contents

Chapter 1

Introduction

Human-robot interaction is usually limited to short-term interaction, as users typically spend less than 10 hours with a robot, before losing interest. However, for applications that require the robot to interact with humans, it is desirable and often crucial for the efficacy of the robot, that the robot is able to sustain long-term interaction. Research of the past decade has focussed primarily on short-term interaction between human and robot. And thus far, long-term human-robot interaction is poorly understood, and there are no design paradigms or algorithms for designing a robot that can maintain a long-term interaction.

In this study, we use a social robot to support a child user in learning and provide entertainment, by playing (educational) games with the child. For the robot to be effective, interaction with the robot should keep the child motivated to play the games with the robot for a longer period of time, when the initial novelty has worn off. This can only be achieved when the robot is able to maintain long-term interaction. Baxter and colleagues (2011) argue that for robots to accommodate long-term interaction, they need to be able to establish a socio-emotional relation with its user. Such a relationship between human and robot can only be established when the human-robot interaction has a feeling of continuity in the long-term. A robot has to remember previous encounters with the user and adapt its behaviour accordingly. Only then can it maintain human-robot interaction after the initial novelty has worn off (Robins et al., 2010). To keep the child motivated to play the games with the robot, the games should be challenging (Deci & Ryan, 1985). The skill amongst children in any given group may vary significantly, so what one child may perceive as challenging another may perceive as being (too) easy. It can be discouraging for children when they perceive the game to be too difficult or too easy. Therefore, the difficulty of the games should be adapted to a personal level, so that each child may play the games at a difficulty that is challenging.

We designed a user-adaptive system with can be utilised by the robot to adapt the difficulty of the games to the child. The robot stores relevant information about the child in a user model, such as

how skilled the child is at playing a game, and uses the stored information to adapt the difficulty of the games and the child-robot interaction. In this study, we investigate whether children are more motivated to play the games with the robot, when the robot adapts the difficulty of the games and its behaviour to the children.

1.1 The ALIZ-E project

This study is part of the ALIZ-E project. ALIZ-E is a European funded research project (FP7-ICT-248116) and aims to move human-robot interaction from the range of minutes to the range of days. The mission statement of the project is "*to develop the theory and practice behind embodied cognitive robots which are capable of maintaining believable multimodal any-depth affective interactions with a young user over an extended and possibly discontinuous period of time*". The scientific methods that are developed for this project will be implemented using a humanoid robot. The robot will be used to support hospitalized children as they learn how to cope with a lifelong metabolic disorder (i.e. diabetes and obesity). The children are eight to twelve years old and have been recently diagnosed. Being hospitalized is often a fearful experience and even more so for children, who lack understanding of what is happening to them and why it is happening. They are suddenly in a different environment, full of strangers and often without their family and friends to keep them company. As a result, the child could feel lonely, depressed, fearful or abandoned. A robotic companion can make the stay in the hospital less discomfortable and also to help the child learn how to cope with their disease.

The ALIZ-E project involves a consortium of seven academic partners and two commercial partners. The role of TNO (Netherlands Organisation for Applied Scientific Research) in the ALIZ-E project is to develop and test user and task models. The user and tasks models are used to endow the robot with the ability to adapt its behaviour, both linguistic and non-linguistic, to the user, the task and the interaction history. For example, when the user is currently feeling depressed, the robot ought to recognize this and respond to it. Also, the robot should be able to give an empathic response to show support and be aware that now may not be the time to start educating the child on a topic the child should improve on.

Chapter 2

Theoretical Background

The user-adaptive system we designed draws upon theories from psychology, human-computer interaction, and artificial intelligence. We utilise motivational techniques to keep the children motivated to play the games with the robot. By adapting the difficulty of the game to the child, each child may play the games at the optimal challenge. A game is optimally challenging, when the difficulty of the game challenges the children, while at the same time the children believe that they are skilled enough to meet the challenge. Playing games at a challenging difficulty can be more satisfactory, as it may provide the child a feeling of competence. In order to adapt the difficulty of the game to the child, the robot has to estimate how skilled the child is at playing the game. There are several approaches to measuring skill. We opted to use a skill-based approach, where the robot estimates how skilled a child is at playing the game and adapts the difficulty of the game accordingly. The robot estimates the skill of a child with the use of a Bayesian rating system.

In this chapter we discuss the theoretical background of each topic. We begin by giving a general introduction to social robotics and user modelling. Next, we discuss the topic of motivation, the different approaches to the optimal challenge, and the Bayesian rating system that we used to estimate the user's skill. Last, we discuss the research questions of this study.

2.1 Social robotics

To date there are several definitions of a social robot, each with a different statement about the features that define it (Breazeal, 2003; Duffy, 2000; Fong et al., 2003; Hegel et al., 2009). All the definitions contain aspects about the appearance of a social robot and how it behaves in a social context. The appearance of a social robot contains features, like a face, which signal that the robot is a social interaction partner. Also, a social robot can behave socially to a certain extent, like being able to communicate with humans. Breazeal defines social robots by the effects of a social robot's appearance and function on a human observer: "*social robots are a class of autonomous robots to*

which people apply a social model to in order to interact with and to understand". People use social models to explain, understand and predict human behaviour. However, sometimes people use social models to explain the behaviour of living creatures or objects, when the observed behaviour is not easily understood in terms of its underlying mechanisms (Reeves & Nass, 1996). For example, we may attribute human characteristics, such as mental states (i.e. feelings, desires or intentions), to a computer, an animal, or a robot, in order to explain their behaviour or actions. This is called anthropomorphising.

People apply a social model to a social robot's behaviour, because its behaviour adheres to the social models; the robot appears to be, or is to a certain extent, socially intelligent. To achieve behaviour that adheres to a social model, the robot has to at least appear socially intelligent (Bates, 1994). In a constrained environment, with limited interaction with people, a robot can appear to be socially intelligent without being social intelligent. But, as the complexity of the environment increases, the social intelligence of the robot will have to scale accordingly. For social robots to perform well in human environments, the robot has to be genuinely socially intelligent, to the extent that a person can interact with the robot as if it were a socially responsive creature, like, for instance, one would interact with an animal. A social robot does not need to have a humanoid appearance in order for a social model to be applicable. What matters is how the robot interacts with people and how they interact with it.

To date, the field of social robotics is still in its infancy. Many of the tasks a social robot can perform can also be performed by other platforms, such as virtual agent or technology embedded in the environment. And compared to these other platforms, robots are generally expensive, are easily damaged, and can only perform a task under specific circumstances. However, a social robot can offer advantages not found in other platform. Robots are different from virtual agents in that robots have a physical body with which they can interact with their environment. As a result, people respond to robots in a different way than to a virtual agent. In a study conducted by Kiesler and colleagues (2008), participants rated the character traits (i.e. trustworthiness, respectfulness) of a robot and a virtual agent. The character traits of the robot were rated more positively than that of the virtual agent. Similar results were found by Komatsu and Abe (2008). Besides rating robots as more trustworthy than a virtual agent, people are also more likely to trust a robot (Naito & Takeuchi, 2009).

Trust is a key factor for the acceptance of technology and is critical for forming interpersonal relationships (Lee & See, 2004). When a robot is used to achieve behavioural change, it is essential that a positive relationship exists between the person and the robot as it allows the robot to be more persuasive (Fogg, 2002). A robot is also more likely to influence a person than a virtual agent, because of its physical proximity (Kidd & Breazeal, 2004; Powers et al., 2007).

Children view robots differently than adults. Tanaka and colleagues (2007) immersed a social robot in a classroom for 45 sessions, each lasting approximately 50 minutes, over a period of five months. They found that the children came to perceive the robot as a peer, rather than a toy. The children exhibited a variety of social and caretaking behaviours towards the robot. In another study, Tanaka and Ghosh (2011) used their social robot in a care-receiver role. The children exhibited various caretaking behaviours, and began teaching the robot when it started making mistakes. Tanaka and colleagues (2007) argue that children perceive a social robot as a peer, because of their stronger tendency to anthropomorphize and increased use of imagination. As a result, children are more likely to suspend disbelief and engage with the robot.

Social robots can perform various roles, such as a motivator, educator, or companion. By using persuasive technology and applying motivational techniques, a social robot can motivate users to change their behaviour or to adhere to some sort of program. In their study, Fasola and Matarić (2012) used a social robot called Bandit (see Figure 1), which was designed to motivate elderly to do physical exercise. The robot incorporated behaviours aimed at increasing the motivation of the participants (i.e. giving positive feedback) and relational discourse (i.e. politeness, humour, or empathy), which contribute to the development of a meaningful relationship between the user and robot. Fasola and Matarić compared a relational robot, which incorporated all the motivational behaviours, with a non-relational robot, which did not incorporate the motivational behaviours. The participants showed a strong preference for the relational robot; they rated the robot higher in terms of enjoyableness, companionship, and as an exercise coach, and the robot was also able to motivate the participants during the exercises. Kidd and Breazeal (2008) designed a social robot, called Autom (see Figure 2), that functioned as a weight loss coach, using the robot's ability to engage the user and

Figure 1. Bandit.

Figure 2. Autom, the weight loss coach.

engendering trust, to motivate the user to reach the goals related to losing weight. They compared the robotic weight loss coach with a computer running identical software and a paper log. The participants using the weight loss coach used it for 50 days on average, while participants with the computer used their system for 36 days on average, and participants with a paper log reached an average of 26 days.

Social robots can also be used for educational purposes. They may be designed to teach others (Hashimoto, Kato & Kobayashi, 2010), assist a teacher (Chang et al., 2010), or serve as an educative companion (Kanda et al., 2004). The use of social robots as educators is promising, as effective methods used in computer-based learning can be combined with an increased engagement, persuasiveness, and motivation which a social robot can offer. Saerbeck and colleagues (2010) found that engaging in social interaction, compared to just focusing on knowledge transfer, had a positive effect on learning. They developed a robotic tutor, using Philip's iCat (see Figure 3), that could help children learn an artificial language. In addition, the robot was capable of socially supportive behaviours, such as empathic and motivational responses. They compared how well the participants learned the language, using either the socially supportive robot or the robot that did not show socially supportive behaviour. The children who learned the artificial language with the socially supportive robot scored better on the assessment test than the participants who studied with the non-socially supportive robot.

Social robots can serve as a companion to a person, in order to increase a person's health and psychological well-being. Using a companion robot can benefit people who are, for example, depressed, going through a hard time, or feel isolated. An example of a social robot used as a companion is the seal-like Paro robot (See Figure 4), which was used by Wada and colleagues (2005).

Figure 3. The iCat.

Figure 4. The seal-like Paro.

In their study, elderly people residing in a health service facility interacted with Paro for one hour per day, over two weeks. At the end of the study, the elderly showed improved moods, improvements in depression, a decrease in stress level, and they communicated more often.

Researchers have successfully designed social robots to perform the roles of a motivator, educator, and companion, which illustrates the potential of social robots. However, today's social robots all have severe limitations that limit their effectiveness. They are still unable to robustly perceive and understand humans and the environment, which severely limits the human-robot interaction. Moreover, social robots have been unable to engage in social exchanges extending beyond the scale of minutes and to adapt their interactive behaviour on the basis of previous encounters with a person. If social robots are to enter our daily lives, they have to be able to maintain long-term interaction if they are to be effective.

2.2 User-adaptive systems

A social robot's ability to adapt its behaviour is a key factor for maintaining human-robot interaction over longer periods of time (Gockley et al., 2005)**.** The robot has to adapt its behaviour to the user, the environment, the task, and the interaction history. To do so, the robot needs to observe the behaviour of the user and make generalizations and predictions about the user. The acquired information is used to build a user model[1], which is a model that contains user information associated with a specific user and represents the that user. The process of building up and modifying a user model is referred to as user modelling. User models may be used for various goals (for a review, see Taatgen & Johnson, 2005), such as predicting user behaviour, or to gain knowledge about the user. A user-adaptive system is a user modelling system that utilised the user model to adapt a system to the user.

A user-adaptive system consists of three major steps, namely the acquisition of information, the representation of information, the evaluation of the information.

2.2.1 Acquisition of information

A user-adaptive system requires information about the user and the user's environment to determine when the system should be adapted, what part of the system should be adapted, and how it should be adapted. Information can be acquired explicitly, by asking the user questions, or implicitly, by observing the user's behaviour and making inferences based on stored knowledge. Any

[1] In some studies, the term "user model" is used to refer to the user-adaptive system as a whole. However, in this study the term is only used to refer to the model that contains the user information related to a specific user.

information can be used by a user-adaptive system, as long as it has enough predictive value to be of use, and can be measured with enough accuracy. Often, user-adaptive systems are based on information about the user (user data) and the context. User data can include the user's knowledge, skills, capabilities, interest, preferences, goals, plans, or demographic variables. Information about the context can also be a source for adaption and includes information about the task and the interaction history. Some kinds of the information can be acquired directly, such as observing the user's sex or asking the user's name. However, most kinds of information have to be inferred from observable behaviours and other variables, such as the user's motivation, before they can be used for adaptation.

2.2.2 Representation of the information

In its simplest form, a user model is a set of variables with certain values. More complex user models can use various methods to structure the information, so that secondary inferences can be drawn. Secondary inferences operate on the contents of the user model. There are many different user modelling techniques that can be employed to structure the information and draw inferences from it (Kobsa, 2001). Logic-based methods of representing information can be used for deductive reasoning (Pohl, 1999). For instance, if a user is motivated to play game A, the user will also be motivated to play game B. If the user model contains an entry that the current user is motivated to play game A, it can then be inferred that the current user will also be motivated to play game B. A shortcoming of logic-based methods is that they are generally not able to deal with representing uncertainty and maintaining what is true at a certain point in time. On the other hand, probabilistic user model representations are able to represent uncertainty. These include methods like Bayesian networks, linear parameters, or fuzzy logic.

The information about the current user can also be compared with that of similar users, in order to predict unknown characteristics. This is known as clique-based filtering and may operate according to a correlation-based approach, clustering algorithm, vector-based similarity technique, or a Bayesian network.

2.2.3 Evaluation of the information

When there is enough information about the user and environment, the user-adaptive system must decide on how it should respond, given the current information. Some information can be used without evaluation, like the user's name, or the user's preferences. Other information needs to be evaluated before they can be used for adaption, for which selection rules can be used. Selection rules are conditional statement that specify that when the condition of the statement is true, given the available information, adapt X.

For more complex reasoning, a decision model may be employed, such as a rational agent system. A rational agent is an autonomous entity which observes and acts upon the environment and directs its activity towards achieving goals (Russell & Norvig, 2003). Rational agents have a decision making component that governs its decisions based on its informational (beliefs/distributions/knowledge) and motivational attitudes (goals/desires/utilities/preferences) (Dastani, 2011). A rational agent has to find a balance between pursuing its goals (proactive behaviour) and reacting to the environment (reactive behaviour).

The agent's sub-goals can sometimes contradict each other. The contradiction can be solved by comparing which sub-goal aids the main goal the most. For example, when the main goal of the agent is to educate the user, and its sub-goals are to keep the user motivated to play games and to play the game at which the user performs under par. When the agent beliefs that the user is not motivated to play games anymore, it has to decide whether the sub-goal of playing the game at which the user performs under par will still further the main-goal of educating the user, or whether increasing the user's motivation to play games should be prioritised. To answer this question, the agent will make a decision based on its informational and motivational attitudes.

2.3 Intrinsic motivation

The goal of our user-adaptive system is to keep children motivated to play games with the social robot. To achieve this goal, we need to know what factors can influence a child's motivation and how these factors can be influenced by a social robot.

People have different amounts of motivation, but also different kinds of motivation. Two types of motivation can be distinguished, namely intrinsic and extrinsic motivation (Ryan & Deci, 2000). Intrinsic motivation refers to participating in an activity because it is inherently interesting or enjoyable. On the other hand, extrinsic motivation refers to participating in an activity because it leads to a separable outcome, such as a reward or the approval of others.

Deci and colleagues (1999) conducted a meta–analysis of 128 studies to examine the effects of extrinsic rewards on intrinsic motivation. They concluded that enhancing extrinsic motivation by rewarding desirable behaviour, such as playing games with a robot, is effective for promoting the desired behaviour in the short-term, but may have a negative effect on promoting the desired behaviour in the long-term, as it undermines a person taking responsibility for motivating or regulating oneself. The goal of our study is to motivate users to play games in the long-term. Therefore, we focus on enhancing the intrinsic motivation of the user to play games with the robot.

Children will only be intrinsically motivated for activities that hold intrinsic interest for them (i.e. activities that are novel, challenging, or hold aesthetic value). For such activities, the social environment can either facilitate or forestall intrinsic motivation. According to the Self Determination

Theory (SDT; Deci & Ryan, 1985), a macro theory on human motivation and personality, people have three innate psychological needs that are the basis for self-motivation, namely the need for competence, autonomy, and relatedness. The Cognitive Evaluation theory (Deci & Ryan, 1985), a sub-theory of the SDT, states that activities that support the psychological needs of people can facilitate intrinsic motivation, given that the activity holds intrinsic interest to begin with. Conversely, thwarting the psychological needs of people can forestall intrinsic motivation. The SDT is also applicable to children, as studies have shown that a child's perception of competence is positively related to the child's intrinsic motivation (Boggiano, Main & Katz, 1988; Gottfried, 1990). Furthermore, children are more likely to be intrinsically motivated when the context is characterized by a sense of security or relatedness (Grolnick & Ryan, 1986).

Several factors have been identified that may serve to enhance intrinsic motivation. According to the Cognitive Evaluation Theory, praise can be used to enhance intrinsic motivation by promoting a greater perceived competence. Anderson and colleagues (1976) found that giving praise to children increased the time the children would engage in a task, relative to baseline and to groups of children that were given money or symbolic rewards. Henderlong and Pepper (2002) conclude that provided the attributional message is perceived as sincere, praise is likely to enhance intrinsic motivation under certain conditions, namely when the praise prevents maladaptive inferences, when autonomy is promoted, when perceived competence and self-efficacy are heightened without undue use of social comparison, and when realistic standards and expectations are conveyed.

Competition can also be used to enhance intrinsic motivation, as it can possibly give a person a feeling of competence. The competition can be direct, when the competition is between people, or indirect, when a person competes against an ideal outcome, such as that person's high score on a game. A direct competition can both increase and decrease a person's intrinsic motivation, depending on the outcome of the competition (Reeve & Deci, 1996). In case a person won, intrinsic motivation was increased, and for those who lost, intrinsic motivation was decreased. Furthermore, a direct competition may cause a person to feel pressured to perform well, which decreases intrinsic motivation. Indirect competition has been shown to increase user enjoyment in an otherwise non-competitive task (Weinberg & Ragan, 1979).

According to Deci and Ryan (1985), people are intrinsically motivated under conditions of optimal challenge, because it promotes a greater perceived competence. When a person starts playing a game, he or she may find playing a game at the basic level is satisfactory, because it matches the person's skill level. But one cannot enjoy playing the game at the same level of difficulty for long. For a skilled-based game, a person can become more skilled at playing the game. In turn, the game will become boring, because the game has become too easy. Conversely, when a person starts playing a game at a difficulty that proofs to be too difficult, the person might feel anxious due to the poor

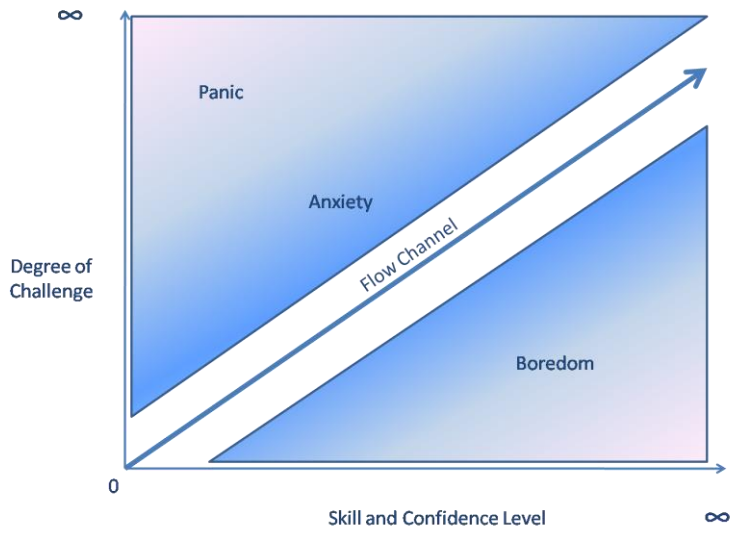


Figure 5. The relation between perceived skill and challenge, according to the Flow Theory.

performance. Rather, the difficulty of the game must be challenging, but manageable for a person. As people generally get more skilled at a game, the longer they play it, the difficulty of the game must be continuously updated to reflect the skill of a person.

The flow theory (Csikszentmihalyi, 1990) describes the relation between how skilled a person perceives himself to be and the degree of challenge, as can be seen in Figure 5. When an activity is challenging and a person beliefs that his or her skill is enough to meet the challenge, it is possible to experience flow. Flow can be described as a state in which a person is so involved in an activity that nothing else seems to matter. The activity is enjoyable to such an extent that a person is intrinsically motivated to participate in the activity. People experience flow when participating in an activity for which they have a sense that one's skill is adequate to cope with the challenges of the activity. One is fully concentrated on the activity, so that no attentional resources are left to think about anything irrelevant. And as a result, self-consciousness disappears and the sense of time becomes distorted.

Although it is possible to experience flow while engaged in any activity, some activities are more likely to elicit flow than others. Csikszentmihalyi argues that a person will enjoy an activity for its own sake, when an activity is able to limit the stimulus field so that the person can act in it with total concentration, responding to greater challenges with increasing skills, and when it provides clear and unambiguous feedback.

2.4 Different approaches to the optimal challenge

The optimal challenge can be assessed with affect-based or skill-based models. Affect-based model adjust the difficulty of the game to keep the user's affective state around a certain level, which is deemed as the optimal level. For example, the state of flow is characterized by low levels of

anxiousness and boredom, and a high level of engagement. The affective state of a user can be detected via different modalities, such as the user's facial expressions, vocal intonation, gestures, body language, and physiological responses (Calvo & D'Mello, 2010). Liu and colleagues (2009) adapted the difficulty of a game based on the anxiety level of a person. They used regression trees to determine the intensity of the affect states from a set of features derived from physiological signals, including the user's electromyographic, cardiovascular, and electrodermal activity. Their model was able to differentiate between three levels of anxiety, and recognised the levels correctly for 78%. Liu and colleagues compared their affect-based model with a basic model that adjusted the difficulty of the game based on the user's performance. They observed lower anxiety levels, a greater improvement in performance, and a greater subjective sense of challenge when the game was adjusted by the affect-based model.

While we believe that affect-based methods show promise in providing users with the optimal challenge, there are several limitations that make affect-based models unsuitable to be applied to a robotic platform. The machine learning techniques that are used to recognise the user's affective states generally require a training session before the affective states can be recognised. Also, recognition of the user's affective states is not very reliable and precise, as only a few levels of intensity of the affective states are distinguished and are recognised correctly for only 60% to 90%. Furthermore, when physiological signals are used, the user has to wear physiological sensors, which can restrict movement and may be (very) uncomfortable.

Skill-based models adapt the difficulty of the game based on how skilled the user is at playing the game. The user's skill is an example of a latent trait (a trait that cannot be directly measured) and therefore has to be estimated. Skill can be estimated as a holistic construct, or be decomposed into the procedural knowledge (i.e. strategies) and declarative knowledge (i.e. facts) that can contribute to the user's performance. The latter approach is used by Intelligent Tutoring Systems (ITS). ITS have a domain model, which contains the procedural and declarative knowledge required to solve a problem, and contains all problems that can be encountered. A problem (from now on referred to as an item) may be analogue to a question which the user needs to answer, an equation that needs to be solved, or difficulty setting such as the speed of the game. Besides a domain model, an ITS also models to what extend the user has mastered a piece of procedural or declarative knowledge: the student model. By comparing what the user knows and what is required to solve a certain item, it is possible to estimate what the likelihood is that the user will solve the item. This way, challenging items can be selected and presented to the user. The domain and student models have proven to be successful in estimating how likely it is that the user will solve an item. For example, Koedinger and colleagues (1997) developed an ITS with which students could learn algebra. They found that

students that used the ITS in classroom settings performed much better on traditional math tests than students that followed the same curriculum without the ITS.

The benefit of decomposing the construct of skill into procedural and declarative knowledge is that a greater insight into the user's skill can be gained. Theoretically, this may lead to more accurate estimations compared to when skill is estimated as a holistic construct. In practice, the accuracy of the estimations depends for a large part on how well the domain is modelled. Especially for complex domains this may proof a difficult and time consuming task. This is of practical concern to how easily new games can be developed and raises the question of whether this approach actually leads to more accurate estimations.

Rating or ranking system can be used to estimate the skill of the user as a holistic construct. These models use a numerical rating to represent a user's skill level. The difficulty of the game is set based on the estimated skill of the user (the user rating). After each instance of the game (e.g. answering an item or finishing part of the game), the user rating is adjusted based on the outcome of the instance. If the outcome is correct, then the user rating will be increased, and if the outcome is incorrect, the user rating will be decreased. Thus, if the preceding item was answered incorrectly, it may have been too difficult and the present item will be less difficult, as the estimate of the user's skill level will have been adjusted downwards. This way, rating systems can be used to adjust the difficulty of the game based on the skill of the user.

Rating systems are used in sports, such as chess, football and basketball, and online video-games, such as League of Legends, World of Warcraft, to selects players of equal skill to play with or against each other. Rating systems are not only used to match players with other players, but can also be used to match a player with an item of a certain difficulty. Klinkenberg, Straatmeier and Van der Maas (2011) let children practice math with a computerized educational game, and used a rating system to select items for which the children had a 75% chance of answering the item correctly. They found that 33% of the items were answered after school hours and during the weekend, which suggests that the children were motivated to play the game. Also, a child's skill at math did not appear to have any effect on how frequent the child played the math game.

For our study, we opted to use a rating system to estimate the user's skill, because rating systems can have several advantages that make it suitable to be applied to a robotic platform. Rating systems are capable of achieving a high measurement precision (Glickman, 1999; Klinkenberg, Straatmeier & Van der Maas, 2011), and the measurement precision will generally increase the more items are answered. Thus it should be possible to accurately assess which item is optimal in a certain circumstance. Also, the user's skill can be estimated covertly, so that the child-robot interaction is not interrupted. This is critical to keep the child-robot interaction as naturalistic as possible. Furthermore, rating systems are non-domain specific and thus can be applied to any skill-based

application by changing a few parameters. However, for some applications getting reliable item ratings can proof difficult, namely when answering an item takes a long time, as each item will have to be answered many times before the item rating is reliable. Thus, while in theory a rating system can be applied to any skill-based application, in practice it may not be feasible to use a rating system for certain skill-based applications. Another advantage of using a rating system is that the user's skill is estimated based on the outcome of an instance, which is easily measured, compared to the advanced techniques that are required to estimate the user's affective state.

2.5 Rating systems

We will first discuss the Elo Rating system (Elo, 1978) which forms the foundation for contemporary rating systems, followed by the Bayesian Glicko rating system (Glickman, 1999), which is the rating system used in this study. These two ratings systems are not the only rating systems that exist, as there are other rating systems which have been designed for different purposes. For example, Microsoft's TrueSkill (Herbrich, Minka & Graepel, 2007) is a rating system that is specifically designed to match a group of players with another group, based on the players' individual skill.

2.5.1 The Elo Rating System

The Elo rating system is a probabilistic model for estimating skill levels, and was originally designed to rank chess players and pair them with opponents based on the ratings. The Elo rating system works as follows. All users start out with a certain numerical rating, which represent that user's estimated skill level. If any user information that correlates with the user's skill is available (i.e. the age of the user), then that information can be used to set the initial rating to increase its accuracy. If no such information is available, a default rating is used. A rating is also assigned to each item, which represents its level of difficulty. The ratings generally range from 0 to 3000, with higher ratings meaning a higher skill/difficulty level.

When the initial ratings are set, the user is paired with an item, based on a selection algorithm that uses the ratings (e.g. minimizing the difference between the user's rating and item's rating). After an item has been answered, the rating of both the user and item are updated, based on the outcome of the instance. The rating (r) is updated using the following formula:

$$r_{\text{new}} = r_{\text{old}} + \text{K}(s - \text{E}(s|r))$$

where K is a constant that governs how much a rating can change in one instance, *s* is the outcome of the instance, which can be 1, when the user answers the item correctly, 0 when the user answers incorrectly, or 0.5 when the answer is neither correct nor incorrect, and E(s|r) is the expected

outcome. For the user, the expected outcome is the probability of answering the item correctly, given the rating of the item. And in case of the item, the expected outcome is the probability of the user answering incorrectly, given the user's rating. The expected outcome for the user can be calculated using the following formula:

$$\mathrm{E}(s|r_{\mathrm{user}}) = \frac{10^{\frac{r_{\mathrm{user}}}{400}}}{10^{\frac{r_{\mathrm{user}}}{400}} + 10^{\frac{r_{\mathrm{item}}}{400}}}$$

The same formula can be applied to calculating the expected outcome for the item, when r_{user} is substituted by r_{item} and vice versa.

When the discrepancy between the user's rating and the item's rating is small ($r_{user} \approx r_{item}$), the probability of the user answering correctly will be close to 0.5; the user is expected to give the correct answer approximately 50% of the time. When the discrepancy becomes larger, it is estimated that one side (the user or the item) has a greater probability of "winning". Winning means the user answering the item correctly, in case the user had the higher rating, or the user answering the item incorrectly, in case the item had the higher rating. In Figure 6 shows the relation (the s-curve) between the difference between the user and item rating and the probability.

The estimated probability is taking into account by the rating update formula. It does so by increasing the difference between the old and new rating, when the discrepancy becomes larger. For example, when the user has to answer a difficult item (an item with a higher rating than the user) the odds will be against the user. As a result, the user will be "rewarded" with a greater increase in

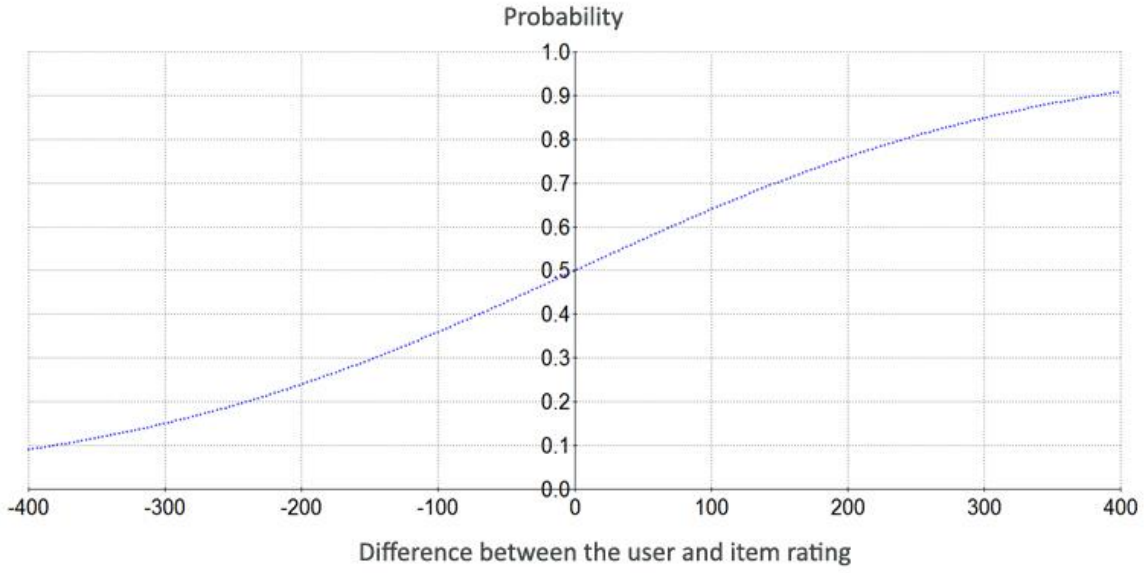


Figure 6. The difference of the user and item rating in relation to the probability.

rating, when giving the correct answer. Also, the decrease in rating will be diminished when the user answers incorrectly. For easy items (items with a lower rating than the user), it is the other way around; a greater decrease in rating when the user answers incorrectly, and a smaller increase in rating when the user answers correctly. The accuracy of the expected outcome depends on the accuracy of the rating of the user and of the item (e.g. how close they are to the user's true rating/item's true difficulty). Because the ratings are adjusted after each instance, the Elo rating system is a self-correcting system. That means that the ratings will generally become more reliable estimates the more instances occur.

A simulation of the Elo rating system can be seen in Figure 7. The simulation shows how the rating (the black line) develops on average the more items are answered. The simulation simulated the data of 30000 simulated users. The initial rating was set at 1500, while the simulated users had a true rating of 1700. The true rating (the yellow dotted line) is the actual (unknown) numerical representation of a user's skill: the user rating is the estimate of the user's true rating. The true rating is used to generate the outcomes of the instances. For each instance, the user was paired with an item of exactly the same rating as the user. The green line shows the spread of the individual ratings (±2*standard deviation). For this simulation, it is assumed that no learning occurs over time;

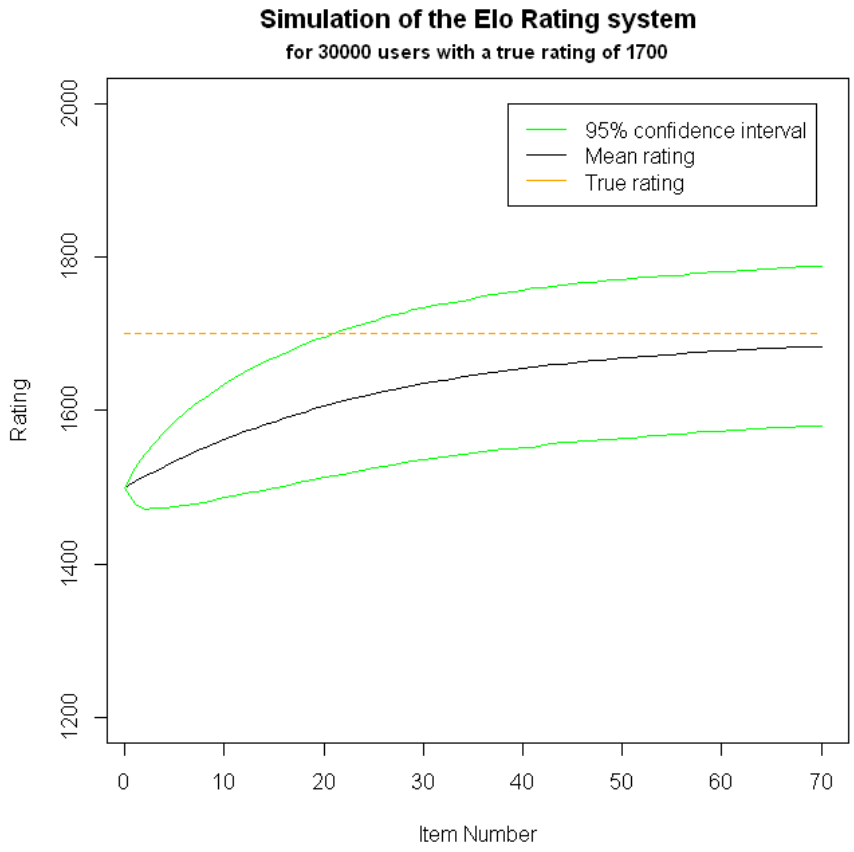


Figure 7. A simulation of the Elo rating system for a user with a true rating of 1700.

the true rating remains at 1700. The simulation shows that, on average, the ratings will rise steadily from 1500 rating to approximate the user's true rating. To close a difference between the initial rating and the user's true rating of 200, the rating system needs about 70 items answered on average. At this point, the difference between the estimated probability and the actual probability of the user answering an item correctly differs by approximately 1.5%.

2.5.2 The Glicko rating system

The Glicko rating system extends the Elo rating system by taking the uncertainty about the user's and item's rating into account. The uncertainty is represented by the rating deviation (RD), which is the estimated standard deviation of the rating. A high rating deviation indicates that the user has not played the game (much), or that it has been a long time since the user last played the game. A low rating deviation indicates that the user has played the game to such an extent that the rating is assumed to be reliable.

In the Glicko rating system, all users and items start out with an initial rating and rating deviation. If any user information is used to improve the accuracy of the initial rating, then the rating deviation can be adjusted downwards, because there is more certainty that the user's rating is close to the user's true rating. The rating deviation is decreased after each instance, because with each instance more information is gained regarding the true skill of the user and the true difficulty of the item. As time passes and the user has not played the game, the user could have become more skilled or less skilled at the game. To reflect the increase in uncertainty regarding the user's true skill, the rating deviation increases as time passes by. The rating deviation of items does not increase due to the passage of time, unless it is assumed that certain items can become more or less difficult over time. For example, when a certain class of math equations is no longer included in the curriculum, it can be expected that such equations become more difficult over time.

The rating updating formula of the Glicko rating system takes the rating deviation of both the user and item into account. If the user's deviation is large, the difference between the old and new rating will be larger, because there is still much uncertainty regarding the true skill level of the user. This allows ratings to increase or decrease quickly when the rating deviation is high, which is especially useful when the initial rating differs greatly from the true rating. For example, when the initial rating of a user is set to 1500, while the user's true rating is 2700, the Elo rating system will take a long time to approximate the user's true rating, because the increase in rating only depends on the difference between the user's rating and the rating of the item. As a result, the user will have to answer a lot of easy items, before the items are of a difficulty that matches the user's skill. When the Glicko rating system is used, the true rating can be approximated much earlier, because the increase in rating is much larger due to the high initial rating deviation.

When the user rating is adjusted after an instance, the rating deviation of the item is also taken into account. If the user's rating deviation is small and the item's rating deviation is large, the instance will have a smaller effect on the user's rating, because the predicted outcome may not be a reliable estimate. The reverse is true for adjusting the rating of the item after an instance. If the item has a small rating deviation, while the user's rating deviation is large, the change in rating for the item will be diminished.

The Glicko rating system uses the following formula's to adjust the rating and rating deviation of the items and users. If the user has not played the game before, a default deviation is used. If the user has played the game before, the user's old deviation is used and updated using the following formula:

$$\mathrm{RD} = \min\left(\sqrt{\mathrm{RD}_{\mathrm{old}}^2 + c^2 t}, 350\right)$$

where *t* is the number of time intervals since the last time the user played, *c* is a constant that controls the increase of variability over time, and 350 is the default rating deviation.
After each instance, the rating deviation is updated with the following formula:

$$\mathrm{RD}_{\mathrm{new}} = \sqrt{\left(\frac{1}{\mathrm{RD}_{\mathrm{old}}^2} + \frac{1}{d^2}\right)^{-1}}$$

The adjusted rating can be computed with:

$$r_{\mathrm{new}} = r_{\mathrm{old}} + \frac{q}{1/\mathrm{RD}_{\mathrm{user}}^2 + 1/d^2} g(\mathrm{RD}_{\mathrm{item}})(s - \mathrm{E}(s|r_{\mathrm{user}}, r_{\mathrm{item}}, \mathrm{RD}_{\mathrm{item}}))$$

where

$$q = \frac{\log 10}{400}$$

and *g*(RD) is the variable that reduces the change in rating, due to the uncertainty of the "opponent", which can be computed with:

$$g(\mathrm{RD}) = \frac{1}{\sqrt{1 + 3q^2(\mathrm{RD}^2)/\pi^2}}$$

and $\mathrm{E}(s|r, r_{\mathrm{i}}, \mathrm{RD}_{\mathrm{i}})$ is the expected outcome, given the rating, the rating of the opponent and the deviation of the opponent. The formula for the expected outcome is:

$$\mathrm{E}(s|r, r_{\mathrm{i}}, \mathrm{RD}_{\mathrm{i}}) = \frac{1}{1 + 10^{\frac{-g(\mathrm{RD}_{\mathrm{item}})(r_{\mathrm{user}} - r_{\mathrm{item}})}{400}}}$$

The variability that is due to the outcome of the instance can be computed with:

$$d^2 = \left(q^2\left(g(\mathrm{RD}_{\mathrm{item}})\right)^2 \mathrm{E}\left(s|r_{\mathrm{user}}, r_{\mathrm{item}}, \mathrm{RD}_{\mathrm{item}}\right)\left(1 - \mathrm{E}\left(s|r_{\mathrm{user}}, r_{\mathrm{item}}, \mathrm{RD}_{\mathrm{item}}\right)\right)\right)^{-1}$$

A simulation of the Glicko rating system can be seen in Figure 8. Similar to the simulation of the Elo rating system, the simulation of the Glicko rating system contains the data of 30000 simulated users, starting at a rating of 1500, and with a true rating of 1700. The simulated users started out with 350 rating deviation, and all the items had a rating deviation of 30. For each instance, the user was paired with an item of exactly the same rating as the user. For this simulation, it is assumed that no learning occurs over time; the true rating remains 1700. The black line is the average rating, the dotted yellow line is the user's true rating, and the green line shows the spread of the individual ratings (±2*standard deviation). The Glicko rating system requires about 20 items to be answered by the user, in order to close a rating difference of 200 between the initial rating and a user's true rating, given that the initial rating deviation is set at 350. Compared to the Elo rating system, fewer items have to be answered by the user when there is a difference between the initial rating and a user's

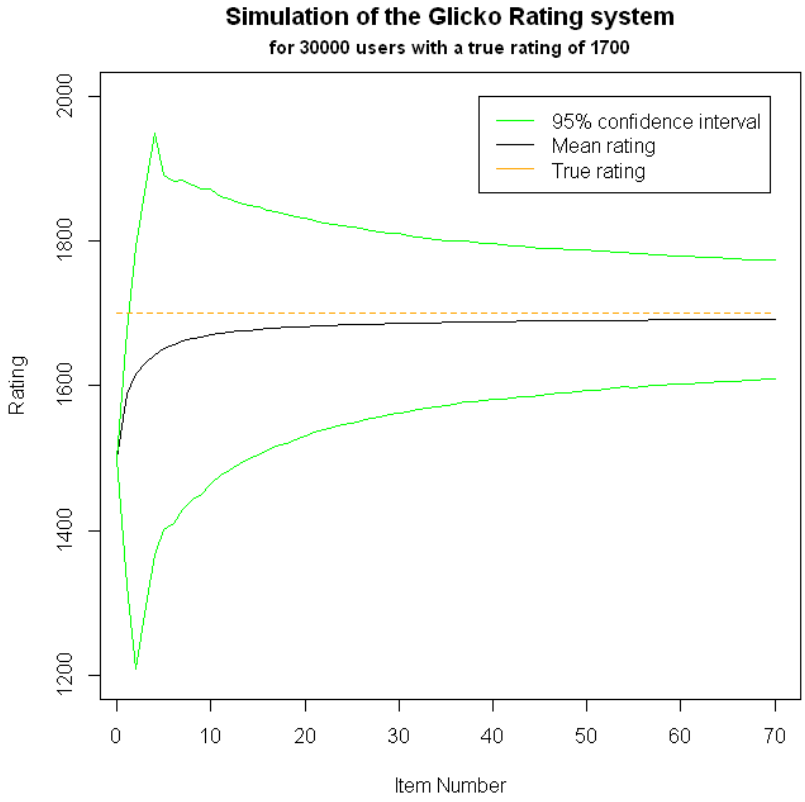


Figure 8. A simulation of the Glicko rating system for a user with a true rating of 1700.

true rating. This advantages comes at the cost however; the individual user ratings can deviate from the true rating to a larger extend, as can be seen by the large spread of the user ratings when few items have been answered.

2.6 Research Questions

The research questions this study attempts to answer are:

1) To what extend will children be intrinsically motivated to play games with a social robot, when the games are optimally challenging and exceptional performance is praised by the social robot?
2) To what extend is a sudden drop in performance related to the child's motivation to play the current game with the social robot?

To this end, we developed a user-adaptive system that can be utilised by a social robot to adjust the difficulty of the games to the child, in order to provide the child with the optimal challenge. Moreover, the system can be used to discern when a child is performing exceptionally well or exceptionally poor. The robot will praise the child when the child is performing exceptionally well, and will assume that the child is no longer motivated to play the current game when the child's performance is exceptionally poor. The goal of providing children with the optimal challenge and praising them for exceptional performance is to facilitate intrinsic motivation, so that the children will stay intrinsically motivated to play the games with the social robot once the initial novelty has worn off.

To answer the first research question, we will measure how intrinsically motivated children are when they play games with a social robot that utilises the user-adaptive system, and compare it to how intrinsically motivated children are when the games are overly challenging and no praise is given. We hypothesise that significantly more children will be intrinsically motivated when the user-adaptive system is utilised by the social robot, to provide the children with the optimal challenge and praise them when they perform exceptionally well.

To answer the second research question, we will relate exceptionally poor performance with the motivation of the child. We hypothesise that when a child's performance is exceptionally poor, given the skill of the child, the child is no longer motivated to play the current game with the social robot.

Chapter 3

Design of the User-Adaptive System

The goal of our user-adaptive system is to keep children intrinsically motivated to play games with the social robot for a longer period of time. The user-adaptive system consists of three components. A rating system with which the skill of the user can be estimated, a user model for the user specific information, and a rational agent that decides how the robot should adapt. In this chapter, we will discuss how we designed each of these components and what parameters we used. Furthermore, we will discuss the two games, a math game and an imitation game, that can be played with the robot.

3.1 The rating system

The user-adaptive system makes use of the Glicko rating system to estimate the skill of the user and the difficulty of the items. In theory, the user has to answer a relatively small number of items for the Glicko rating system to reliably estimate the user's skill. Thus, it should be possible to provide the user with the optimal challenge relatively fast. However, this is only possible when estimates of the difficulty of the items are reliable. The less reliable the estimation of the difficulty of the items, the more items the user will need to answer before the user rating is a reliable estimate of the user's true rating. The Glicko rating system is a self-correcting system, as it will adjust the ratings after each instance, and thus, it can accommodate changes in difficulty and skill. This is important, because for a system that is to be used for a longer and possibly discontinuous period of time, changes in the skill of a user or the difficulty of an item are expected to occur.

The initial rating deviation for users was set at 350, because we assume that the user's true rating lies within 700 rating of the initial rating. We make this assumption based on the spread of the initial item ratings (which are discussed in section 3.5). For the math game, each 300 rating corresponds to a year of learning math at an elementary school. And for the imitation game, the movement sequence is (generally) increased with one additional movement per 300 rating. Thus, setting the rating deviation of the user at 350 gives us a large margin of error for the initial user rating. The initial

user ratings were set at different values per person and per game and were based on the available user information. The *c* parameter, used for calculating the rating deviation when a new session is started, is set at 18.132. We chose this value so that the rating deviation will equal the default rating deviation when the user has not played the game for a year or more.

3.2 User model

The user model in our adaptive-user system is very basic. It contains a few user-specific variables and no user modelling techniques are used to draw secondary inferences. The user model contains the following information: the user's rating and rating deviation, the date the user last played a game, the user's name, and a list of items that have been answered during the current session. The user's rating, rating deviation, and the date the user last played a game are variables that are used by the Glicko rating system. The user's name is memorised so that the robot can address the user by the user's name. And by storing which items the user has during the current session, the robot can avoid asking the same item again within a session.

3.3 Rational agent

The decisions on when, what, and how to adapt, are made by a rational agent, which was programmed using GOAL. GOAL (Goal-Oriented Agent Language) is a programming language designed for programming rational agents, which derive their choice of action from their beliefs, knowledge and goals (Hindriks, 2009). Together, the beliefs, knowledge and goals form the mental state of the agent. Knowledge is static and will not change during runtime. An example of knowledge is that the robot knows which games can be played. Beliefs, on the other hand, are dynamic and may change during runtime. For example, the agent has the belief that the child is currently at 2350 rating for a game. A GOAL agent can have one or more goals, that each specify a state of the environment that the agent wants to achieve. To realise these goals, the GOAL agent selects actions based on action rules. An action rule consists of a reference to the corresponding action specification and a mental state condition that indicates when the action can be selected by the agent. The action is said to be applicable when the mental state of the agent matches the mental state condition of the action rule. For example, when the agent beliefs that the user is not motivated to play a certain game, the agent can consider taking the action of switching to a different game, which is an action that is only applicable when the user is no longer motivated. An action specification contains a precondition, postcondition, and the action. When the precondition is met, the action is said to be enabled. An example of a precondition for switching a game could be that the user should have played the game for at least five minutes, before a game switch can be considered by the agent. A GOAL agent will

only consider actions for execution that are both applicable and enabled. When more than one action is applicable and enabled, the agent randomly selects one of the actions.

A GOAL agent is connected with its environment via a perceptual interface. The interface specifies which percepts the agent receives from the environment. The percepts are handled by the event module, which uses the percepts to update the agent's mental state.

The main task of the GOAL agent is selecting the difficulty of the items. Psychometrically optimal selection of items means that items will have to be selected with a difficulty matching the user's ratings. When such a selection method is used, the rating system can make the most reliable estimations of the user's rating. However, if the item rating equals the user rating, it is estimated that the probability of the user answering the item correctly is 50%. Answering about 50% of the items correctly is experienced as discouraging, as the game will be too difficult. Therefore, instead of using psychometrically optimal selection, the GOAL agent selects items based on what percentage of the items the user answered correctly, so that, on average, a user will answer close to 70% of the items correctly. A success rate of 70% is generally considered to be optimal for facilitating intrinsic motivation. To influence the percentage of correct answers, the GOAL agent can select either an easy, moderate, or difficult item. An easy item is an item that on average will be answered correctly 70% of the time and is selected when the user answered less than 70% correct. A difficult item is selected when the user answered more than 80% of the items correctly, and is an item which will be answered correctly 30% of the time on average. An item of a moderate difficulty is an item that on average will be answered correctly 50% of the time on average, and is selected when the user answered between 70% and 80% of the items correctly.

Eggen and Verschoor (2006) showed that increasing the percentage of correct answers from 50% to 70% comes at a cost of measurement precision. Therefore, the user will have to answer more items before the user rating is a reliable estimate of the user's true rating.

The GOAL agent also keeps track of the user's performance and responds to exceptionally well and poor performance. The user's performance is defined as the discrepancy between the expected outcome and the actual outcome. We used a basic algorithm to calculate when the user is performing exceptionally well:

$$\prod_{i=1} \mathrm{E}(s|r_{user},\ r_{\mathrm{item}}, \mathrm{RD}_{\mathrm{item}})_i < 0.10$$

Where $\mathrm{E}(s \mid r_{\mathrm{user}}, r_{\mathrm{item}}, \mathrm{RD}_{\mathrm{item}})$ is the expected outcome given the estimated user rating, the difficulty of the item, and the rating deviation of the item. Each time the user correctly answered an item, the probability of a correct answer was stored, provided that the user's rating deviation was less than

125. The cumulative probability is calculated by multiplying the probabilities of answering each item in the sequence correct. The cumulative probability is reset when the user answered incorrectly. When the cumulative probability was smaller than 0.10, the GOAL agent responds by complimenting the user on doing well. The cumulative probability was set at 0.10 so that the compliment would likely be perceived as sincere, and that it was likely that this action rule would be triggered about once or twice during the experiment. Because the probabilities depend on the difficulty of an item, the user will have to answer fewer items when they are difficult items than when they are easy items, in order to receive a compliment. For example, the user only has to answers two difficult items correctly in a row, in order to be given a compliment, compared to seven easy items.

We used the same algorithm to estimate when the user was performing exceptionally poor, except that the cumulative probability had to be smaller than 0.05. The agent responds to exceptionally poor performance, by suggesting playing another game. We argue that when there is such a large discrepancy between the expected outcomes and the actual outcome, the user might not be motivated to play the game anymore, which can result in the user putting less effort into the game.

The GOAL agent also kept track of time during the experiment and tracked which games were played. The agent suggested switching to another game, when the user had played the game long enough and had just answered an item. Furthermore, the agent initiated the end-of-experiment dialogue, when the time for the experiment was up.

For this study, the use of a GOAL agent is useful for structuring the process of reasoning. However, it was not necessary to use a rational agent, nor did we fully exploit the functionalities offered by GOAL, like handling multiple conflicting goals. The main reason we used a GOAL agent is to make the user-adaptive system compatible with other systems developed for the ALIZ-E project. Because the complexity of reasoning will increase the more systems are integrated into the social robot, it makes sense to use rational agents to handle the social robot's reasoning.

3.4 The games

Three games have been designed for the ALIZ-E project, namely a quiz, a math game and an imitation game. The games are designed to be both fun and educational, embracing the concept of "learning by playing". For the experiment, we used the math game and the imitation game, which we will now discuss.

3.4.1 The math game

For children, it can sometimes be difficult to solve arithmetical problems encountered in real-life situations. For example, a child with diabetes needs to be able to calculate how many carbohydrates he or she has consumed since the last insulin injection, so that the required amount of insulin can be

estimated and injected. For the self-efficacy of the child, it is important that the child has sufficient skills in arithmetic to be able to calculate the required amount of insulin without the help of an elder or caretaker. With the math game, the children can become better at arithmetic by solving basic arithmetic assignments. The robot will select an arithmetic assignment and ask the child to solve it. The assignment will also be displayed on a monitor standing next to the robot. Once the child gives the answer, the robot will tell the child if the answer is correct or incorrect. The game is designed to allow the child to practice on arithmetic, rather than teaching the child how to solve the assignments.

The arithmetic assignments are selected from an item bank; a repository containing 310 unique arithmetic assignments. The complexity of the assignments ranges from the very basic (e.g. "1 + 1") to complex assignments (e.g. "11858 / 98"), which require multiple operations to solve. The item bank includes addition, subtraction, multiplication and division assignments. For a complete overview of all the arithmetic assignments, see Appendix B. The difficulty of each of the assignments was set using the levels of difficulty used in the study of Janssen and colleagues (2011). The levels of difficulty are based on two instruction books (Goffree & Oonk, 2004; Borghouts et al., 2005) and have been verified by an elementary school teacher. In total, there were 29 different levels of difficulty which have been converted to ratings, using the same order. For example, assignments of level 3 were converted to a rating of 300, and assignments of level 4 were converted to a rating of 400, etcetera. All the assignments were given an initial rating deviation of 150.

3.4.2 The imitation game

The imitation game is designed to have the children do physical exercise. In the game, the robot executes a sequence of arm movements, which the child then has to memorize. Once the robot has finished the sequence it is the child's turn to reproduce the sequence. If the child wants to get better at the game, he or she will need to find new strategies to memorize the sequences efficiently.

The initial ratings for the sequences are based on the length of the sequence and modified by the complexity of the movement(s), and the presence of similar subsequent movements. Every movement has to be memorized, and thus, the more movements have to be memorized, the more difficult the sequence will be. For each movement in the sequence, the rating increased by 300. Thus, a sequence of one movement has a rating of 300, a sequence of two movements has a rating of 600, etcetera. The sequences could contain eight different arm movements; left arm down ("BL"), right arm down ("BR"), left arm up ("TL"), right arm up ("TR"), both arms down ("BLBR"), both arms up ("TLTR"), left arm down and right arm up ("BLTR"), and left arm up and right arm down ("TLBR"). We assumed that some of these movements are easier to memorize than others, because the movements are not equal in the amount of information that needs to be stored in order to memorize

the movement. Thus, sequences containing the movements “BLBR” and/or “TLTR” had their rating increased by an additional 100 rating. These movements require two arms, rather than one, but the arms share the same direction. The sequences containing the movements “BLTR” and/or “TLBR” had their ratings increased by 200 rating, because the movements require two arms, rather than one, and do not share the same direction. All the sequences were given an initial rating deviation of 200. The complete item bank can be found in Appendix B.

Chapter 4

Methodology

4.1 Participants

The seventh grade class of a primary school (The 'Griftschool' in Woudenberg, the Netherlands) was selected to participate, and all 22 of the children in the class participated. Of the 22 participants, 14 were male and 8 were female. They were between 10 and 12 years old, with an average age of 10.59. A seventh grade class was selected because children of a seventh grade class are generally between 10 and 11 years old, which falls within the targeted age of the ALIZ-E project of 8 to 12 years old. In return for allowing the experiment to take place at the school, the school received toys (K'NEX) as a gift. Furthermore, the class received a lesson about robotics, which took place a month after the experiment. As a reward for participating, all participants received a picture of themselves and the robot.

4.2 Measures

Intrinsic motivation was measured using the free-choice method (Deci, 1971). This is a widely used method to study the intrinsic motivation of both adults and children (e.g. Vallerand, Gauvin & Halliwell, 1986). During the experiment, the participants are presented with a period in which they are free to choose what activity they want to do. To ensure that the participants choose an activity at their own volition, they are led to believe that the free-choice period is not part of the experiment. The participant is believed to be intrinsically motivated to engage in the chosen activity, because the participant selected the activity out of interest/enjoyment and was free to do so (Deci & Ryan, 1985). The time that the participant engages in a target activity during the free-choice period can then be used as a measure for intrinsic motivation. In our experiment, the free-choice period was formalized as follows. After playing the two games, the participants were led to believe that the experiment was over, but that they had to stay in the room while the experimenters checked the data. During that time, the participants were free to choose in which activity they wanted to engage. The available

activities were: continue playing with the robot (target activity), read one of the five comics (alternative activity) or play a popular game on the laptop (alternative activity). We measured the time the participant engaged in one or more of the activities, to a maximum of five minutes. The participants were free to switch from the initially chosen activity to another activity. However none of the participants switched to an alternative activity to the robot or vice versa; they either played 5 minutes with the robot or 5 minutes with one/two of the alternative activities. Thus, this variable is considered to be dichotomous rather than continuous.

After the free-choice period, the participant had to fill in a questionnaire (see Appendix D), with question about how they experienced interacting with the robot. The Self-Assessment Manikin (SAM; Bradley & Lang, 1994) was used to measure how the participant felt (happy/sad) and how tense they felt during the experiment. The SAM is a non-verbal pictorial rating system that can be used to obtain self-assessments of experienced emotions on the dimensions of arousal, pleasure, and dominance. It consists of three rows of manikins, which each measure one of the dimensions on a 5-points scale. Children readily identify themselves with the SAM figure, and they can easily understand what emotional dimensions it represents (Lang, 1985). The participants also had to rate how fun playing with the robot was, and how fun playing the math game and the imitation game was, on a scale from 1 (terrible) to 5 (amazing). For these questions, smileys were used to explain the scale in a graphical way. The smileys are used to measure fun, and can be easily interpreted by children (Read, MacFarlane & Casey, 2002). Finally, the participants had to indicate which game they enjoyed the most, and how difficult they considered the games.

All the sessions have been recorded on video. The recordings were made using a Sony Handycam (DCR-SR55E). The camera was mounted on a tripod for extra stability. Because the experimenters could not see the participant directly, the camera was used to see the participant. By connecting the video camera to a monitor, the experimenters could see the images of the video camera in real-time. Because the robot used in the experiment was not fully autonomous, the experimenters partly operated the robot. With the real-time video feed, the experimenters could both see and hear the participant, and make the robot respond to the participant as if it perceived what the participant did and said. The recordings are also used for post-hoc analysis and are used in other studies within ALIZ-E.

4.3 Materials

The NAO robot

The robotic platform used within the ALIZ-E project is Aldebaran's NAO (see Figure 9). NAO is a 57 cm tall humanoid robot, with 25 degrees of freedom. For this experiment, we used a NAO robot v32.

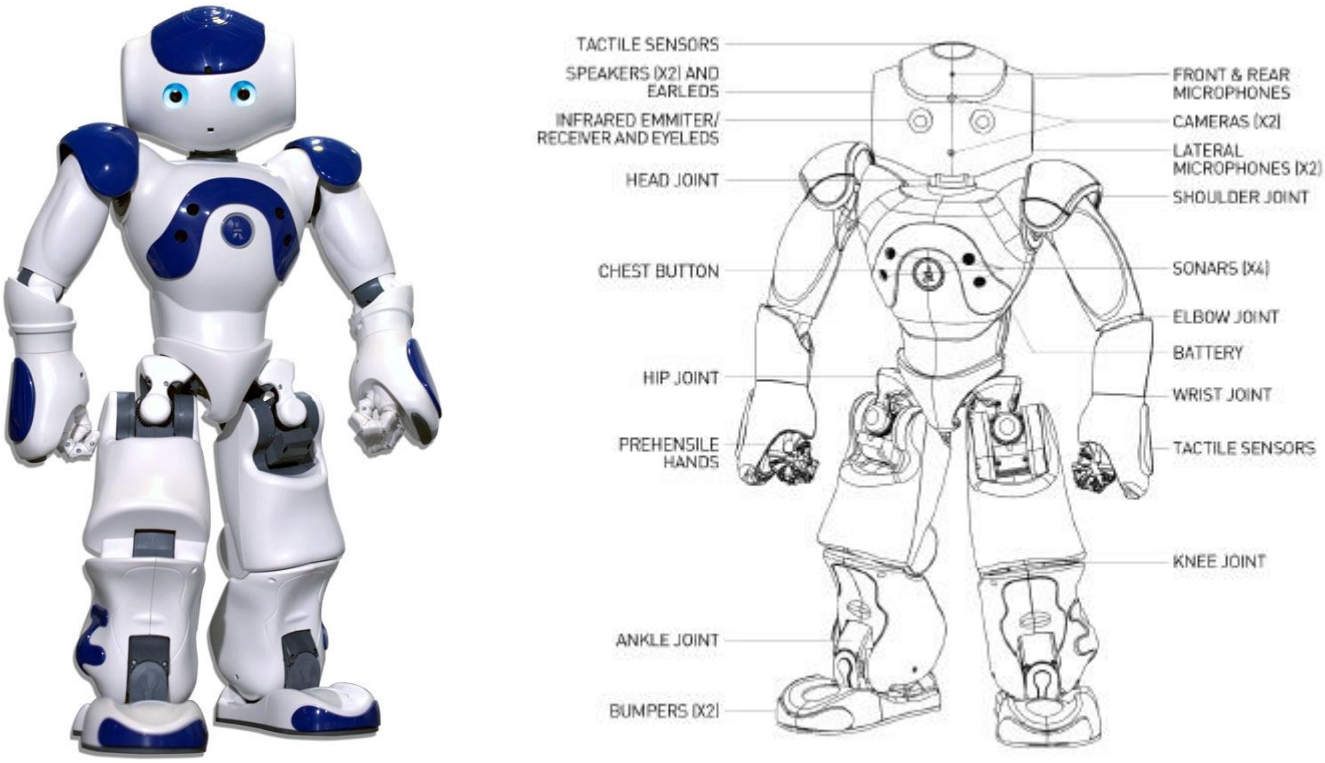


Figure 9. Aldebaran's NAO.

However, we started out with a different NAO (v28), but it malfunctioned after having tested four participants. For the remainder of the experiment the NAO v32 was used. The two NAOs differed from each other in colour and in the voice it used. The first NAO was coloured red, while the second was coloured blue. The two NAOs also used different text-to-speech (TTS) modules to convert text into speech. The TTS module allows the robot to talk. First, a text message is sent to the TTS module, where it is converted into speech. Then the produced speech can be pronounced using the robot's speakers. The first NAO used Acapela TTS[2] (v7.0, using the voice "Femke22Enhanced") to convert text to speech, and spoke with a mature woman's voice, while the second NAO had a child's voice and used the program Fluency TTS[3] speech editor professional (v4.0, using the voice "Fiona"). Other than these two differences, the two robots were identical. Both robots were provided with the name "Lola". The robot and the laptops communicated with each other using a wireless router.

Software

In this experiment we wanted to study the interaction between a child and an autonomous robot. However, for the robot to function autonomously, it needed to recognize and understand the child's speech and motions flawlessly, as any mistake in communication could disrupt the interaction

[2] http://www.acapela-group.com/index.html
[3] http://www.fluency.nl/international.htm

between the child and robot. Because such software was not available, we adopted the Wizard-of-Oz framework (Green, Hüttenrauch & Eklundh, 2004). The child was to believe that he or she interacted with an autonomous robot, while, in truth, the wizard (the experimenter) was partially controlling the robot. In this way, the interaction between a child and an autonomous robot was studied without the use of a fully autonomous robot.

For this experiment, we used the existing Wizard-of-Oz interface that was developed for the ALIZ-E project (see Figure 10). With this interface, we had full control of the experiment and over the robot's speech and motion. The experimenter processed the child's behaviour and selected the appropriate responses. This included judging the correctness of a given answer, by comparing the child's answer with the correct answer shown on the screen, and pressing the "correct" or "incorrect" button, and interpreting the child's speech. While the experimenter selected what response to give, the content of the responses were hardcoded, and in case there were more variations of one response, the robot would randomly select one of the variations. Once the appropriate response was selected, the robot would communicate the response to the child. With the Wizard-of-Oz interface we could also controlled the flow of the experiment, in case it needed to be altered due to an unexpected event (e.g. starting halfway the experiment when the robot crashed). However, normally the GOAL-agent would tell the experimenter what action to take, and

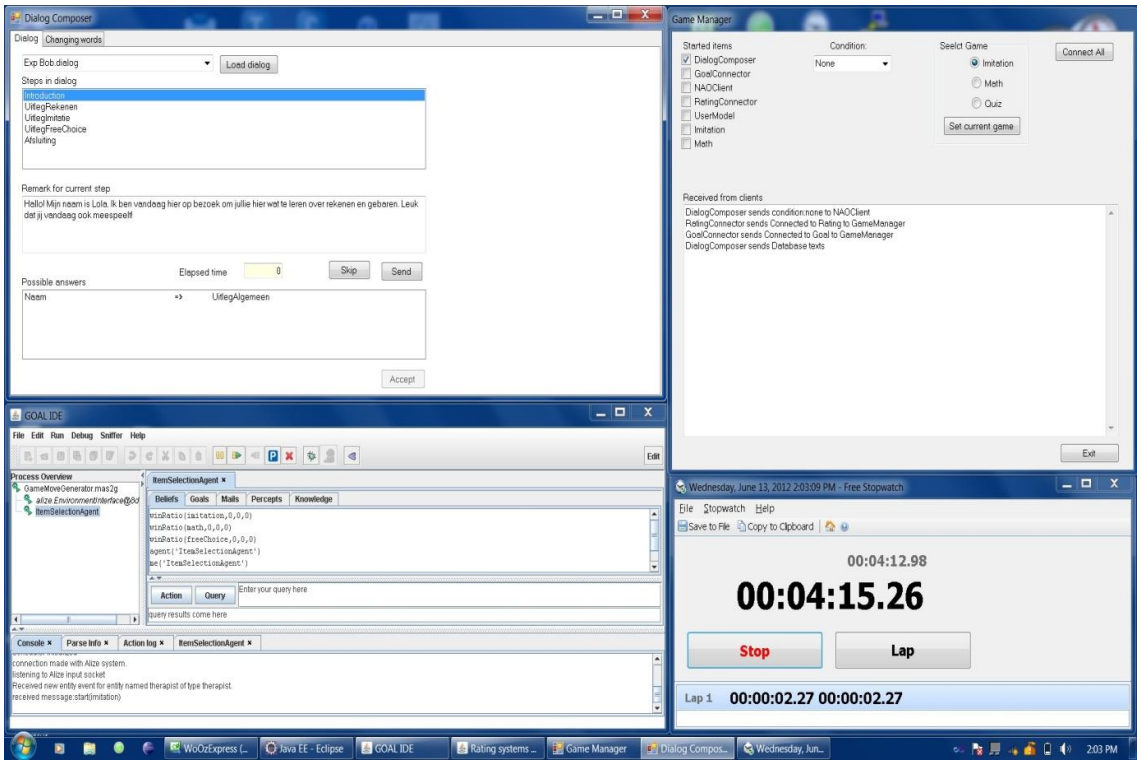


Figure 10. The interfaces used by the experimenter: the Wizard-of-Oz interface (top right), the dialog tool (top left), GOAL (bottom left) and the stopwatch (bottom right).

the experimenter would then select the action. Furthermore, the Wizard-of-Oz interface included a dialog tool, which could be used to create new sentences on the fly, and send those to the robot to pronounce them, in case the premade dialog options (see Appendix A) were insufficient. Furthermore, the interface for both games included standard dialogs which could be used to deal with common situations, such as when the experimenter could not hear what the child said. We used a virtual stopwatch to keep track of time during the free-choice period.

The Wizard of Oz program was connected to the robot, GOAL (v3959) and the rating system. The GOAL agent kept track of time during the experiment up to the free-choice period. We used a virtual stopwatch to keep track of time during the free-choice period and as a failsafe in case the GOAL agent crashed or had to be restarted.

Silent mouse

A mouse that made no sound when you click with it (a Nexus Silent Mouse SM-70008) was used by the experimenter who was controlling the Wizard of Oz, to ensure that the participants would not relate hearing a mouse click with the robot performing an action.

Alternative activities

For the free-choice period, the children could choose to continue playing with the robot, read a comic or play a game on the laptop. There were five comics from which the children could choose, namely two comics of "Donald Duck", one of "Dirk Jan", one of "Asterix" and one of "Kid Paddle". All the comics are popular and well known amongst 10 year olds. Two games could be played on the laptop, namely a free-to-play version of "Bejeweled"[4] and of "Bubbles"[5]. Both are well known and popular games.

For a detailed overview of all the used equipment, see the experiment protocol in Appendix C.

4.4 Experimental design

For the experiment, a between-subject design was used. The participants were randomly assigned to either the experimental group or the control group. For the experimental group, the percentage of correct answers was regulated to be close to 70%, by asking items with an easy, moderate or hard difficulty. Furthermore, the robot reacted when the performance of the participant was exceptionally poor or exceptionally well. The robot responded to the latter by giving the participant a compliment on his or her performance. When a participant's performance was exceptionally poor, the robot

[4] http://www.spellensite.nl/spellen-spelen.php?type=spellen&spellen=Bejeweled&id=1116
[5] http://www.funnygames.nl/spel/bubbels_3.html

would switch to another game. For the control group, the robot only asked items of a moderate difficulty and did not react to the performance of the participant. The order in which the participants played the two games was counterbalanced; half of the participants started with the math game, followed by the imitation game, and the other half of the participants started with the imitation game, followed by the math game.

The experiment consisted of two sessions. The first session lasted approximately 25 minutes. For the second sessions, we reduced the time the participants played a game from 7 minutes to 5 minutes and therefore the second session lasted approximately 20 minutes. Having two sessions, rather than one, served three purposes. First, by playing one of the games for seven minutes, a reliable estimate can be obtained regarding skill of the participant on that game. Also, the answers can be used to adjust the item ratings of the corresponding items, making them more reliable. This in turn will make the estimated likelihood on a correct answer more reliable, as its reliability is dependent on the reliability of the item's rating and the reliability of the participant's rating. The more reliable the rating of the participant and item, the more reliable the estimated likelihood will be and the more efficient the item selection will be. Second, earlier research (Janssen et al., 2011; Robben, 2011) indicates that children will find the robot to be very fun to play with, regardless of the functions of the robot. These studies found that in a second session, the novelty effect was reduced sufficiently as to prevent a ceiling effect. And third, the participants will learn how to play the game during the first session, so that they can play the games correctly right from the start during the second session.

4.5 Experimental Setup

The experiment took place at the Griftschool, in one of the offices. The experimenters had to be in the same room as the participant, because no other room was available. The office was rearranged for the experiment, so that the experimenters would attract as little attention as was possible. The experimental setup can be seen in Figure 11. The robot stood on the desk, at eye level. The participant would sit or stand in front of the robot and the experimenters would sit behind the participant. The protective foam of the case in which the robot was transported was used to block the line of sight between the participant and the experimenters. The reason to block the line of sight was to prevent the participant from seeing the experimenter control the robot. Furthermore, this would prevent the participant from feeling stared at, which could make the participant feel uneasy. We placed the camera away from the robot (the primary focus of the participant's gaze), to minimize the possible unease of being filmed.

Figure 11. The experimental setup.

4.6 Procedure

Prior to the experiment, the teacher of the participants assigned each of them to one of three groups (above average, average, and below average) depending on how skilled they are at math. These groups were used to set the initial rating for the math game. The participants with above average math skills started out at 2300 rating. Those with average math skill started at 2100 rating. And those that have below average math skills started at 1900 rating. For the imitation game, the participants all started out at 1500 rating. During the first session, the estimated likelihood on a correct answer may be off, because the accuracy of the participants' rating depends on how well the experimenter estimated the participants' skill level. For the second session, the difficulty of the items was manually updated, using the answers of the participants during the first session. While the rating system would normally update the difficulty of the items after each given answer, this option had to be turned off so that the rating system would be the same for each participant

Entering an office with two adult strangers can be scary for children. Therefore, the two experimenters introduced themselves and the robot to the class a week before the experiment. They told the class about the ALIZ-E project and about the experiment, what they could expect and what was expected of them. See 2.1 in Appendix C for all the things we mentioned during the introduction.

Prior to the experiment, one of the experimenters lead the participant to the office and explained the following:

1) The experiment is to test the robot, not the participant's skill.

2) The robot may sometimes show erratic movements, which they are to ignore.

3) The experimenters are there only in case the robot fails and the participant can act as if the experimenters are not in the room.

4) The participant can decide at any point to stop participating in the experiment.

5) That the experiment is over after playing the two games, but that they could freely choose an activity, while the experimenters checked the data.

The experiment started with making acquaintance with the robot. The robot introduced itself and asked the child a few questions, such as "what is your name". The answers were added by the experimenter to the user model. After the introduction, the robot explained how the first game is played. The participant then played the game for five minutes. When the five-minute mark was reached, no new items were asked, and the robot waited for the participant to answer the current item, before introducing the second game, which would also be played for five minutes. After the second game, the robot would announce that the participant could choose what activity to do next. The participant had five minutes to engage in the chosen activity. Following the free-choice period, the robot would ask if the participant liked playing with the robot and say goodbye. Finally, the participant had to fill in the questionnaire and could then return to class. A schematic overview of how the experiment was scheduled can be seen in Table 1.

Table 1

The timetable of the experiment

Time	Action
0:00 – 2:00	The robot makes acquaintance with the participant, and explains the experiment and the first game.
2:00 – 7:00	Play the first game for ~5 minutes.
7:00 – 8:00	The robot explains the second game.
8:00 – 13:00	Play the second game for ~5 minutes.
13:00 – 14:00	The robot explains the free-choice period.
14:00 – 19:00	Play the free choice period.
19:00 – 20:00	The robot explains this is the end of the experiment and says goodbye.
20:00 – 23:00	Ask the participant to fill in the questionnaire.
23:00 – 25:00	Thank the participant and make a picture (only in session 2) of the participant with the robot.

4.7 Analysis

For the analysis, we categorized the items to levels of difficulty, based on the item rating. For example, items with a rating of 1950 to 2049 were assigned to the level of difficulty of 2000, and those with a rating of 2050 to 2149 were assigned to the level of difficulty of 2100, and so forth. None of the participants switched from one activity to another during the free-choice period. Therefore, the time the participants spend playing with the robot is considered to be a dichotomous variable, rather than a continuous variable.

To evaluate the reliability of the user ratings, we stored the participants user rating after each instance. The reliability is assessed by analysing how the user ratings develop over time and how much the initial user rating differs from the user rating at the end of the session. To analyse how reliable the item ratings are, the following variables were stored: the rating of the item, the outcome of the item (correct/incorrect) and the response times. We compare what percentage of the items was answered correctly with what percentage was to be expected according estimations of the rating system, for each of the levels of difficulty. Furthermore, for the math game, the item's rating will be evaluated by comparing the response time to the different levels of difficulty.

The independent variable of the experiment is the percentage of correct answers on both games. For the control condition, the percentage should be approximately 50%. And for the experimental condition, the percentage should be approximately 70%. To check if the manipulation was successful, we calculated what percentage of the items was answered correctly, using the outcomes of the presented items.

To measure whether the participants in the experimental condition were more intrinsically motivated to play games with the social robot, we tested if there was a significant difference between the two conditions in the number of participants that continued playing with the robot during the free-choice period.

4.8 Pilot

A pilot experiment was conducted to check the procedure, the software, the hardware, the parameter settings, and whether the dialogs were understandable and sufficient. Two girls, aged 10 and 11, voluntarily participated in the pilot. Both participants lived and went to school in Soesterberg. The pilot took place at the TNO building located in Soesterberg, in an office similar to the office used in the experiment. One program was not ready at the time of the pilot, was the program 'natural behaviour', which makes the robot occasionally shift its weight and look around. The Wizard of Oz program still had to be attuned to the natural behaviour program, so that both programs could run without causing the robot to crash.

After playing with the robot for about 25 minutes, both participants started showing signs of being agitated (e.g. frequently shifting in the seat). Therefore, we reduced the time of the experiment from 30 minutes to 25 minutes, by reducing the time the participants had to play the games and the free-choice period from 7 minutes to 5 minutes. We could not find any oddities in the development of the ratings, plotted over the time of the pilot, or in the percentage of correct answers. So we decided not to make any changes to the formulae of the rating system and its parameters.

Chapter 5

Results

The results are discussed as follows. First, the performance of the rating system is analysed. Were the user ratings reliable estimates of the participants' true ratings? And were the item ratings indicative of the difficulty of the items? To answer these questions, we analysed how the user ratings developed when more items were answered. And we analyse whether the estimated probability on a correct answer was a reliable estimate of the actual probability.

Second, we analyse whether the rating system was able to provide the participant with the optimal challenge. To this end, we discuss what percentage of the items the participants answered correctly, and compare these percentages to simulated data.

Last, we discuss whether more participants were intrinsically motivated to play games with the social robot, when the robot utilised the user-adaptive system to provide the participants with the optimal challenge and praise the participants when they performed exceptionally well. Also, we analyse whether exceptionally poor performance could be detected and if there was a relation between exceptionally poor performance and a low motivation to play the current game with the social robot.

Of the 22 participants, 21 finished both the first and second session. One participant was not feeling well and quit during the second session.

Analysis of the performance of the rating system

5.1 Development of the user ratings on the math game

Theoretically, the user rating will become more reliable the more items are answered and, at some point, the user rating will have approximated the participant's true rating. The question is how many items have to be answered before the user rating is a reliable estimate of the participant's true rating. To answer this question, we analyse how the user ratings developed when the participants

answered more items. When the user rating is a reliable estimate of the participant's true rating, we expect that the user rating will become relatively stable and will fluctuate around a certain rating, which we assume to be the participant's true rating. At this point, the rating system can make reliable predictions on the outcome of an instance, given that the item ratings are also reliable estimates of the items' true difficulty. For many of the participants, such fluctuations of the user's rating can already be found in the first session (see Figure 12). For instance, the user rating of participant number 17 (the black dotted line) fluctuates around 2750 rating after the participant has answered 10 items. Not all user ratings have stabilised at the end of the first session. An example is the user rating of participant 6 (the red line), who started at a rating of 2100. For this participant, the user rating increased for each of the last eight items, which means that all these items have been answered correctly. This suggests that the user rating was not yet a reliable estimate of the participant's true rating, because we did not expect to see such large gains in user rating when the user rating approximates the participant's true rating.

On average, the user rating increased by 447 during the first session; the initial user ratings proved to be too low. The large increase in user rating does suggest that the item ratings are indicative of

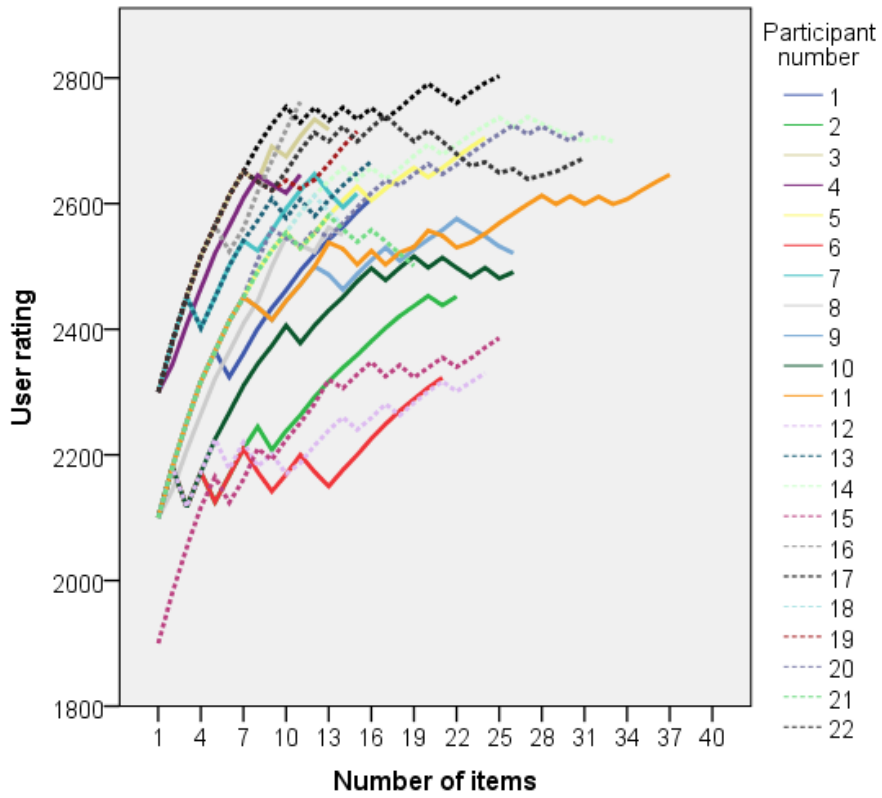


Figure 12. Each line represents the user rating on the math game of one participant during the first session. The figure shows how the user ratings developed as more items were answered by the participant.

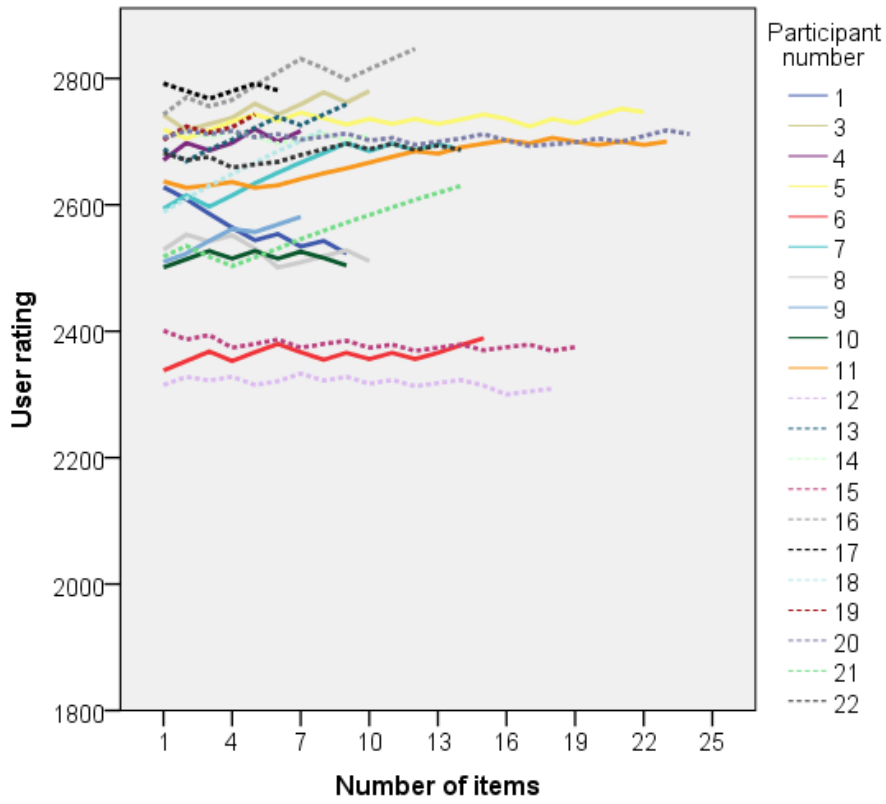


Figure 13. Each line represents the user rating on the math game of one participant during the second session. The figure shows how the user ratings developed as more items were answered by the participant.

the difficulty of the items, because such an increase in user rating should not occur when the item ratings have no predictive value on the difficulty of the items.

Figure 13 shows the user ratings on the math game during the second session. For about half of the participants (i.e. participant number 6 or 19) the user rating fluctuated from the beginning of the session. This suggests that for those participants the user rating was a reliable estimate of the participant's true rating. For the other half of the participants, the user rating still shows large increases or decreases. However, the average change in user rating was small. On average, the user rating changed by 29 during the second session. We found no significant difference between the control condition and the experimental condition in the average change in rating ($t(1) = 1.51$, $p = 0.132$). This means that the user ratings were not more stable in one condition than the other. Thus, controlling what percentage of the items the participant answered correctly did not appear to have affected the stability of the rating.

5.2 Evaluation of the difficulty of the math items

The initial item ratings for the math game were set based on two instruction books. We assume that the item ratings are precise enough to support an ordinal level, and to some extend an interval level. The rating system operates on an interval scale, which means that the difference in item rating between two items not only imply which is more difficult, but also how much more difficult the item is than the other. To evaluate the reliability of the item ratings, we compared the response times of the participants to the level of difficulty. Also, we analysed how reliable the estimated probabilities of a correct answer were. The reliability of the estimated probability depends on the reliability of the user rating and of the item rating. However, because the user ratings appear to have been reliable estimates of a participant's true rating during the second session, we attributed the difference between the estimated percentage of correct answers and the actual percentage for a large part to the reliability of the item ratings.

In order to compare the response times to the level of difficulty, we have to make a couple assumptions. We assumed that the more difficult the item, the longer it will take the participant to answer. Furthermore, we assumed that the time it takes the participant to answer follows a logistic curve (S-curve). We base this assumption on the relation between the difference between the user and item rating and the probability (see Figure 6), which also follows a logistic curve, and on the assumption that there is a maximum amount of time a participant is willing to spend before giving up. However, because the participants will not be asked to answer items for which they have an estimated probability of less than 30% to answer correctly, there will be little to no data on the upper tail of the logistic curve. Each participant has a different true skill, meaning that the logistic curve of each participant lies on a different point on the rating scale. For some participants, answering a math item of a difficulty of 2300 is easy and will take 20% of the maximum time the user is willing to spend on one item, while for others it is a difficult item and will take 90% of the maximum time to answer. Given that the participants are all from one class and roughly the same age, the assumption was that the skill of the participants at math follows a normal distribution. When this assumption was tested, we found that it did not hold, as the user ratings on the math game are slightly skewed to the left ($W(12) = 0.9$, $p = 0.035$). This could have resulted in items of a difficulty of less than 2600 being answered slightly faster than would be expected. Nonetheless, it is still expected that response times follow a logistic curve, given the difficulty of the items. And, again, because there is little to no data on items for which the participant has a probability of less than 30% of answering correctly, the upper tail of the logistic curve will not be visible in the data. As a result, we will not be able to discriminate between an exponential curve and a logistic curve. If any of the levels of difficulty

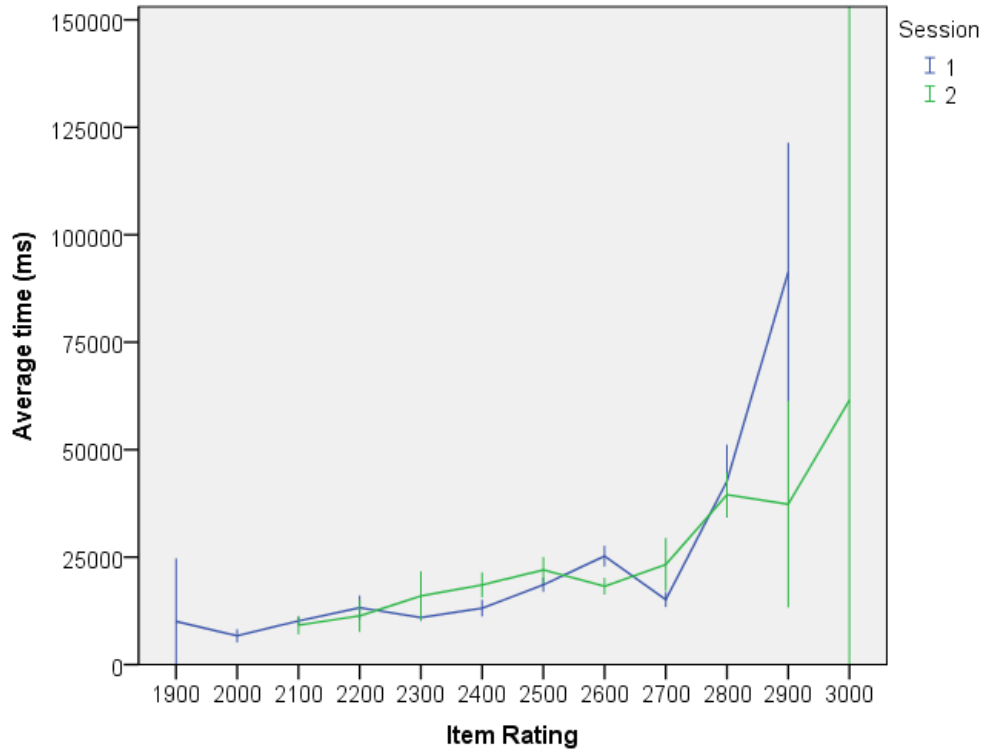


Figure 14. The average response time (in milliseconds) for answering a math item, for math items with an item rating of 1900 to 3000. The bars show the 95% confidence interval.

deviate from the expected logistic curve, then that can be an indication that the items are more or less difficult than their rating suggests.

Figure 14 shows the average response times for a math item of a certain level of difficulty. As expected, the response time increases exponentially with the difficulty of the items. There is one exception, namely items with that had an initial item rating of 2700. In the first session, the response times were lower than expected. An example of an item with a rating of 2700 is “how much is 49 times 12”. In the second session, the update of the item ratings corrected the ratings of these items by adjusting the ratings downwards to an item rating of approximately 2600. As a result, the lower response times shifted from occurring at 2700 to occur at 2600 during the second session and had become less substantial.

In Figure 15, the estimated and actual percentages of correct answers per level of difficulty are shown. The discrepancy for items with a difficulty of 2600 to 2700 is almost three times as large as the discrepancy for other levels of difficulty. Given that the response time was also lower than expected, the items appeared to have been easier than was expected. Items with a rating of 2100 to 2200 were also easier to answer correctly than was expected. An example of an item with a rating of 2200 is “how much is 80 times 25”. Only one item of the level of difficulty of 3000 has been answered, which explains the large discrepancy for this level of difficulty. The relatively large

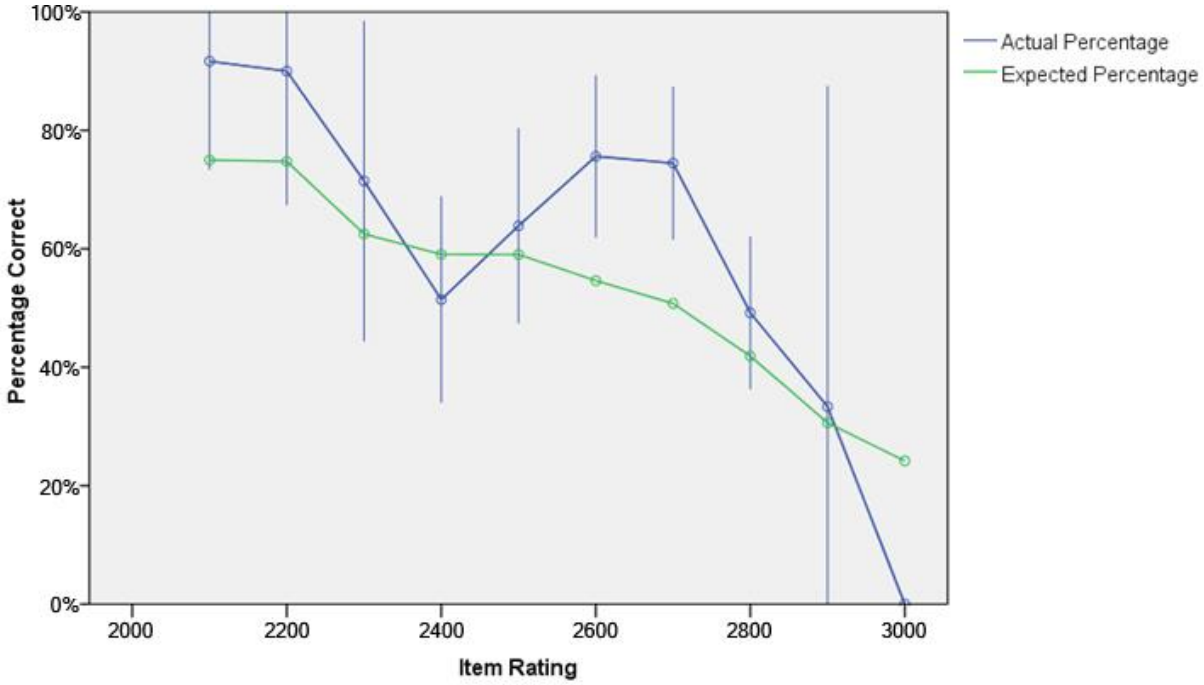


Figure 15. The estimated (the green line) and actual percentages (the blue line) of correct answers per level of difficulty, for the math game during session 2. The vertical bars represent the 95% confidence interval.

discrepancies between the expected and actual percentage of correct answers will be discussed when we discuss what percentage of the items were answered correctly by the participants (section 5.5).

The item ratings were adjusted in between the first and second session, based on the answers given by the participants. For the math game, 462 items were answered during the first session, divided amongst 78 different items. On average, each of the 78 items was answered 5.9 times, with a standard deviation of 4.9. While adjusting the item ratings should have made the item ratings more reliable estimates of the item's true difficulty, too few items have been answered to assume that the adjusted item ratings were reliable estimated during the second session.

Overall, the item ratings of the math items appear to be accurate enough to support an ordinal level. Based on the response times and the discrepancy between the estimated percentage of correct answers and the actual percentage, it appears that items with an initial item rating of 2700 were easier than expected. And possibly also items with an item rating of 2600.

5.3 Development of the user ratings on the imitation game

We analysed the user ratings on the imitation game in the same way as we analysed the user ratings on the math game. Answering an imitation item takes more time than answering a math item. Therefore, fewer items have been answered for the imitation game, than for the math game. As a

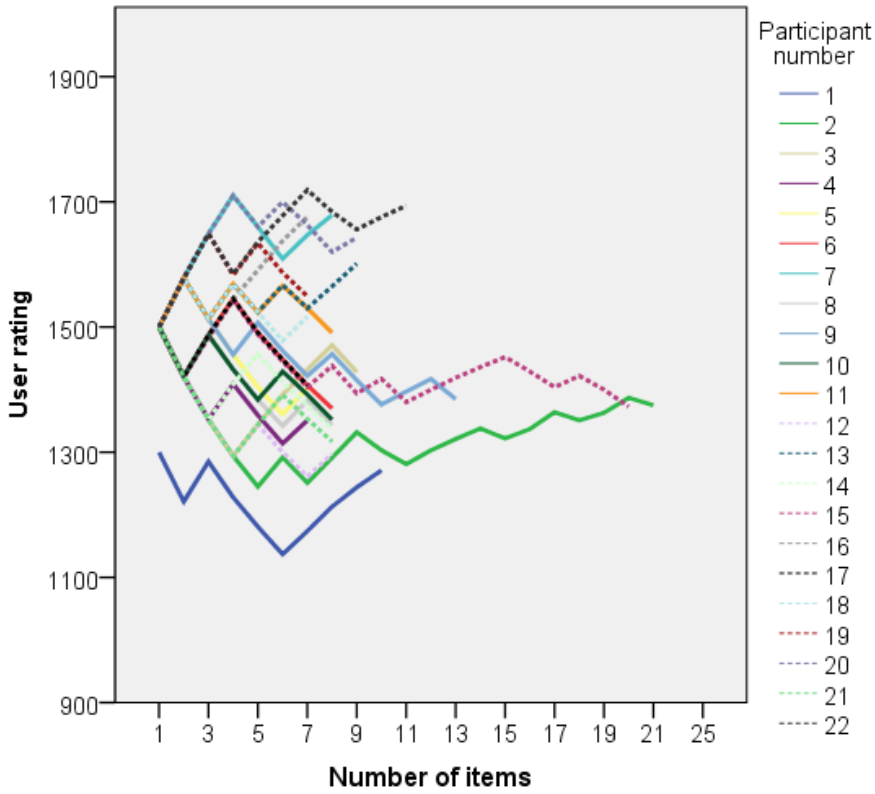


Figure 16. Each line represents the user rating on the imitation game of one participant during the first session. The figure shows how the user ratings developed as more items were answered by the participant.

result, we expect that the user ratings are less stable and only fluctuate around a certain rating for a few participants. In other words, we expect that the user ratings are less reliable estimates of the user's true rating imitation game.

The development of the user ratings on the imitation game can be seen in Figure 17. Too few items were answered during the first session to state whether or not the user ratings fluctuate around a certain rating. There are two participants that answered around twenty items, namely participant 2 and 15. The user rating of participant 2 steadily increased after the initial drop, rather than to fluctuate around a certain rating. For participant 15, the user rating did fluctuate (around 1400) after the initial drop in rating. On average, the user rating changed by 112 during the second session. Given that the user started with an initial rating deviation of 350, this is considered a relatively small difference.

Figure 17 shows the development of the user ratings of the imitation game for the second session. For many participants, the user ratings fluctuated around a certain rating. On average the user rating changed by 37 during the second session. However, the participant's rating deviations were still relatively large because the participants answered only a few items during the first session.

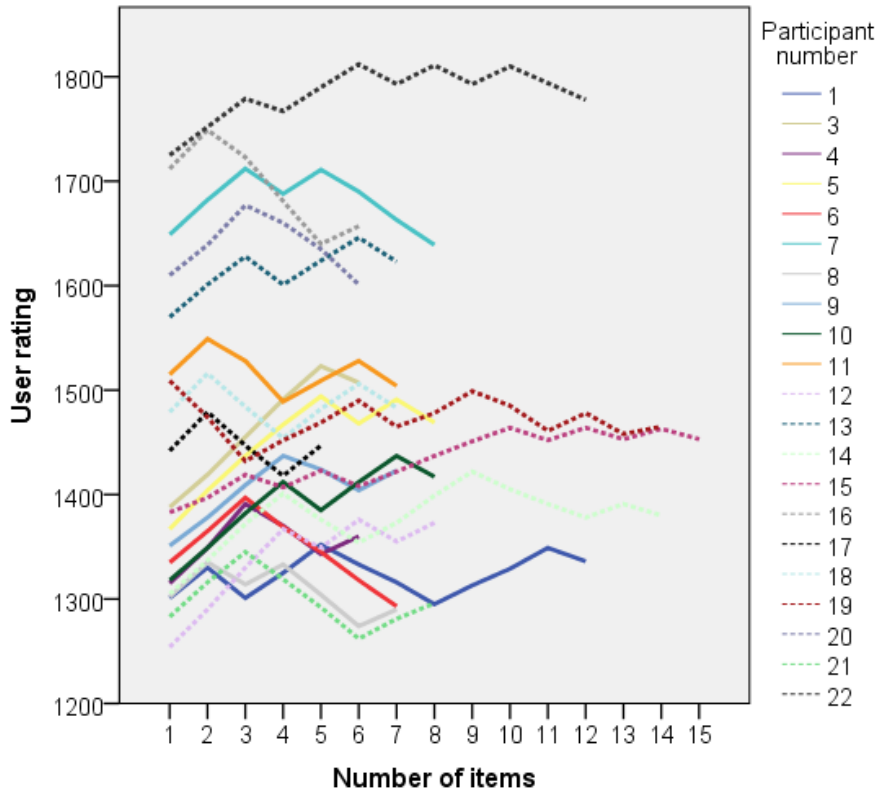


Figure 17. Each line represents the user rating on the imitation game of one participant during the second session. The figure shows how the user ratings developed as more items were answered by the participant.

Therefore, relatively large changes in rating occurred during the second session. While it appears that the user ratings fluctuated around a certain rating, the participant's rating deviation was too large to speak of a reliable estimate of the participant's true rating.

5.4 Evaluation of the difficulty of the imitation items

For the evaluation of the item ratings of the imitation items, we analysed the difference between the estimated percentage of correct answers and the actual percentage of correct answers. The response time cannot be used to evaluate the initial item ratings, as more difficult sequences generally contained more movements and thus took longer to complete.

Figure 18 shows the expected and actual percentage of correct answers for the first session. Items with an item rating of around 1300 to 1700 were answered correctly approximately 40% of the time, which is lower than expected. Items with a difficulty of approximately 1200 differed from this pattern, as these items were answered correctly for 81%.

Figure 19 shows the expected and actual percentage of correct answers for the second session. The confidence intervals are large, due to the small number of items that were answered. Two levels of

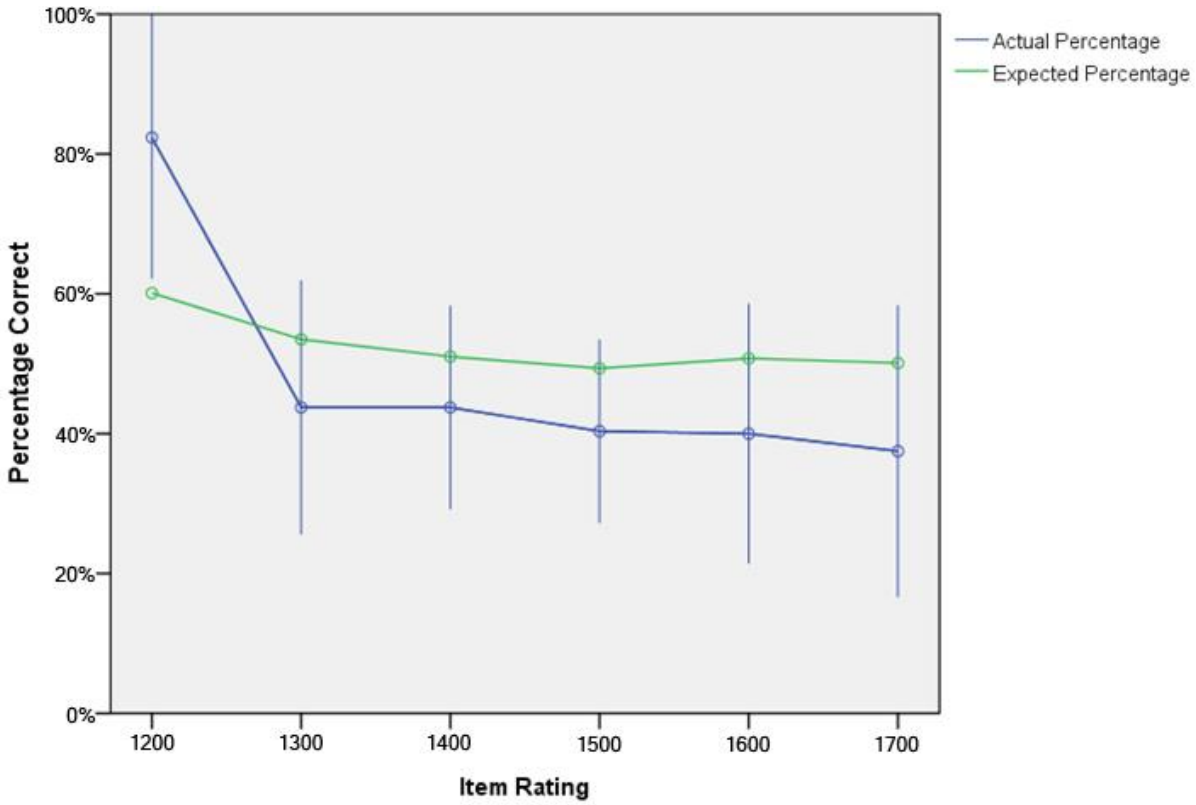


Figure 18. The percentage of correct imitations for different levels of difficulty during the first session. The vertical bars represent the 95% confidence interval.

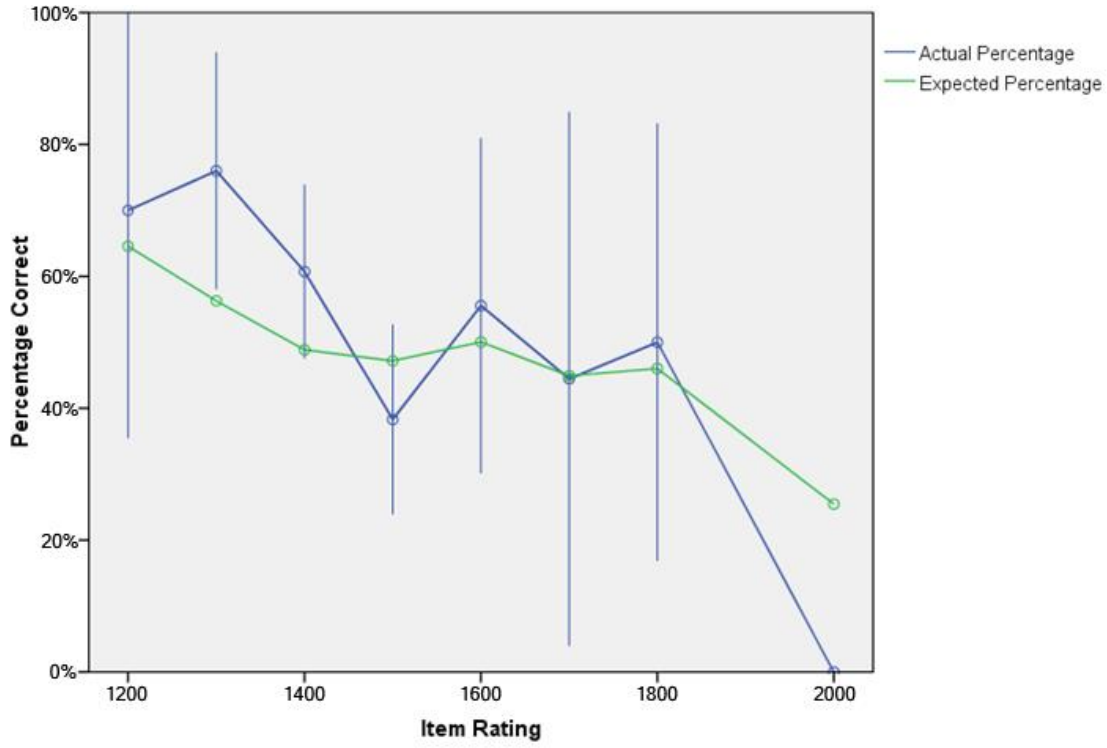


Figure 19. The percentage of correct imitations for different levels of difficulty during the second session. The vertical bars represent the 95% confidence interval.

difficulty deviate from what is expected, namely items with an item rating of 1300 and 1400. These items had a higher percentage correct than is expected. This suggests that these items were easier than the item rating suggested. Items with an item rating of 1300 to 1400 generally contain four movements, including one movement that requires both hands. Compared to the first session, the difference between the actual and expected percentages of correct answers has become smaller after the adjustment of the item ratings.

The item ratings for the imitation game were adjusted in between the first and second session. 208 imitation items were answered during the first session, divided amongst 37 different items. On average, each of the 37 items was answered 5.6 times, with a standard deviation of 4.2. Like with the math game, too few items have been answered by the participants to assume that the adjusted item ratings are reliable estimates of the item's true rating.

For the math game, the user ratings showed a large increase at the beginning of the first session. But, unlike for the math game, the user rating for the imitation did not show such a pattern during the first session. However, it is likely that the item ratings of the imitation game are related to the difficulty of that item, because there is a strong correlation between the ratings on the games. For the second session, there is a correlation of 0.54 ($n = 21$, $p = 0.012$) between the median rating on the math game and the median rating on the imitation game. Participants with a high user rating on the math game also had a high user rating on the imitation game. And, because the item ratings of the math items are related to the difficulty of those items, it is likely that the item ratings of the imitation items are also indicative of their difficulty.

Analysis of the use of the rating system

5.5 Manipulation check

For the experimental condition, items were selected on the basis of the current percentage of correct answers. With a set of selection rules, the GOAL agent attempted to keep the percentage of correct answers around 70%. Without these rules, the rating system has a natural tendency to select items for which the estimated probability of a correct answer is approximately 50%. Therefore, we expected that the participants in the control group answered 50% of the items correctly on average, and the participants in the experimental group answered close to 70% of the items correctly on average.

To get a better idea of what percentage of correct answers we can expect, given the size of the experiment and without the intervention of the GOAL agent, we constructed a simulation of the experiment. In the simulation each item had a 50% chance of being answered correctly. The simulation simulated data for the same number of participants as were in the control condition of

Table 2

Percentage of Correct Answers for the Simulated Data

			95% CI	
	M	*SD*	*LL*	*UL*
Mean	49.96%	4.94%	40.25%	59.62%
Standard Deviation	15.28%	3.60%	8.61%	22.62%

Note. CI = confidence interval; *LL* = lower limit; *UL* = upper limit.

the experiment and the same number of items answered by the participants. The simulated data can be seen in Table 2. Provided that the rating system is able to select items that the user will answer correctly 50% of the time, it can be expected that the percentage of correct answers in the control condition lies between 40 % and 60%.

In the experiment, the participants in the control condition answered on average 64% of the math items correctly and 52% of the imitation items, during the second session (see Table 3). For the math game, this percentage is not what is expected if the rating system was able to select items that would be answered correctly 50% of the time. It is likely that the estimates of the rating system were not accurate enough. Thus, while the rating system estimated that the probability of the participant answering the item correctly was close to 50%, the actual probability may have been much higher. While the rating system estimated that the participant had a 50% chance, they actually had a higher chance to answer correctly.

For the experimental condition, we expected the participants to answer close to 70% of the items correctly. In the experiment, the participants answered 64% of the math items correctly and 53% of the imitation items correctly. Especially for the imitation game, the GOAL agent was unable to select the difficulty of the items so that the participants would answer approximately 70% of the items correctly. It could be that the selection rules are insufficient, or that we used the wrong parameters for the selection rules. However, it is more likely that the rating system was also unable to select items for which the user had a probability of 30% or 70% to answer that item correctly.

There is no difference between the two groups on the math game ($F(1,19) = 0.063$, $p = 0.805$), nor on the imitation game ($F(1,19) = 0.12$, $p = 0.912$). Given that there are no differences between the

Table 3

Percentage of Correct Answers for Both Conditions

	Session 1				Session 2			
	Control		Experimental		Control		Experimental	
Game	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Math	77%	9%	72%	7%	64%	20%	65%	15%
Imitation	40%	15%	47%	19%	52%	13%	53%	12%

control group and experimental group in the percentage of correct answers, the manipulation did not appear to have worked.

In short, the rating system and item selection rules were not able to select items so that the participants in the control group would answer 50% of the items correctly, while the participants in the experimental group would answer 70% of the items correctly. It is unlikely that the rating system was able to reliably estimate the probability of a correct answer. Given the analysis of the user and item ratings, it is more likely that the item ratings, and to a lesser extend the user rating, were not reliable enough to support reliable estimations. Therefore, the GOAL agent was unable to increase the percentage of correct answers from 50% to 70%.

5.6 The free-choice period

Of the 21 participants in the second session, 9 chose to continue playing with the robot during the free-choice period. Of the 9 participants that chose to continue playing with the robot during the free-choice period, 5 belonged to the experimental group and 4 to the control group (see Figure 20). The participants in the experimental condition did not chose to continue playing with the robot during the free-choice period more often than the control condition ($F(1,19) = 0.058$, $p = 0.813$). Considering that the manipulation of the percentage of correct answers did not work, no differences between the two groups are expected. However, there were two participants in the experimental condition that answered more than 65% of the items correctly for both games. One of them chose to

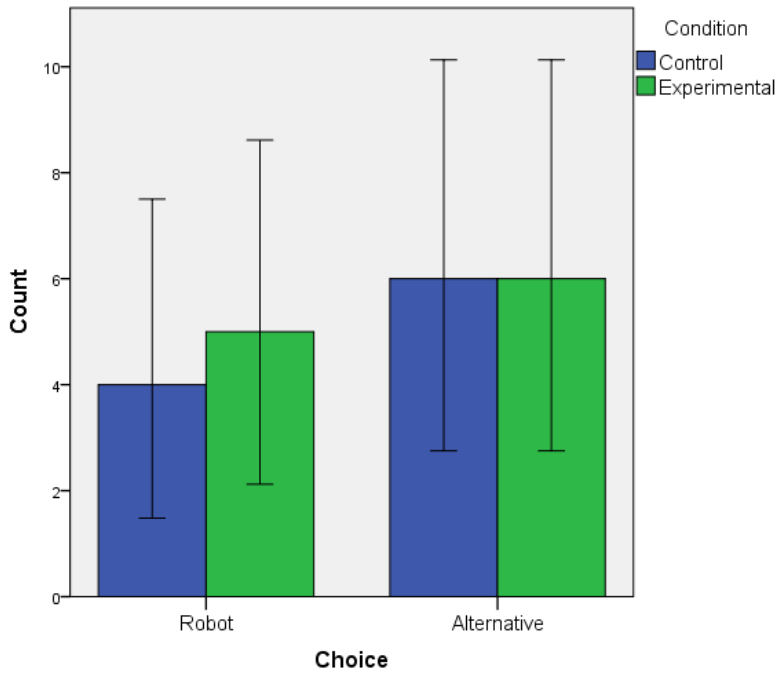


Figure 20. The choices the participants made during the free-choice period for both groups.

continue playing with the robot during the first session, but neither continued playing with the robot during the second session.

As a subjective measure of how the participants liked the robot, they were asked to rate the robot on a scale from 1 (terrible) to 5 (amazing). On average the participants in the control condition rated the robot with a 4.30, and the participants in the experimental group rated the robot with a 4.45, which is not significant ($F(1,19) = 0.269$, $p = 0.610$).

Of the 9 participants who continued playing with the robot, 5 participants continued playing the imitation game, and 4 continued with the math game. There was no effect of the order in which they played the games ($F(1,7) = 0.071$, $p = 0.798$). The math game was considered to be the most fun of the two, rated 3.81 on average. The participants rated the imitation game 3.76 on average.

The participants were also asked to rate the difficulty of the games. We expected that the participants in the control condition would rate the games to be more difficult than the participants in the experimental condition. However, since the manipulation of the percentage of correct answers did not work, no different should be expected. As can be seen in Figure 21, no difference between the two conditions was found in how difficult they perceived the games.

5.7 Detection of exceptional performance

The estimates of the rating system were used to estimate the participant's performance. When the robot had the belief that the participant was performing exceptionally well, it would complement the

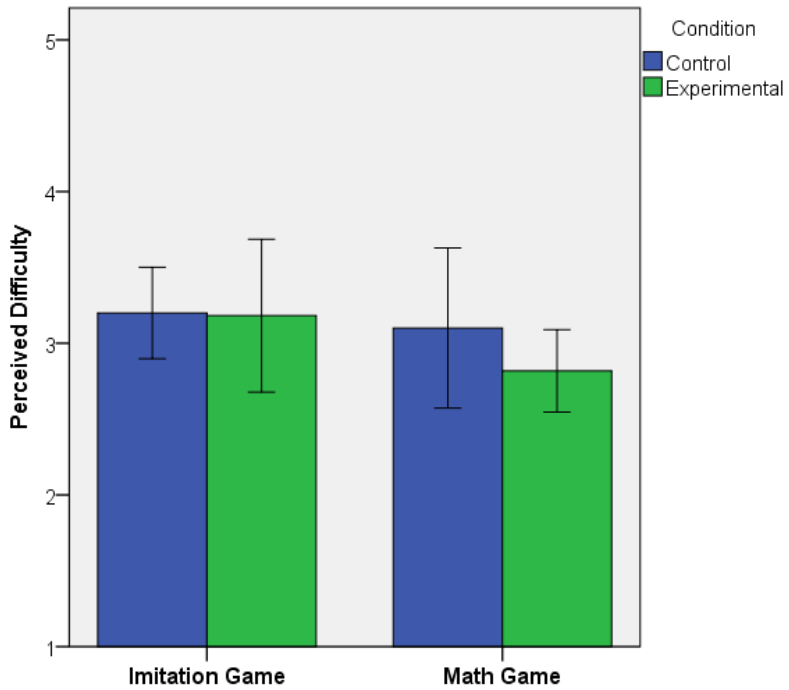


Figure 21. The perceived difficulty of the two games for both conditions.

user on doing so well. On average, the robot gave 0.90 compliments per participant. Especially the participants that continued playing with the robot during the free-choice period received compliments, because answering more items increases the likelihood that the participant performs exceptionally well at some point.

When the robot had the belief that the performance of the participant was exceptionally poor, the robot would assume that the participant was no longer motivated to play the current game. Such an event took place once during the experiment, and took place during a free-choice period in the first session. At the beginning of the free-choice period, the participant was asked whether he wanted to continue playing games with the robot, or do one of the alternative activities. The participant chose to continue playing the math game. The subsequent five items were then answered incorrectly (See Figure 22, the 22th item to the 27th item). All five questions had a difficulty equal to the participants' level of skill, meaning that the participant was expected to answer 50% of the items correctly. The probability of answering all five questions incorrectly in a row was estimated to be 2%. During the second session (the items 28 to 33), the rating of the participant on the math game went up again, up to the rating prior to the drop in rating. Given that the participant had a choice to do something different rather than playing with the robot, it is unlikely that the participant had lost the motivation to play with the robot. Furthermore, we used the video recordings to review the session. The participant did not appear to be unmotivated and tried to answer the item correctly. This is reflected

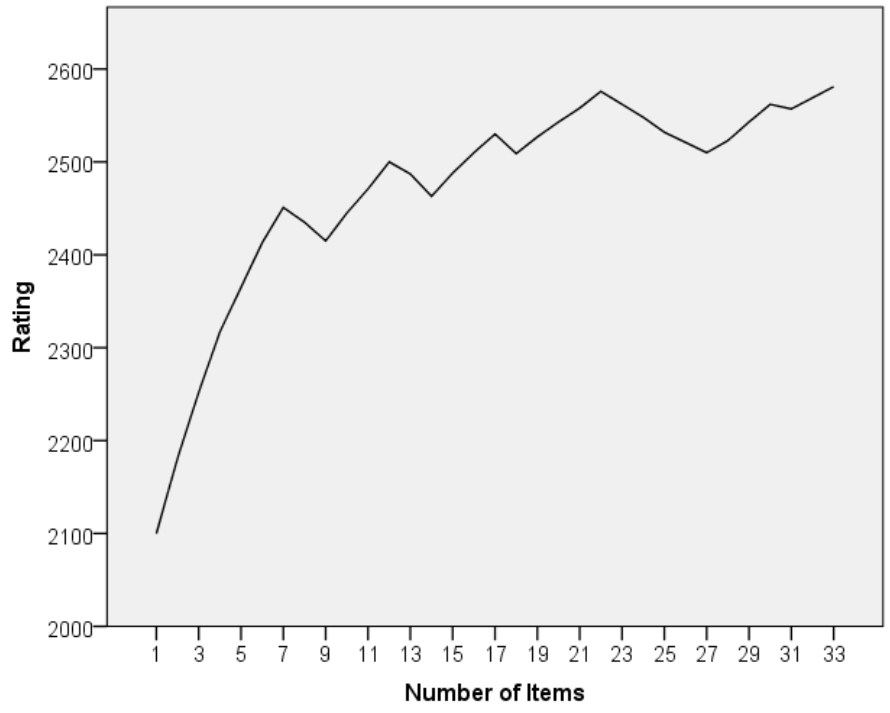


Figure 22. The drop in rating, starting at the 22th item and ending at the 27th item.

in the participants response time for these five items, as it took the participant 28.6 seconds on average (sd = 9.8s) to answer. However, during the second session, the participant took 41.5 (sd = 31.4s) seconds to answer items of the same difficulty. Rather than explaining the sudden drop in rating by a drop in motivation, this case may be better explained by a loss in attention.

Another way of using a rating system to predict the motivation of the user is by taking the response time into account. When the user takes adequate time to answer the item, and answers incorrectly, than the user tried to answer the item to the participant's best ability. However, when the user answer the item incorrectly and takes little time doing so, given the difficulty of the item, then that might be an indication that the user is no longer motivated to play the game. Figure 24 shows the response time for math items that were answered incorrectly, where all the dots represent an instance. The central black line is the average time it took the participants to answer and the top and bottom black lines represent the 95% confidence interval. In this example, we use the confidence interval to discriminate between instances for which the participants did try to answer, but failed, and instances for which the participant did not try hard enough to get to the correct answer. This simple method results in three instances for which the participant took little time to answer. We used the video recordings to analyse these three cases. In one case, the participant looked at the assignment ("what is 95 times 84") and gave up without trying. The other two cases, both

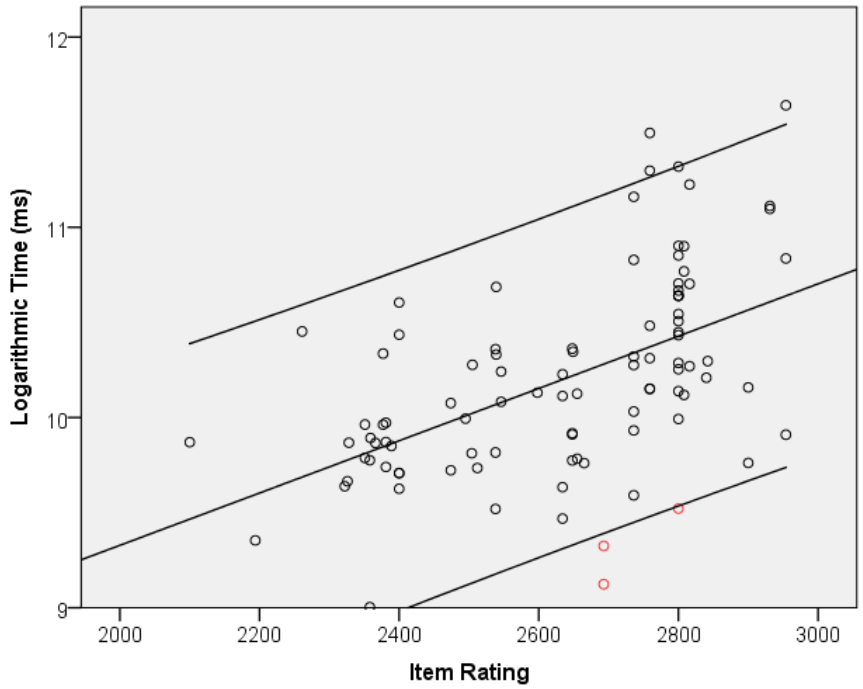


Figure 23. The logarithmic time for incorrect answers of different difficulties. The middle line is the fit line, and the lower and upper bar represent the individual 95% confidence interval. The red dots are cases for which the participant took very little time to answer, given the difficulty of the item.

participants had to answer the assignment “what is 21 times 60” and answered it with “180”. Given that both participants gave the same incorrect answer, they may have applied the wrong operations to solve the assignment, and gave the first answer that came to mind. For both these participants, the assignment was the first math item they had to answer. Given that they took little time to calculate the answer, we interpret these two cases as a lack of effort. The method we used to discriminate is not optimal, as the response time is related to the difficulty of the item, in respect to the user’s skill. In Figure 23, we have not taken the individual user ratings into account. The figure is only meant to illustrate how response times could be used to detect low motivation.

5.8 The subjective experience of the participants

During the lesson about robots, which took place a month after the experiment, the participants were asked to write down two things they liked about the robot and two things they did not like. Most of the participants reported that the robot and the games were (very) fun. Also, participants liked it that the robot memorized their names and used the names to address the participants. The participants reported that the NAO was kind to them, and that it didn’t mind wrong answers. Some participants found the robot to be funny, because of its robotic characteristics.

The participants were divided in their opinions on the voice of the robot. Some participants mentioned that they did not like the NAO’s voice. They said the voice sounded strange to them. Some others described the voice as ‘dull’, because the voice does not convey any emotion. But other participants liked the voice and described it as a ‘funny’ voice.

There were three things that the participants did not like about the robot. First, the robot was sometimes slow to respond. There are several reasons why the robot could not always respond quickly. If the robot had to pronounce a new message, while it was not yet finished pronouncing the previous message, then the robot would crash. Therefore, the experimenters had to wait for the robot to be ready to receive a new message to pronounce, which was usually right after the robot was finished pronouncing a message. Sometimes however, there was a delay of one to two seconds after the robot was finished talking, before the robot was ready to receive a new message to pronounce. This sometimes caused the robot to be slow to respond. Second, some participants mentioned that the squeaking noises, made by one of the joints of the robot, distracted them. And lastly, there was a bug that the participants found annoying. Sometimes during the imitation game, when the robot was performing a set of movements, the robot suddenly made erratic movements, which startled the participants. The experimenters warned them beforehand that this could happen and that should try to ignore it.

Chapter 6

Discussion

The field of social robotics aims at developing intelligent robots that can communicate and interact with people in a personal way, and to understand and relate to people (Breazeal, 2005). The idea of social robots capable of assisting us in our daily lives is becoming more real every day. However, for the long-term acceptance and utility of social robots, they have to be able to perceive, understand and respond to human social behaviour. To date, the efficacy of social robots is constraint by their inability to sustain social interactions beyond the scale of minutes. Therefore, human-robot interaction is usually limited to short-term interaction, as users typically stop interacting with a social robot once the initial novelty has worn off. This raises the question of how can we improve the social intelligence of a social robot, so that a person is motivated to interact with a social robot for a longer period of time.

In this study, we addressed one condition that can enhance a child's motivation to interact with a social robot, namely by providing the child with a feeling of competence when playing games with the robot. The robot adapts the difficulty of the games based on how skilled the child is, so that each child may play the games at a level of difficulty that is optimally challenging. The child's skill was estimated with the use of a Bayesian rating system. Moreover, the robot could recognise when the child was performing exceptionally well and give a compliment to the child on the performance. The robot could also recognise when the child was performing exceptionally poor, given how skilled the child was.

In this chapter, we will discuss the research questions, the performance of the rating system, suggest improvements to the user modelling system, and draw conclusions from this study.

6.1 Research questions

The first research question was "to what extend will children be intrinsically motivated to play games with a social robot, when the games are optimally challenging and exceptional performance is

praised by the social robot?". After analysing the results, we have to conclude that we cannot answer this research question. The rating system proved to be unable to control the percentage of correct answers to ensure that the participants in the control group would answer 50% of the items correctly and the participants in the experimental group answer 70% of the items correctly. The reason for this was that the item ratings, and to a lesser extend the user ratings, were not reliable enough.

In between the first and second session we adjusted the item ratings based on the answers that were given, so that we would have more reliable item ratings for the second session. The items that were answered during the first session were generally answered by less than five participants. Therefore, the reliability of the item ratings largely depended on the initial ratings. Moreover, many items that received little to no adjustment of its rating were selected during the second session, reducing the effectiveness of the adjustment of the item ratings.

Based on the information about the difficulty of the items available to us, we were able to distinguish 29 levels of difficulty and set the initial item ratings accordingly. Each level of difficulty contained 10 different items, with the exception of items with a rating of 2000, of which there were 30. Each item within a level of difficulty had the same rating, and thus, the same probability of being answered correctly. Because no distinction could be made between items of a certain level of difficulty, a level of difficulty was selected, rather than a specific item. By default, the first item of a level of difficulty would be selected, followed by the second item and so on. This resulted in the first few items of a level of difficulty being answered more often, than the items at the end of a level of difficulty, which were answered a few times to none. This is turn led to a greater accuracy of the first few items of a level of difficulty after the update of the item ratings, while the items at the end of a level of difficulty received a small or no update of their rating.

During the second session, the user ratings were relatively stable. As a result, the participants answered most, or all, of the items of a certain difficulty, including the items that received a small to no adjustment. Thus, the effectiveness of the adjustment of the item ratings was reduced. The effectiveness of the item selection algorithm was further reduced in the second session because the item bank of the math game contained too few items. Because the user ratings were relatively stable, they answered many items with an item rating around a certain rating (i.e. all items with an item rating of around 2800). With each item they answered, the number of items of a certain difficulty that were eligible for selection decreased until there were no more items of a certain difficulty that the participant had not yet answered. This caused the discrepancy between the targeted likelihood of answering an item correctly and the estimated likelihood to increase from a discrepancy of ~1% to a discrepancy of up to 10%. The problem of not having enough items of a certain difficulty can be solved by increasing the number of items in the item bank.

There was no shortage of items for the imitation game. For this game, the reliability of the item ratings was mainly reduced due to the inaccuracy of the initial ratings. There was no data on which we could base the initial ratings, so the initial ratings were calculated based on a couple of assumptions. While we believe that this was enough to impose a rank order on the imitation items, the intervals between items ratings were meaningless. For example, we could not say that item #4 was X times more difficult than item #5. Only that item #4 is more difficult than item #5. The adjustment of the item ratings in between the first and second session did not improve the item ratings sufficiently to make the intervals between item ratings meaningful, because only a small number of imitation items was answered in the first session.

In short, the item ratings were not reliable enough to support controlling the percentage of items answered correctly. A solution would have been to have more sessions, or more participants to participate in the experiment. Because the ratings system is a self-correcting system, the item ratings become more reliable the more times they are answered. However, using more sessions or participants was not feasible for this study due to time constraints. Another solution is to use item banks that have already been calibrated by a rating system.

The second research question of this study is “to what extend is a sudden drop in performance related to the child’s motivation to play the current game with the social robot?”. The experiment was not specifically designed to answer this research question, as then all participants would have to play the games until they are bored. With such an experimental design it is not possible to use the free-choice paradigm, which we used to answer the first (primary) research question. The length of playing the games was set so that approximately half of the participants would be bored with the games and choose one of the alternative activities during the free-choice period. Thus, it was possible that enough participants would be bored enough, and would to no longer put the required effort into answering items correctly, to answer this research question.

During the experiment, one participant performed exceptionally poor at one point. The participant took less time to answer the items, than when items of the same difficulty were answered previously and afterwards (in the next session). However, it is unlikely that the drop in performance was primarily the result of a loss of motivation to play games with robot as the drop in performance occurred in the free-choice period, where the participant had chosen to continue playing with the robot and specifically chose to play the math game. A better explanation for the drop in performance is that the participant no longer had the attention to solve the math assignments, nor enough motivation to compensate for the loss of attention. Nonetheless, a drop in performance was detected, and a robot should consider responding to such an event. However, the adequate response to the performance drop may be different depending on the underlying cause of the drop in performance.

Given that only one drop in performance occurred, we do not have enough data to answer to what extend a sudden drop in performance is related to the child's motivation to play the current game with the robot. However, we believe that the detection system can be useful when it is combined with other measures, such as the time it takes a user to answer, the answer that is given, or the beliefs about how much the user likes each game. Such a system will be discussed in 6.2.6.

6.2 Improving the user modelling system

There are several ways in which the user-adaptive system can be improved. We will discuss several options that can improve the measurement precision of the rating system, an alternative approach to providing the user with the optimal challenge, and improvements to the detection system that detects when the user is no longer motivated to play the current game.

6.2.1 Improve measurement precision by taking the response time into account

The main shortcoming of our study was that the item ratings were not reliable enough to support manipulating the percentage of correct answers. Research by Eggen and Verschoor (2006) showed that when a rating system is used to increasing the percentage of correct answers, it will result in a loss of measurement precision. This is due to the rating system natural tendency to select items that has the same rating as the user. As a result, a user will need to answer more items in order to obtain a reliable estimate of the user's skill. Maris and van der Maas (2012) propose a novel measurement model that incorporates response times, as well as the outcome of an instance. The model only works if there is a time limit for answering items. If the rating of the item is lower than the rating of the user, then a fast response is more likely to be correct, whereas a slow response is more likely to be incorrect. When the rating of the item is higher than the rating of the item, the reverse is true. Fast responses are more likely to be incorrect, and slow responses are more likely to be correct. Thus, the response times can indicate how difficult a certain item was. This can be taken into account by increasing the change in rating when the user answers quickly and reducing the change in rating when the user answers slowly.

Klinkenberg and colleagues (2011) incorporated response times in their rating system. They found that incorporating response times resulted in much better measurement precision when the user only answers easy items (e.g. items for which the user has a high chance of answering it correctly). This allowed them to increase the number of correct answers from 50% to 75%, without a great loss of measurement precision.

Incorporating response times could improve our user-adaptive system. For the math game, response times could be included when a time limit is set. Since more difficult items generally take more computational operations to solve and/or more time to read the assignment, it can be

expected that more difficult items take longer to answer regardless of the skill of the user. If this true, then the time limit can be adjusted to reflect the difficulty of the item and the minimal time that is required to solve the assignment. For the imitation game, incorporating response times can be problematic because the response time depends on the length of the movement sequence, the time it takes the user to perform one movement, which depends on the speed at which the user moves his or her arms, how far the arms are moved. Also, the robot has to recognise the movements for which it is not desirable that the user perform the movements as fast as possible, and possibly incomplete, as that will likely make it more difficult to recognise which movement is performed by the user.

6.2.2 Adding a competition

The flow theory (Csikszentmihalyi, 1990) does not quantify when a person is optimally challenged. We quantified that a child user will generally be optimally challenged when they answer approximately 70% of the item correctly. As our study showed, controlling the percentage of correct answers requires a high measurement precision, which, as our study showed, can be problematic. A different approach to providing the user with the optimal challenge is to make the user compete with the robot. The robot and the user will take turns in answering an item, and whoever answered the most items correctly wins. The competition can be designed in a way that the user will win the competition in 70% of the time, even though the user answered only 50% of the item correctly. In this way the degree of challenge is not primarily determined by the difficulty of the items, but is determined by the difficulty of the opponent. The user may feel confident that he or she is skilled enough to beat the opposing robot.

Taking turns has already been implemented for some of the games used within the ALIZ-E project, but for our study, we removed the turn-taking component, because we wanted the participants to answer as many items as possible in a short period of time, so that we would have more data. Also, previous studies (Robben, 2011; Janssen et al., 2011) indicated that turn-taking made the games more fun for the user. For our experiment, it was critical that the games were not too fun to play, as then most or all participants would continue playing the games during the free-choice period.

6.2.3 Selecting items on content and difficulty

Within the ALIZ-E project, one of the goals of playing educational games with the robot is to teach the children how to manage their health independently by teaching them relevant knowledge and skills. Managing the difficulty of the game items makes the game more fun to play, because the games are challenging. Furthermore, it increases the efficiency of learning, as the children will play the games at their own skill level. For some games it may be beneficial to select items not only based

on the desired difficulty, but also on the content. For example, a child is playing a math game and is proficient at solving assignments involving subtraction, but poor at solving assignments involving multiplication. It may be best for improving the child's self-management skills to have the child practice with multiplication assignments, rather than with assignments about any of the knowledge domains incorporated in the game.

The rating system can be used to model the user's knowledge on each of the domains, by assigning a user rating to each of the domains. Each knowledge domain has a separate item bank. Items can then be select for the domain the user knows the least about by selecting the lowest user's rating on all the knowledge domains. In theory, separate items banks for each domain can lead to a greater measurement precision, because each individual item bank will be more homogeneous. However, the user will have to answer several items for each domain before the rating of this user becomes reliable. This problem can be partly solved by using the correlation between the knowledge domains to set the initial ratings, making them more reliable. The knowledge domains of a math game are straightforward, but for games where the knowledge domains are not as straightforward (i.e. a quiz about a person's disability), it may be better to use an additional model to model the domain knowledge of the user (for an overview on such models see Brusilovsky & Millán, 2007). Also, not all games are suitable for selecting items based on content and difficulty, as for some games, no distinction can be made between different knowledge domains. The imitation game, used in this study, is an example of this.

6.2.4 Improving the initial user rating for the second game

For the experiment, we based the initial user ratings on the pilot experiment and the recommendation of the teacher of the participants. However, in a real scenario, the robot will not have access to such information. The initial user rating can be set based on the age of the user. Also, because we found a strong correlation between the user ratings on both games, the user rating of the game played as second can be set based on the user's rating on the first game and the user's age.

6.2.5 Item reuse

In our design, we avoided the repetition of item; items that were answered would not occur later in the same session. For longer sessions, where the user answers more items, it may be more efficient to reuse items within a session. In this way, the item bank can be smaller, compared to when items are not reused with in a session. Also, because each item will be answered more often, the item ratings will be more reliable estimates of the items true difficulty.

6.2.6 Interpreting poor performance

We used a rating system for our social robot for two purposes. First, it is used to estimate the user's skill and the difficulty of the items, which allows us to adapt the difficulty of the game to the user. Second, we use the estimations, calculated by the rating system, to make inferences about the user's motivation. Poor performance can be an indication that the user is not motivated to perform well or that the user pays enough attention to the game. Low motivation or attention can be a reason to initiate (or refrain from) other functionalities of the robot, like educating the user about a certain topic.

Based on the results of the experiment, we conclude that the detection of low motivation can be improved by using additional information. The expected response time can be estimated based on how difficult the item is for the user and on the average response time on the item. More difficult items will generally take more time to answer than items that are easy for the user. The average response time can be used to normalize the response times, which is necessary as each item has a minimal amount of time that is required to answer it. The more difficult items are generally more complex and, therefore, take more time to be answered, regardless of the skill of the user.

The estimated response time can be compared to the actual response time, and when the user is performing exceptionally poor, but takes adequate time to answer the items, then the user appears to have the motivation and attention required to perform well. In this case, the poor performance may be the result of bad luck or an overestimated user rating. In case of the latter, the rating deviation can be increased to reflect the increased uncertainty about the user rating. When the user is performing exceptionally poor, and is answering consistently faster than expected, then the user is not putting enough effort into the game. This can be due to a loss of attention, or a loss in motivation to play the game. The difference between the two is that in the case of a loss of attention the user is still wants to play the game, but is finding it difficult to pay enough attention to the game (i.e. due to being tired), while in the case of a loss of motivation the user would rather do something else. The loss of attention can be the result of the user being preoccupied with something else, like having received bad news. In this case, it may be beneficial for the robot to ask a certain question so that it may discern the underlying cause of the loss of attention and respond accordingly. For example, the robot may ask "You seem preoccupied. Is there something on your mind you wish to talk about?". In the case of a loss of motivation, the robot may suggest playing another game, or motivate the user to put more effort into playing the current game. In the end, the best course of action has to be decided based on all current goals and beliefs. The user's state is taken into account, so that does not necessarily mean that the robot should adapt its behaviour to please the user. A rational agent can evaluate the situation and make a decision on how to deal with poor performance, in relation to the current goals the agent wants to achieve and the other beliefs it holds.

6.3 General conclusion

In this master's thesis, we have presented our design of a user-adaptive system that estimates the user's skill and performance with the use of a rating system. The user-adaptive system is used to adapt the difficulty of a game to the user's skill in order to provide the user with the optimal challenge. The assumption was that because the games will always be challenging, the games will remain fun to play for a longer period of time. Furthermore, the user-adaptive system was capable of detecting exceptionally good and poor performance, to which the robot responded accordingly. The robot gave the user a compliment when the user was performing exceptionally good, and changed the game when the user's performance was exceptionally poor.

In general, using a Bayesian rating system to estimate a user's skill and performance has four major advantages. First, studies have shown that a high measurement precision can be achieved (Glickman, 1999; Klinkenberg, Straatmeier & Van der Maas, 2011), leading to reliable estimations of the user's skill and of the difficulty of the game items. Second, the robot could possibly detect events that are exceptional given the user's skill and respond to such events, increasing its social intelligence. Third, the rating system is a self-correcting system that continuously adjusts the estimated skill of the user and the difficulty of the game items. When the user has become more or less skilled, or a game item has become more or less difficult, the ratings are adjusted without the need for human intervention. And last, the difficulty of a game can be adapted without needing to explicitly ask the user for information regarding the skill of the user or the preferred difficulty. Thus, the interaction between the robot and user is not interrupted. However, due to methodological shortcomings, we were not been able to fully utilise the advantages that the use of a rating system can offer.

We could not determine to what extend the participants who played the games at the optimal challenge were intrinsically motivated. In the end, we did not have enough data to improve the measurement precision up to a point that it was possible to control the percentage of items answered correctly. The measurement precision of the rating system can be improved by using item banks that already have reliable item ratings. The rating system itself can also be improved by incorporating response times. Also, selecting items based on content and difficulty could improve the measurement precision. Alternatively, creating a competition between the user and the robot is a different approach to providing the user with the optimal challenge. It may be possible to shift the focus from answering each item correctly to winning the game, by having the user compete with the robot. This way, the optimal challenge is dependent on the "skill" of the robot, instead of the difficulty of the items.

The system to detect exceptional performance has merit in its use and ability to reliably detect exceptional performance by the user. We found some evidence that the probability of the user answering an item correctly can be used to determine when the user's performance is exceptionally

poor. However, further research is needed to conclude if, and how effective, exceptional performance can be detected. The detection system can be improved by including response times. Also, by adding additional behaviours, like asking a certain question when the user's performance is poor, the robot may be able to discern the underlying cause of the poor performance and respond accordingly.

References

Anderson, R., Manoogian, S. T., & Reznick, J. S. (1976). The undermining and enhancing of intrinsic motivation in preschool children. *Journal of Personality and Social Psychology*, Vol. 34, pp. 915–922

Baxter, P., Belpaeme, T., Cañamero, L., Demiris, Y., Enescu, V., Hiolle, A., Kruijff-Korbayová, I., Looije, R., Nalin, M., Neerincx, M. A., Sahli, H., Sommavilla, G., Tesser, F., & Wood, R. (2011). Long-Term Human-Robot Interaction with Young Users. In Miyake, N., Ishiguro, H., Dautenhahn, K. & Nomura, T.(eds.), *Proceedings of the Workshop Robots with Children: Practices for Human-Robot Symbiosis, at the 6th ACM/IEEE International Conference on Human-Robot Interaction*.

Bates, J. (1994). The role of emotion in believable agents. *Communications of the ACM*, Vol. 37 (7), pp. 122-125.

Bradley, M. M., & Lang, P. J (1994). Measuring emotion: The Self-Assessment Manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, Vol. 25 (1), pp. 49-59.

Breazeal, C. (2003). Toward sociable robots. *Robotics and Autonomous Systems*, Vol. 42 (3-4), pp. 167-175.

Breazeal, C. (2005). Socially Intelligent Robots. *Interactions ACM*, Vol. 12, pp. 19-22.

Brusilovsky, P., & Millán, E. (2007). *User Models for Adaptive Hypermedia and Adaptive Educational Systems*. In P. Brusilovsky, A. Kobsa, & W. Nejdl, The Adaptive Web (pp. 3-53). Heidelberg: Springer-Verlag.

Boggiano, A. K., Main, D. S., & Katz, P. A. (1988). Children's preference for challenge: The role of perceived competence and control. *Journal of Personality and Social Psychology*, Vol. 54, pp. 134–141.

Borghouts, C., Buter, A., Dekker, J., Hoogenberg, E., Kopmels, D., & van Oostenbrugge, M. (2005). *Interactie in Rekenen.* Bazalt Educatieve Uitgaven.

Calvo, R. A., & D'Mello, S. (2010). Affect Detection: An Interdisciplinary Review of Models, Methods, and their Applications. *IEEE Transactions on Affective Computing*, Vol. 1 (1), pp. 18–37.

Chang, C. W., Lee, J. H., Chao, P. Y., Wang, C. Y., & Chen, G. D. (2010). Exploring the Possibility of Using Humanoid Robots as Instructional Tools for Teaching a Second Language in Primary School. *Educational Technology and Society*, Vol. 13 (2), pp. 13-24.

Csikszentmihaly, M. (1990). *Flow: the psychology of the optimal experience.* New York: HarperCollins Publishers.

Dastani, M. (2011). Programming Multi-Agent Systems. *Proceedings of the 12th International Workshop on Agent-Oriented Software Engineering*, pp. 23-52.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology*, Vol. 18 (1), pp. 105-115.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behaviour*. New York: Plenum.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, Vol. 125 (6), pp. 627-668.

Duffy, B. (2000). *The Social Robot*. PhD Thesis, Department of Computer Science, University of Dublin.

Elo, A. E. (1978). *The Rating of Chess Players Past and Present*. New York: Arco.

Fasola, J., & Matarić, M. M. (2012). Using Socially Assistive Human-Robot Interaction to Motivate Physical Exercise for Older Adults. *Proceedings of the IEEE - Special Issue on Quality of Life Technology*, Vol. 100 (8), pp. 2512-2526.

Fogg, B. J. (2002). *Persuasive technology: Using computers to change what we think and do*. San Francisco, CA: Morgan Kaufmann.

Fong, T. W., Nourbakhsh, I., & Dautenhahn, I. K. (2003). A Survey of Socially Interactive Robots: Concepts, Design, and Applications. *Robotics and Autonomous Systems*, Vol. 42(3-4), pp. 142-166.

Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society Series C-Applied Statistics*, Vol. 48, pp. 377-394.

Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M., Mundell, A., Rosenthal, S., Sellner, B., Simmons, R., Snipes, K., Schultz, A. C., & Wang, J. (2005). Designing robots for long-term social interaction. *Proceedings of the International Conference on Intelligent Robots and Systems 2005*, pp. 1338-1343.

Goffree, F., & Oonk, W. (2004). *Reken Vaardig: op weg naar basale en professionele gecijferdheid.* Noordhoff Uitgevers B. V.

Gottfried, A. E. (1990). Academic intrinsic motivation in young elementary school children. *Journal of Educational Psychology*, Vol. 82, pp. 525–538.

Green, A., Hüttenrauch, H., & Eklundh, K. S. (2004). Applying the Wizard-of-Oz Framework to Cooperative Service Discovery and Configuration. *Proceedings of the 2004 IEEE International Workshop on Robot and Human Interactive Communication*, pp. 575-580.

Grolnick, W. S., & Ryan, R. M. (1989). Parent styles associated with children's self regulation and competence in schools. *Journal of Educational Psychology*, Vol. 81, pp. 143–154.

Hashimoto, T., Kato, N., & Kobayashi, H. (2010). Study on Educational Application of Android Robot SAYA: Field Trial and Evaluation at Elementary School. *Proceedings of the Third International Conference on Intelligent Robotics and Applications*, pp. 505-516.

Hegel, F., Muhl, C., Wrede, B., Hielscher-Fastabend, M., & Sagerer, G. (2009). Understanding Social Robots. *The Second International Conferences on Advances in Computer-Human Interactions (ACHI)*, pp. 169 - 174.

Henderlong, J., & Lepper, M. R. (2002). The Effects of Praise on Children's Intrinsic Motivation: A Review and Synthesis. *Psychological Bulletin*, Vol. 128 (5), pp. 774-795.

Herbrich, R., Minka, T., & Graepel, T. (2007). TrueSkill(TM): A Bayesian Skill Rating System. *Proceedings of Advances in Neural Information Processing Systems* 20, pp. 569-576.

Hindriks, K. V. (2009). Programming Rational Agents in GOAL. In R. H.Bordini, M. Dastani, J. Dix & A. E. F. Seghroudchni (Eds.), *Multi-Agent Programming: Languages and Tools and Applications* (pp. 119 – 157). New York: Springer.

Janssen, J. B., van der Wal, C. C., Neerincx, M. A., & Looije, R. (2011). Motivating Children to Learn Arithmetic with an Adaptive Robot Game. *ICSR'11 Proceedings of the Third international conference on Social Robotics*, pp. 153-162.

Johnson, A., & Taatgen, N.A. (2005). User modeling. In Robert W. Proctor and Kim-Phuong L. Vu (Eds.), *The Handbook of Human Factors in Web Design* (pp. 424-438). Mahwah, NJ: Erlbaum.

Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. (2004). Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial. *Human–Computer Interaction*, Vol. 19 (1), pp. 61 - 84.

Kidd, C. D., & Breazeal, C. (2004). Effect of a robot on user perceptions. *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vol. 4, pp. 3559-3564.

Kidd, C. D., & Breazeal, C. (2008). Robots at Home: Understanding Long-Term Human-Robot Interaction. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems 2008*, pp. 3230-3235.

Kiesler, S., Powers, A., Fussell, S. R., & Torrey, C. (2008). Anthropomorphic Interactions with a Robot and Robot-like Agent. *Social Cognition*, 26 (2), 169-181.

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. J. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, Vol. 57 (2), pp. 1813–1824.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in a big city. *International Journal of Artificial Intelligence in education*, Vol. 8, pp. 30-43.

Komatsu, T., & Abe, Y. (2008). Comparing an On-Screen Agent with a Robotic Agent in Non-Face-to-Face Interactions. *Intelligent Virtual Agents, Lecture Notes in Computer Science*, Vol. 5208, pp. 498-504.

Kobsa, A., Koenemann, J., & Pohl, W. (2001). Personalized Hypermedia Presentation Techniques for Improving Customer Relationships. *The Knowledge Engineering Review*, Vol. 16 (2), 111-155.

Lang, P. J. (1985). *The Cognitive Psychophysiology of Emotion: Anxiety and the Anxiety Disorders*. Hillsdale, NJ: Lawrence Erlbaum.

Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors*, Vol. 46 (1), pp. 50–80.

Liu, C., Agrawal, P., Sarkar, N., & Chen, S. (2009). Dynamic Difficulty Adjustment in Computer Games Through Real-Time Anxiety-Based Affective Feedback. *International Journal of Human-Computer Interaction*, Vol. 25 (6), pp. 506–529.

Maris, G., & van der Maas, H. L. J. (2012). Speed-accuracy response models: scoring based on response time and accuracy. *Psychometrika*, Vol. 77 (4), pp. 615-633.

Naito, H., & Takeuchi, Y. (2009). Promotion of Efficient Cooperation by Sharing Environment with an Agent Having a Body in Real World. *Progress in Robotics, Communications in Computer and Information Science*, Vol. 44, pp. 128-133.

Pohl, W. (1999). Logic-Based Representation and Reasoning for User Modeling Shell Systems. *User Modeling and User-Adapted Interaction*, Vol. 9, pp. 217–282.

Powers, A., Kiesler, S., Fussell, S., & Torrey, C. (2007). Comparing a computer agent with a humanoid robot. *Proceedings of the Conference on Human Robot Interaction HRI 2007,* pp. 145-152.

Read, J., MacFarlane, S., & Casey, C. (2002). Endurability, Engagement and Expectations: Measuring Children's Fun. *Interaction Design and Children*, Eindhoven, Shaker Publishing.

Reeve, J., & Deci, E. L. (1996). Elements of the Competitive Situation That Affect Intrinsic Motivation. *Society for Personality and Social Psychology*, Vol. 22 (1), pp. 24-33.

Reeves, B., & Nass, C (1996). *The Media Equation*. Cambridge University Press.

Robben, S. (2011). It's NAO or Never! Facilitate Bonding Between a Child and a Social Robot: Exploring the Possibility of a Robot Adaptive to Personality, *Master's thesis*.

Russell, S. J., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach* (2nd ed.). Upper Saddle River, New Jersey: Prentice Hall

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology,* Vol. 25, pp. 54–67.

Saerbeck, M., Schut, T., Bartneck, C., & Janse, M. D. (2010). Expressive Robots in Education: Varying the Degree of Social Supportive Behavior of a Robotic Tutor. *Proceedings of the 28th International Conference on Human factors in Computing Systems*, pp. 1613-1622.

Tanaka, F., & Ghosh, M. (2011). The Implementation of Care-Receiving Robot at an English Learning School for Children. *Proceedings of the 6th international conference on Human-robot interaction*, pp. 265-266.

Tanaka, F., Cicourel, A., & Movellan, J. R. (2007). Socialization Between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 104 (46), pp. 17954-17958.

Vallerand, R., Gauvin, L., & Halliwell, W. (1986). Negative effects of competition on children's intrinsic motivation, *The Journal of Social Psychology*, Vol. 126 (5), pp. 649-657.

Wada, K., Shibata, T., Saito, T., Sakamoto, K., & Tanie. K. (2003). Psychological and social effects of one year robot assisted activity on elderly people at a health service facility for the aged. *Proceedings of the 2005 IEEE International Conference on Robotics and Automation,* no. April, pp. 2785-2790.

Weinberg, R. S., & Ragan, J. (1979). Effects of competition, success-failure, and sex on intrinsic motivation. *Research Quarterly,* Vol. 50 (3), pp. 503-510.

Appendix A

Dialogs

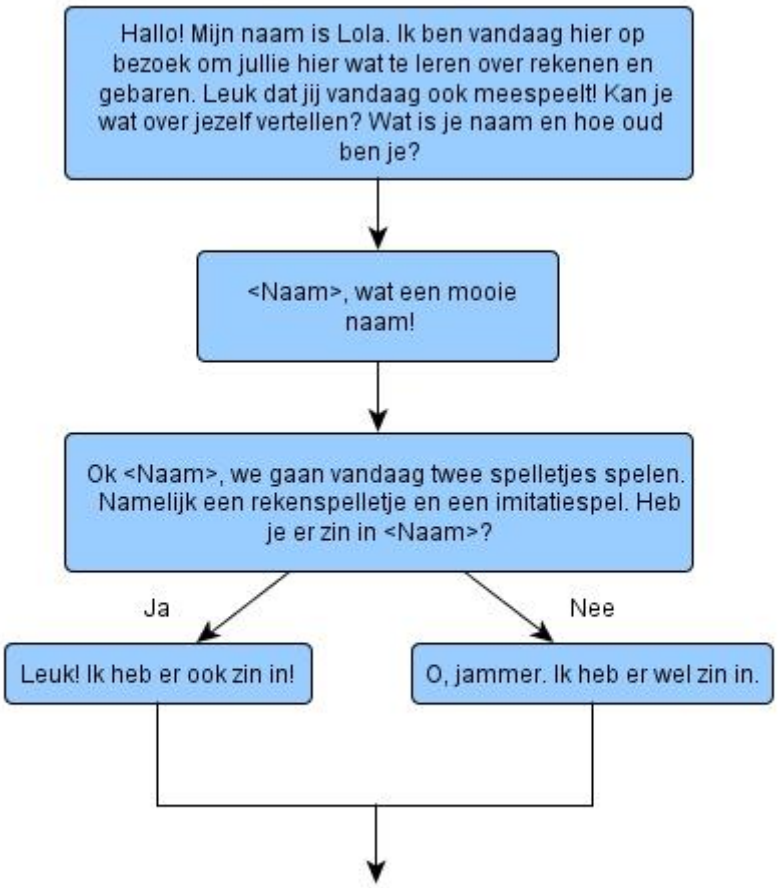


Figure A1. Dialogs for during the introduction.

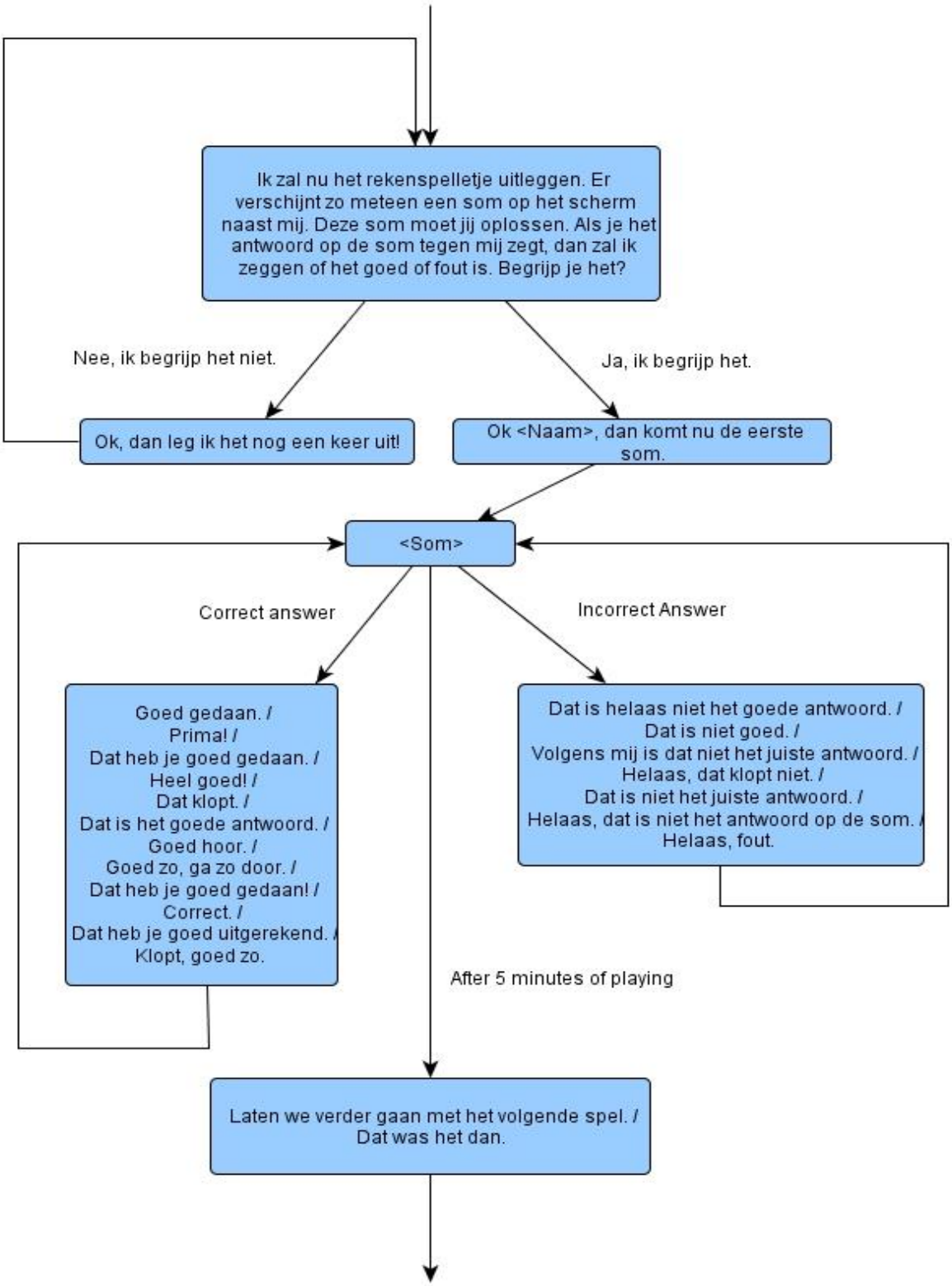


Figure A2. Dialogs for the math game.

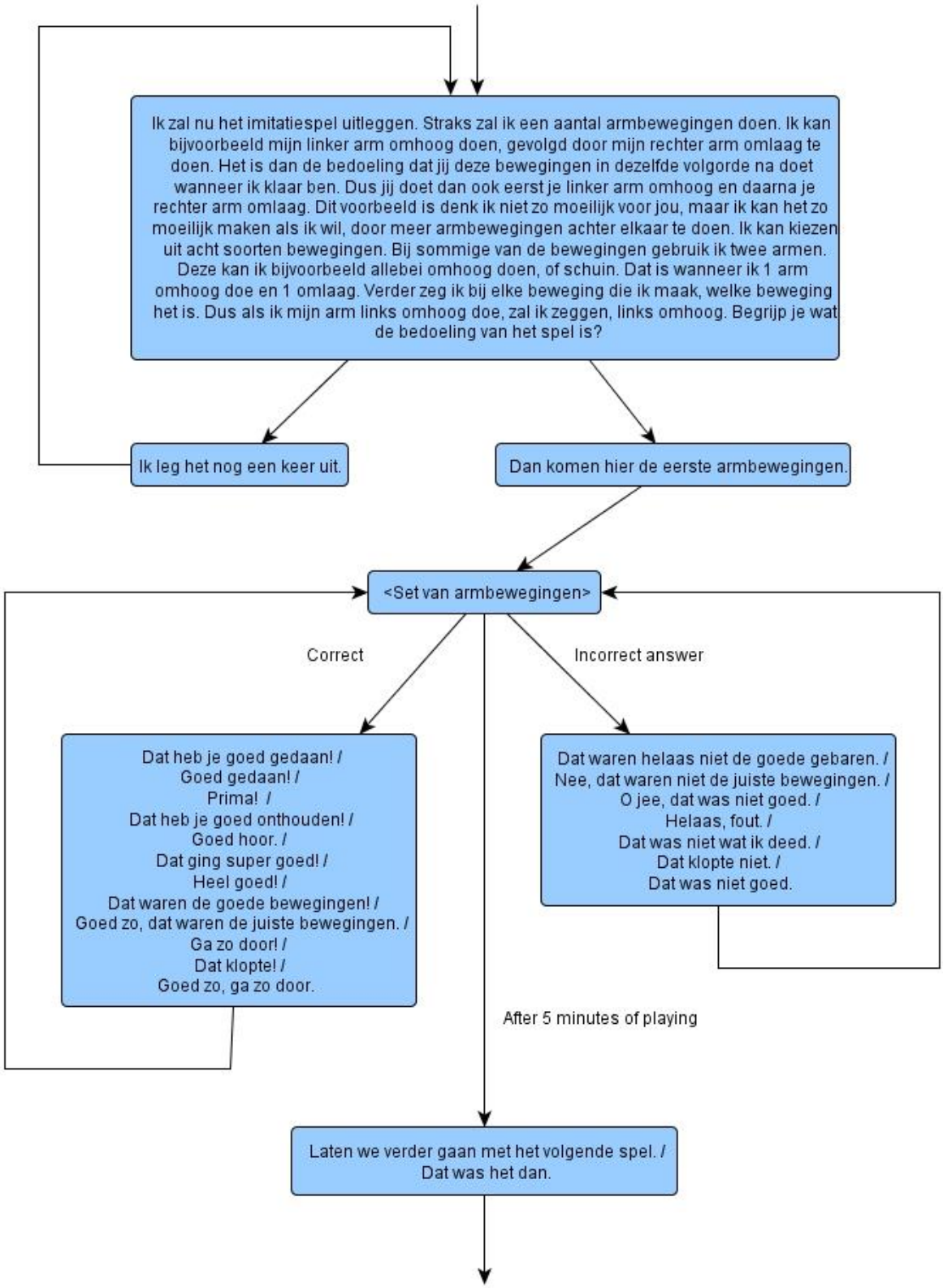


Figure A3. Dialogs for the imitation game.

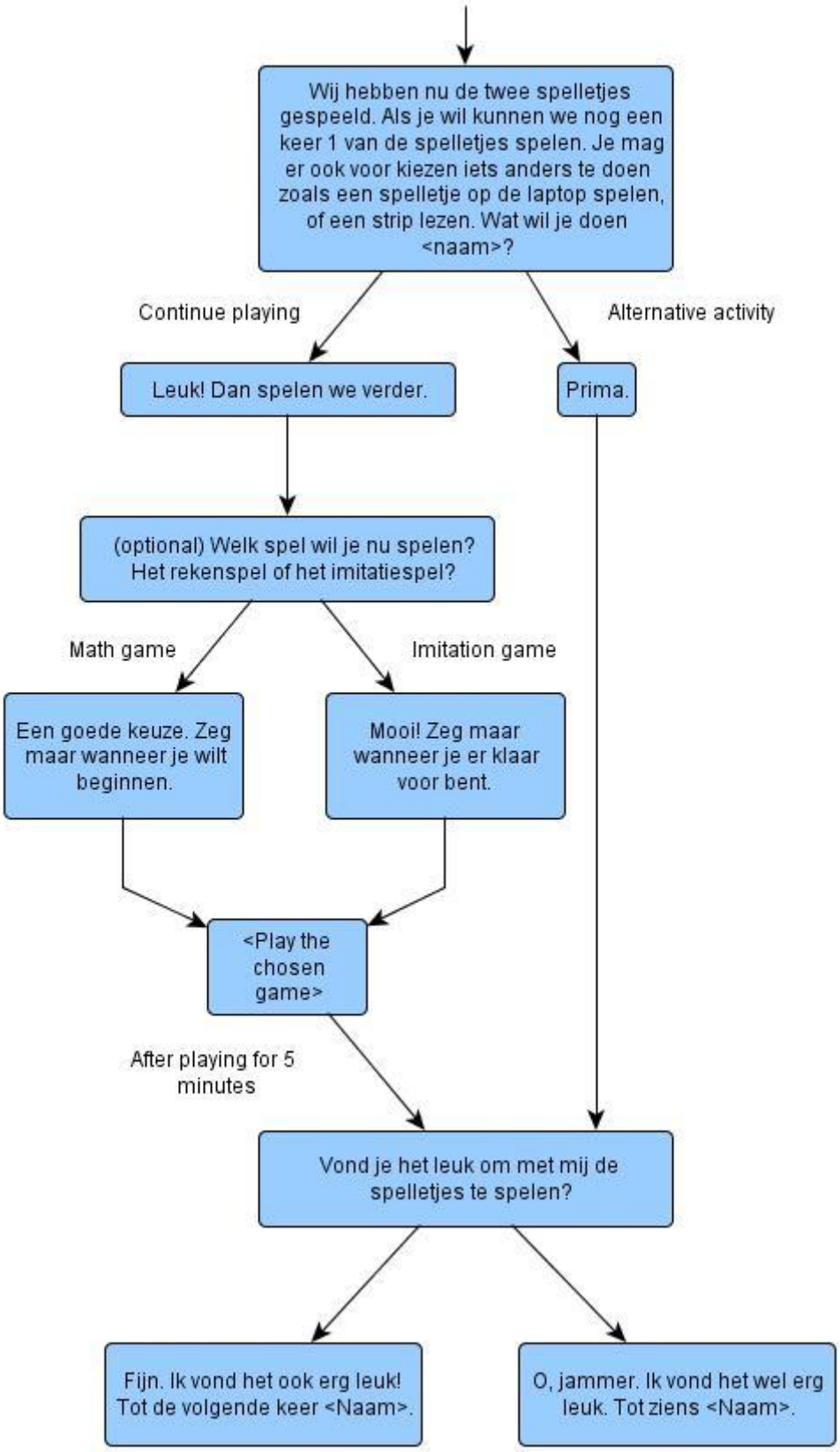


Figure A4. Dialogs for the free-choice period and ending.

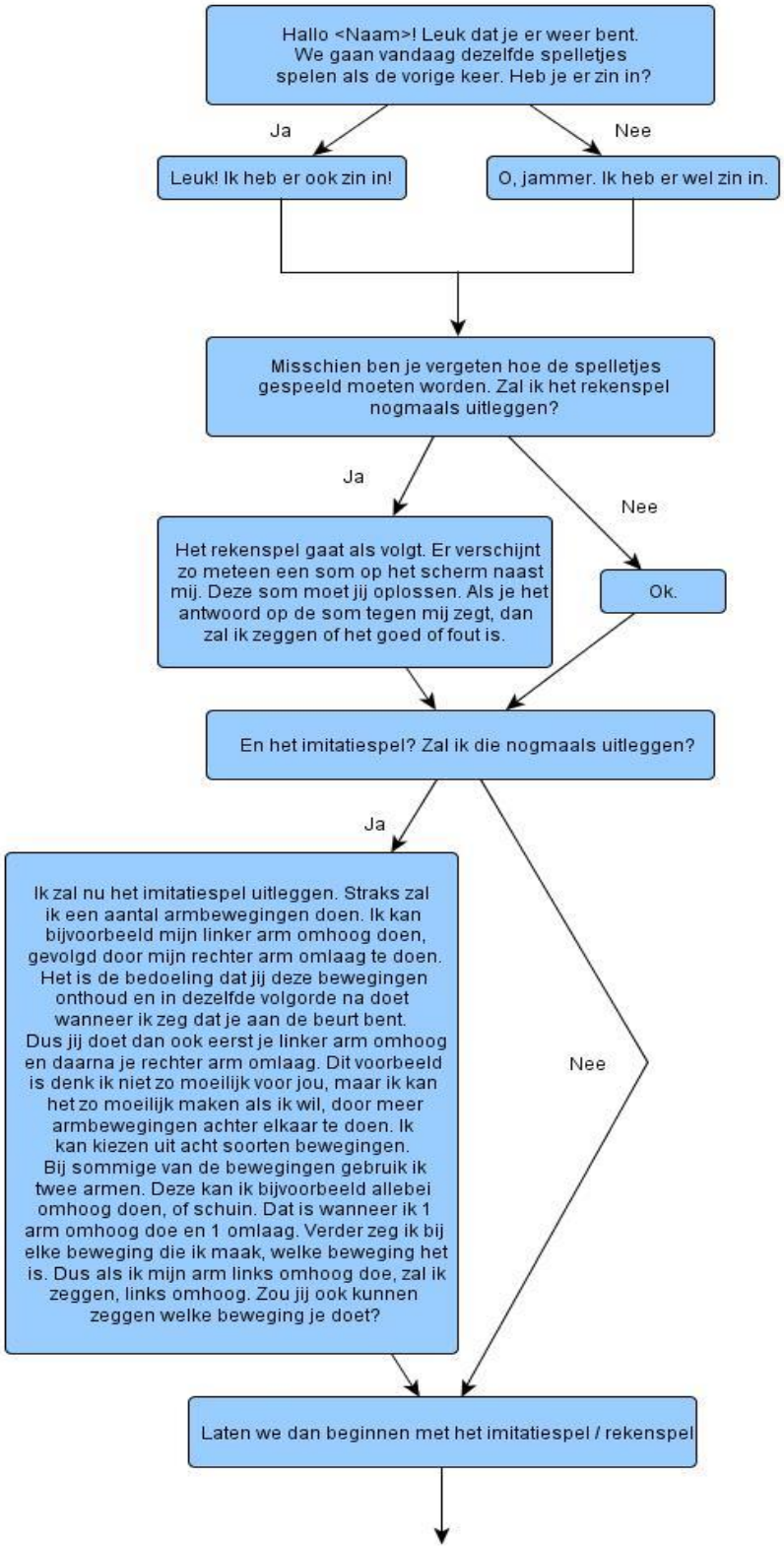


Figure A5. Dialogs for the introduction of the second session.

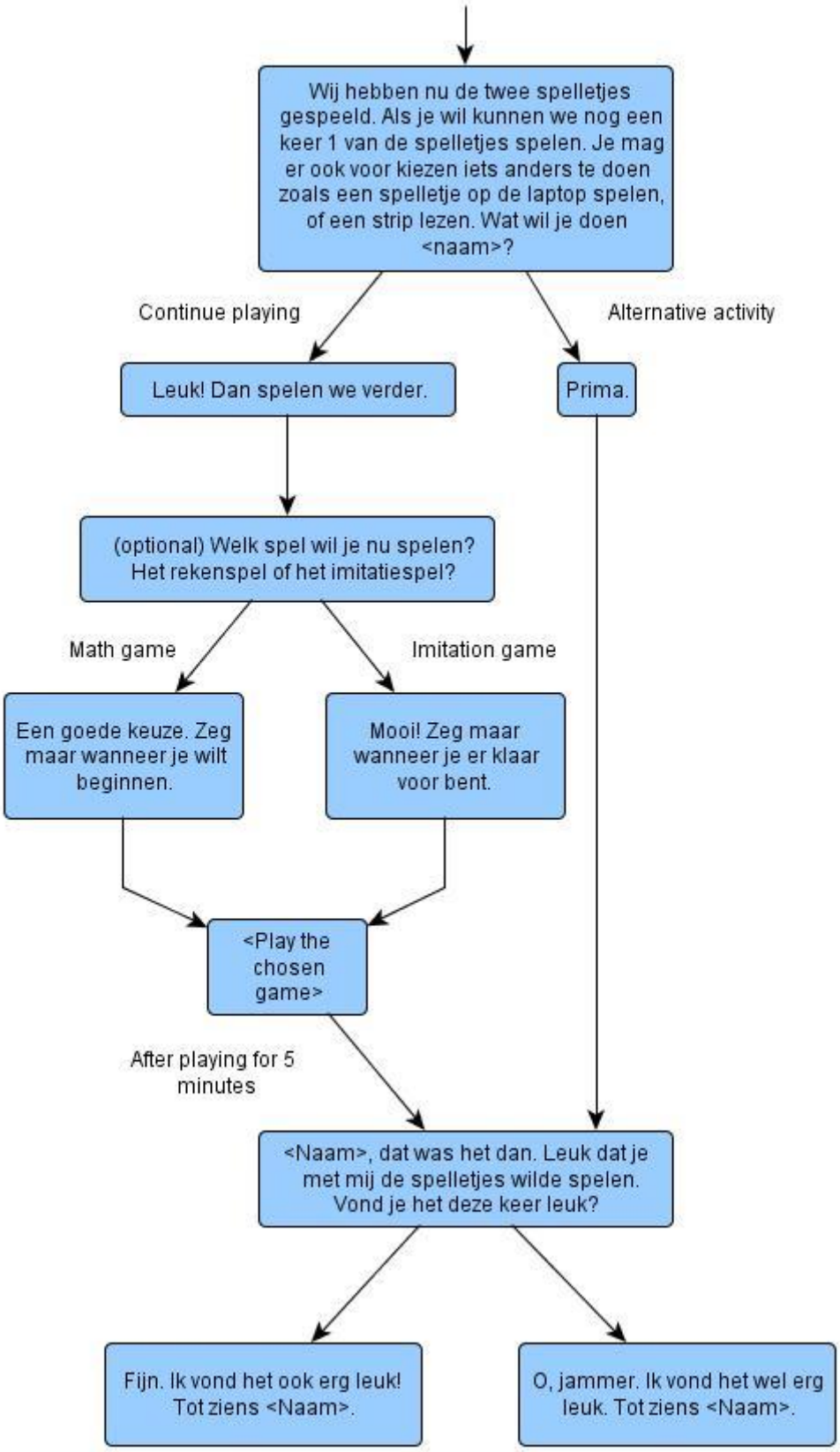


Figure A6. Dialogs for the free-choice period and the ending for the second session.

Appendix B

The Game Items

Table B1

Ratings and Deviations for Each of the Math Items

Item	Rating		Item	Rating	
	Initial (Dev)	Updated (Dev)[a]		Initial (Dev)	Updated (Dev)[a]
1 + 1	100 (150)	-	45 - 4	500 (150)	-
2 + 1	100 (150)	-	89 - 7	500 (150)	-
3 + 4	100 (150)	-	54 - 3	500 (150)	-
5 + 3	100 (150)	-	86 - 6	500 (150)	-
6 + 1	100 (150)	-	22 - 1	500 (150)	-
8 + 2	100 (150)	-	95 - 3	500 (150)	-
5 + 2	100 (150)	-	67 - 5	500 (150)	-
1 + 8	100 (150)	-	65 - 4	500 (150)	-
6 + 4	100 (150)	-	41 - 1	500 (150)	-
3 + 5	100 (150)	-	98 - 6	500 (150)	-
8 - 4	200 (150)	-	36 + 8	600 (150)	-
6 - 2	200 (150)	-	54 + 9	600 (150)	-
9 - 7	200 (150)	-	57 + 6	600 (150)	-
8 - 1	200 (150)	-	23 + 9	600 (150)	-
6 - 6	200 (150)	-	32 + 9	600 (150)	-
5 - 2	200 (150)	-	64 + 8	600 (150)	-
8 - 6	200 (150)	-	12 + 9	600 (150)	-
7 - 2	200 (150)	-	65 + 9	600 (150)	-
3 - 0	200 (150)	-	73 + 8	600 (150)	-
4 - 1	200 (150)	-	48 + 5	600 (150)	-
5 + 7	300 (150)	-	36 - 9	700 (150)	-
6 + 7	300 (150)	-	62 - 8	700 (150)	-
8 + 9	300 (150)	-	94 - 7	700 (150)	-
7 + 8	300 (150)	-	84 - 7	700 (150)	-
7 + 6	300 (150)	-	46 - 9	700 (150)	-
9 + 8	300 (150)	-	26 - 9	700 (150)	-
5 + 6	300 (150)	-	54 - 7	700 (150)	-
6 + 5	300 (150)	-	63 - 8	700 (150)	-
7 + 7	300 (150)	-	84 - 9	700 (150)	-
8 + 7	300 (150)	-	25 - 7	700 (150)	-
64 + 2	400 (150)	-	1 x 5	800 (150)	-
88 + 1	400 (150)	-	3 x 5	800 (150)	-
26 + 3	400 (150)	-	2 x 4	800 (150)	-
27 + 3	400 (150)	-	8 x 4	800 (150)	-
42 + 7	400 (150)	-	9 x 3	800 (150)	-
72 + 6	400 (150)	-	6 x 2	800 (150)	-
31 + 8	400 (150)	-	9 x 4	800 (150)	-
91 + 8	400 (150)	-	7 x 1	800 (150)	-
46 + 4	400 (150)	-	8 x 0	800 (150)	-
90 + 5	400 (150)	-	6 x 3	800 (150)	-

Item	Rating	
	Initial (Dev)	Updated (Dev)[a]
41 + 20	900 (150)	-
55 + 40	900 (150)	-
32 + 60	900 (150)	-
67 + 30	900 (150)	-
28 + 70	900 (150)	-
61 + 20	900 (150)	-
25 + 40	900 (150)	-
12 + 60	900 (150)	-
37 + 30	900 (150)	-
11 + 20	900 (150)	-
39 - 20	1000 (150)	-
21 - 10	1000 (150)	-
45 - 30	1000 (150)	-
73 - 50	1000 (150)	-
34 - 30	1000 (150)	-
69 - 20	1000 (150)	-
41 - 10	1000 (150)	-
35 - 30	1000 (150)	-
83 - 40	1000 (150)	-
64 - 50	1000 (150)	-
126 + 99	1100 (150)	-
226 + 199	1100 (150)	-
326 + 199	1100 (150)	-
426 + 399	1100 (150)	-
245 + 301	1100 (150)	-
255 + 401	1100 (150)	-
265 + 501	1100 (150)	-
275 + 601	1100 (150)	-
226 + 99	1100 (150)	-
444 + 199	1100 (150)	-
700 - 99	1200 (150)	-
700 - 101	1200 (150)	-
800 - 199	1200 (150)	-
500 - 201	1200 (150)	-
500 - 199	1200 (150)	-
499 - 199	1200 (150)	-
456 - 300	1200 (150)	-
456 - 301	1200 (150)	-
456 - 299	1200 (150)	-
800 - 201	1200 (150)	-

Item	Rating	
	Initial (Dev)	Updated (Dev)
6 x 7	1300 (150)	-
8 x 8	1300 (150)	-
8 x 9	1300 (150)	-
7 x 8	1300 (150)	-
6 x 8	1300 (150)	-
8 x 7	1300 (150)	-
7 x 6	1300 (150)	-
9 x 8	1300 (150)	-
7 x 9	1300 (150)	-
6 x 6	1300 (150)	-
63 + 13	1400 (150)	-
63 + 14	1400 (150)	-
63 + 15	1400 (150)	-
63 + 16	1400 (150)	-
63 + 17	1400 (150)	-
52 + 14	1400 (150)	-
52 + 15	1400 (150)	-
52 + 16	1400 (150)	-
52 + 17	1400 (150)	-
52 + 18	1400 (150)	-
91 - 88	1500 (150)	-
92 - 88	1500 (150)	-
90 - 87	1500 (150)	-
91 - 87	1500 (150)	-
92 - 87	1500 (150)	-
72 - 69	1500 (150)	-
72 - 68	1500 (150)	-
73 - 69	1500 (150)	-
73 - 68	1500 (150)	-
75 - 68	1500 (150)	-
500 - 2	1600 (150)	-
801 - 3	1600 (150)	-
302 - 4	1600 (150)	-
700 - 5	1600 (150)	-
905 - 6	1600 (150)	-
201 - 4	1600 (150)	-
503 - 6	1600 (150)	-
802 - 5	1600 (150)	-
405 - 7	1600 (150)	-
904 - 8	1600 (150)	-

Item	Rating		Item	Rating	
	Initial (Dev)	Updated (Dev)[a]		Initial (Dev)	Updated (Dev)[a]
18 : 2	1700 (150)	-	3 x 34	2000 (150)	-
15 : 3	1700 (150)	-	24 x 2	2000 (150)	-
24 : 4	1700 (150)	-	35 x 2	2000 (150)	-
35 : 5	1700 (150)	-	26 x 5	2000 (150)	-
36 : 6	1700 (150)	-	52 x 5	2000 (150)	-
81 : 9	1700 (150)	-	7 x 13	2000 (150)	-
56 : 7	1700 (150)	-	2 x 28	2000 (150)	-
90 : 10	1700 (150)	-	31 x 4	2000 (150)	-
27 : 3	1700 (150)	-	44 x 3	2000 (150)	-
21 : 7	1700 (150)	-	11 x 6	2000 (150)	-
3 x 40	1800 (150)	-	42 x 5	2000 (150)	-
2 x 70	1800 (150)	-	6 x 31	2000 (150)	-
6 x 30	1800 (150)	-	27 x 4	2000 (150)	-
90 x 4	1800 (150)	-	22 x 3	2000 (150)	-
40 x 6	1800 (150)	-	41 x 3	2000 (150)	-
8 x 90	1800 (150)	-	5 x 12	2000 (150)	-
5 x 80	1800 (150)	-	62 x 2	2000 (150)	-
2 x 50	1800 (150)	-	5 x 16	2000 (150)	-
5 x 60	1800 (150)	-	6 x 14	2000 (150)	-
50 x 3	1800 (150)	-	5 x 35	2000 (150)	-
180 : 2	1900 (150)	1831 (135)	350 + 220	2100 (150)	1840 (98)
180 : 20	1900 (150)	-	180 + 320	2100 (150)	1920 (112)
150 : 3	1900 (150)	-	572 + 120	2100 (150)	1986 (124)
150 : 30	1900 (150)	-	285 + 110	2100 (150)	2053 (139)
240 : 4	1900 (150)	-	330 + 150	2100 (150)	-
240 : 40	1900 (150)	-	209 + 191	2100 (150)	-
350 : 5	1900 (150)	-	409 + 111	2100 (150)	-
350 : 50	1900 (150)	-	538 + 161	2100 (150)	-
360 : 6	1900 (150)	-	145 + 141	2100 (150)	-
360 : 60	1900 (150)	-	289 + 111	2100 (150)	-
2 x 44	2000 (150)	1907 (133)	40 x 25	2200 (150)	2173 (99)
3 x 31	2000 (150)	1969 (141)	80 x 25	2200 (150)	2154 (113)
7 x 11	2000 (150)	-	20 x 50	2200 (150)	2075 (110)
4 x 21	2000 (150)	-	40 x 50	2200 (150)	2159 (123)
5 x 51	2000 (150)	-	20 x 25	2200 (150)	2326 (123)
6 x 31	2000 (150)	-	25 x 80	2200 (150)	2156 (123)
4 x 42	2000 (150)	-	50 x 80	2200 (150)	2162 (122)
2 x 27	2000 (150)	-	25 x 40	2200 (150)	2198 (130)
4 x 34	2000 (150)	-	50 x 40	2200 (150)	2105 (130)
5 x 27	2000 (150)	-	50 x 20	2200 (150)	2108 (130)

Item	Rating		Item	Rating	
	Initial (Dev)	Updated (Dev)[a]		Initial (Dev)	Updated (Dev)[a]
220 - 205	2300 (150)	2031 (88)	628 - 357	2600 (150)	2538 (105)
525 - 205	2300 (150)	2040 (95)	951 - 357	2600 (150)	2512 (111)
745 - 244	2300 (150)	2147 (103)	358 - 285	2600 (150)	2546 (116)
789 - 188	2300 (150)	2143 (114)	624 - 589	2600 (150)	2512 (129)
346 - 143	2300 (150)	2261 (123)	485 - 292	2600 (150)	2655 (138)
432 - 128	2300 (150)	2179 (122)	24 x 12	2700 (150)	2665 (105)
861 - 459	2300 (150)	2351 (122)	26 x 12	2700 (150)	2583 (107)
653 - 651	2300 (150)	2210 (129)	25 x 24	2700 (150)	2495 (110)
657 - 155	2300 (150)	2214 (130)	50 x 12	2700 (150)	2561 (122)
321 - 120	2300 (150)	2309 (129)	49 x 12	2700 (150)	2649 (122)
2 x 49	2400 (150)	2182 (87)	40 x 15	2700 (150)	2595 (129)
3 x 29	2400 (150)	2258 (96)	21 x 60	2700 (150)	2693 (129)
4 x 19	2400 (150)	2194 (105)	39 x 30	2700 (150)	2648 (138)
2 x 39	2400 (150)	2220 (108)	50 x 48	2700 (150)	2650 (138)
2 x 29	2400 (150)	2282 (122)	26 x 48	2700 (150)	2759 (138)
2 x 99	2400 (150)	2418 (130)	16 x 37	2800 (150)	2736 (120)
4 x 97	2400 (150)	2358 (139)	25 x 94	2800 (150)	2808 (123)
3 x 98	2400 (150)	2359 (139)	61 x 45	2800 (150)	2860 (133)
8 x 28	2400 (150)	-	23 x 74	2800 (150)	2842 (139)
4 x 69	2400 (150)	-	95 x 41	2800 (150)	2841 (139)
116 + 194	2500 (150)	2322 (86)	62 x 59	2800 (150)	2840 (139)
354 + 187	2500 (150)	2389 (86)	34 x 85	2800 (150)	-
651 + 186	2500 (150)	2328 (98)	12 x 48	2800 (150)	-
387 + 145	2500 (150)	2366 (104)	95 x 84	2800 (150)	-
136 + 195	2500 (150)	2377 (122)	67 x 51	2800 (150)	-
169 + 152	2500 (150)	2634 (122)	1748 : 23	2900 (150)	2866 (131)
566 + 188	2500 (150)	2381 (123)	3752 : 67	2900 (150)	2816 (126)
641 + 189	2500 (150)	2474 (122)	10184 : 19	2900 (150)	2954 (134)
288 + 457	2500 (150)	2505 (129)	12546 : 51	2900 (150)	2931 (141)
631 + 189	2500 (150)	2504 (130)	16482 : 82	2900 (150)	-
864 - 387	2600 (150)	2539 (90)	1005 : 67	2900 (150)	-
654 - 159	2600 (150)	2414 (98)	3478 : 74	2900 (150)	-
284 - 197	2600 (150)	2438 (99)	2256 : 16	2900 (150)	-
637 - 459	2600 (150)	2578 (102)	12126 : 43	2900 (150)	-
548 - 269	2600 (150)	2598 (107)	11858 : 98	2900 (150)	-

Note. [a] Items that did not change in rating are not displayed.

Table B2

Ratings and Deviations for Each of the Imitation Items

Item	Rating	
	Initial (Dev)	Updated (Dev)[a]
BR	300 (200)	-
BL	300 (200)	-
TL	300 (200)	-
TR	300 (200)	-
BL;BL	300 (200)	-
BLBR	400 (200)	-
TLTR	400 (200)	-
BLTR	500 (200)	-
TLBR	500 (200)	-
TL;BL	600 (200)	-
TR;TL	600 (200)	-
TL;TR	600 (200)	-
BL;BR	600 (200)	-
BR;BR;BR;TL	600 (200)	-
BLBR;TLTR	700 (200)	-
BLBR;BR	700 (200)	-
BL;BL;TLTR	700 (200)	-
BR;TR;TR	700 (200)	-
BLTR;TR	800 (200)	-
TL;TR;TLTR	900 (200)	756 (161)
BR;TL;BL	900 (200)	-
TR;TL;BL	900 (200)	-
BR;TL;TR	900 (200)	-
BR;BL;TR	900 (200)	-
BL;TL;BR	900 (200)	-
BR;BL;TL	900 (200)	-
BL;BL;TR;TR;BR	900 (200)	-
BLBR;TL;BL	1000 (200)	914 (176)
TL;TLTR;BR	1000 (200)	-
TL;BLBR;BR	1000 (200)	934 (176)
BLBR;BR;BLTR	1000 (200)	943 (178)
BLBR;TL;BR	1000 (200)	-
BL;TL;BLBR	1000 (200)	-
BR;BR;TR;TL	1000 (200)	-
BR;BL;BL;TR	1000 (200)	-
BL;BL;TR;TL	1000 (200)	-
BLTR;TL;BR	1100 (200)	939 (159)
TLBR;BR;TR	1100 (200)	1012 (176)
TL;TLTR;TLBR	1100 (200)	-
BLTR;TR;TR	1100 (200)	-

Item	Rating	
	Initial (Dev)	Updated (Dev)[a]
TLBR;TL;BR	1100 (200)	-
BR;TR;BL;TR	1200 (200)	1217 (108)
TL;TLTR;BR;BL	1200 (200)	1170 (121)
TL;BR;BL;TR	1200 (200)	1127 (131)
BR;TL;TR;BL	1200 (200)	1163 (140)
TLBR;BR;BR;TL	1200 (200)	1167 (140)
TR;BR;TR;BL	1200 (200)	1094 (176)
TL;TR;BL;BR	1200 (200)	1104 (175)
BR;TL;TR;BR;BR	1200 (200)	1061 (139)
TL;TLTR;TR;TL	1300 (200)	1292 (131)
BL;TR;TR;BLBR	1300 (200)	1140 (139)
TLTR;TR;BL;BR	1300 (200)	1195 (127)
BL;TR;BLBR;TL	1300 (200)	1438 (93)
TLTR;BR;BL;TR	1300 (200)	1295 (98)
BL;BR;BR;BR;BLBR;BR	1300 (200)	1398 (106)
TL;TR;BR;TLBR	1400 (200)	1280 (112)
TR;BL;TLBR;TL	1400 (200)	1409 (95)
BLTR;TR;TL;BL	1400 (200)	1382 (97)
BLTR;TR;BR;TL	1400 (200)	1457 (113)
BLTR;BR;TL;BL;BL	1400 (200)	1423 (96)
TR;TLBR;BR;BR;TR	1400 (200)	1516 (118)
TL;BR;TR;TL;TR	1500 (200)	1580 (94)
BL;BR;TR;BR;BLBR	1500 (200)	1675 (118)
BL;TR;BR;BR;TR	1500 (200)	1584 (93)
TR;BL;TR;BR;BL	1500 (200)	1646 (110)
TR;TL;BL;BR;TL	1500 (200)	1495 (103)
TL;TL;TL;BLBR;TLBR;BR	1500 (200)	1604 (123)
TLBR;TR;BR;TR;TR	1600 (200)	1633 (124)
TL;TR;TLBR;TL;BLBR	1600 (200)	-
BL;TLTR;BR;TLBR;TL	1600 (200)	1511 (130)
TL;BL;TR;TLTR;BL	1600 (200)	1652 (139)
TL;BR;BL;BLBR;TL	1600 (200)	-
TR;TR;BR;TLTR;BR	1600 (200)	1607 (145)
TL;TLBR;TR;BR;TL	1600 (200)	1498 (98)
BL;TR;BL;BLBR;TL	1600 (200)	1578 (131)
BLTR;BR;TLTR;TL;TL	1700 (200)	1746 (176)
TLTR;BL;TR;BR;BLBR	1700 (200)	1745 (177)
BR;BL;TL;BL;BL;BR	1700 (200)	1775 (139)
TR;BR;BR;TL;BL;TL	1700 (200)	1799 (152)
TL;BL;TR;BR;BL;BL	1700 (200)	1460 (123)
BR;BL;BL;TL;TR;BL	1700 (200)	1595 (105)
TR;BR;BL;BL;TR;BR	1700 (200)	1752 (140)
TLBR;TL;BR;BLTR;TLTR	1800 (200)	1811 (194)

Item	Rating	
	Initial (Dev)	Updated (Dev)[a]
BLTR;TLTR;BL;TL;BL;TR	1800 (200)	-
TLTR;TR;BL;TR;TR;BL	1800 (200)	-
BL;BR;TLBR;BL;BLTR;BLBR	1800 (200)	-
BL;TR;BR;TLTR;BL;TR	1800 (200)	-
BL;BR;TR;BL;TL;BL	1800 (200)	-
TR;TL;BR;BR;TLTR;BL	1800 (200)	-
TR;BR;TR;TL;TR;BL	1800 (200)	-
BL;BR;TR;TL;BLBR;TR	1800 (200)	-
BR;BR;TLTR;BL;TR;BLBR	1800 (200)	-
BR;BL;TLTR;TLBR;BL;TL	1800 (200)	-
BR;TR;BR;TR;BL;TR	1800 (200)	-
BR;TLBR;TL;TR;BR;BLBR	1800 (200)	-
TL;BLBR;TR;TLBR;BL;BLTR	2000 (200)	-
TL;TR;BL;TL;TR;TR;BL	2100 (200)	-
BL;TR;TLBR;BLBR;BR;TLTR;TL	2100 (200)	-
BL;TR;BR;TR;BLTR;TLTR;BL	2100 (200)	-
BR;BLBR;BL;TR;BR;TLBR;BLBR	2100 (200)	-
BR;BL;BLTR;TLTR;BL;BR;TR	2100 (200)	-
BR;TL;TL;TLBR;BL;TL;BL	2100 (200)	-
BR;TLTR;TR;BR;BR;BL;BL	2100 (200)	-
TLTR;TR;TLBR;TL;TL;BL;BL	2100 (200)	-
BLBR;TL;BL;BL;TL;TR;TLBR	2100 (200)	-
TR;TLBR;TR;BLTR;BL;BL;BR	2100 (200)	-
BL;TR;TL;BLBR;TL;BR;TL	2100 (200)	-
BR;BL;BR;TL;BR;TLTR;BL	2100 (200)	-
BR;BL;BR;TR;BR;TLBR;BL	2100 (200)	-
TR;BL;TL;TR;BR;TL;TR	2100 (200)	-
TR;BLTR;TR;TL;BR;BL;TLBR	2100 (200)	-
TL;TR;BL;TR;BL;TR;BL	2100 (200)	-
TR;TL;TR;TL;TR;BR;TL	2100 (200)	-
BLTR;BL;TLBR;BR;TL;BL;TLTR	2100 (200)	-
BL;TL;BL;TL;BR;TR;TR	2100 (200)	-
BR;TLTR;TR;BR;TLBR;TL;BL	2100 (200)	-
TR;TL;BR;TR;TL;BLTR;TR;BR	2400 (200)	-
TL;TR;BL;BR;TLBR;BR;BLBR;BLTR	2400 (200)	-
TR;TL;BL;TL;BLTR;BL;BR;BL	2400 (200)	-
BR;BLBR;TL;BL;TR;TL;BL;TLBR	2400 (200)	-
BR;TL;BR;TR;TLBR;BL;TR;BLBR	2400 (200)	-
TL;BL;BL;BR;BLTR;TLBR;TR;TLTR	2400 (200)	-
TR;BL;TR;BLTR;BL;TR;TLTR;BR	2400 (200)	-
TL;BL;TLBR;BL;BLBR;BR;TLTR;TR	2400 (200)	-
TL;BR;TR;BL;BR;TLBR;BL;BLTR	2400 (200)	-
BL;TR;TLBR;TR;BLTR;BR;TL;BR	2400 (200)	-

Item	Rating	
	Initial (Dev)	Updated (Dev)[a]
BR;BL;BL;BR;TR;TL;TR;TLTR	2400 (200)	-
BR;TR;BR;TL;BL;TR;TLTR;BL	2400 (200)	-
TR;BL;BR;BR;TL;TR;TL;TLBR	2400 (200)	-
TR;BR;BR;TLBR;TL;TR;TL;BR	2400 (200)	-
BL;TL;TLTR;TLBR;BR;TL;BL;BL	2400 (200)	-
BL;TR;BLBR;BR;BR;BR;TR;TL	2400 (200)	-
TL;BR;TL;TL;TR;TR;BR;BL	2400 (200)	-
BR;BL;TL;TR;BL;TR;BL;BL	2400 (200)	-
BR;TR;BL;TR;TL;TL;TL;BLTR	2400 (200)	-
BR;TL;TLBR;TR;TL;BR;BLBR;TL	2400 (200)	-
TL;BL;BL;TR;BLTR;TR;TR;BR;BL	2600 (200)	-
BL;BL;TL;BR;BR;TR;BR;TR;BR	2600 (200)	-
BLBR;TLBR;TL;TL;BR;BL;BL;BR;BR	2600 (200)	-
TLTR;TR;TR;TL;BLTR;TL;BR;BLTR;TR	2600 (200)	-
BR;TR;TL;BL;TL;TR;BR;BR;BL	2600 (200)	-
BLTR;TL;BLBR;TL;TL;BLTR;TR;BR;BL	2600 (200)	-
BR;TR;TL;BR;TR;TL;TL;BL;BL	2600 (200)	-
BLTR;BL;TR;TL;BL;BR;TR;BR;BLTR	2600 (200)	-
TL;TL;TLBR;TR;TR;BR;BR;TL;TR;TL	2700(200)	-
BLBR;BL;TR;TLBR;TL;BL;BL;BL;TR;BR	2700(200)	-
TR;BLBR;BL;BLTR;BLBR;BL;TL;BL;BR;TR	2700(200)	-
TL;TR;TR;BLBR;BL;BL;TR;TR;TL;BR	2700(200)	-
TL;BR;TR;TL;BL;TLBR;BLBR;BR;TR;TR;TLTR	2800 (200)	-
TL;BL;TL;TL;TR;BLTR;BR;BL;TLBR;TLTR;TR	2800 (200)	-

Note. TR = top right; TL = top left; BR = bottom right; BL = bottom left; TLTR = both top left and top right; BLBR = both bottom left and bottom right; TLBR = both top left and bottom right; BLTR = both bottom left and top right.

[a]Items that did not change in rating are not displayed.

Appendix C

Protocol

Protocol

for the experiment "Motivating children to play games with a robot, using Bayesian reasoning"

B. R. Schadenberg

April 2012

1. Required equipment

- 1 NAO robot + adapter + protective foam.
- 1 Laptop for the experimenter + adapter + silent mouse + extra AA-batteries.
- 1 Laptop for the participant (which is also used to display the mathematical assignments) + adapter + regular mouse.
- 1 Monitor with VGA connector, which can be connected to the participant-laptop.
- 1 Monitor + A/V connector cable + BNC cable.
- 1 Router + adapter.
- 1 Video camera + adapter + tripod.
- 1 Photo camera.
- 2 Extension sockets.
- 3 UTP-cable.
- Printed out questionnaire.
- Printed out participants list.
- Printed out observation form.
- 5 Comics.
- 3 Pens.
- USB-stick

2. Preparation

2.1 Introduction experimenters

Prior to the experiment, the experimenters should introduce themselves to the participants and tell them what they can expect.

- Start off with introducing yourself, and then introduce the robot.
- Explain for what company or institution the experiment is being carried out.
- Tell something about the ALIZ-E project. What is it about and why is important.
- Tell the participants that the experiment is to test the robot, and not to test the participant's skills.
- Explain what the participants can expect:
    - How long is the experiment going to take.
    - What the participants will be doing.
    - That after playing two games, the experiment is done and they are free to choose between the different options.
    - That there will be two sessions.
- Explain that only researchers of the ALIZ-E project may view the videos and that the videos will not be distributed.
- Explain that the participants can be completely honest with the robot and the experimenters.

2.2 Setup

- Put the NAO on a table, facing the spot where the participant will be sitting/standing. And set the extra screen on the left hand of the NAO.
- Choose a different table that can be used by the experimenters. Make sure that the participant cannot see the laptop and monitor used by the experimenters. Place the protective foam between the participant and the experimenters, so that the experimenters are hidden from view (as much as possible) for the participant.
- Place the camera (mounted on a tripod) so that both the robot and the participant can be seen. Choose a spot where the camera is less likely to attract attention.
- Place the comics, the participant laptop and a pen near the participant, but without them drawing the attention of the participant.
- Place the router somewhere out of sight of the participant.
- Make sure the questionnaire, observation form and participant list are at hand.

2.3 Startup procedure

1) **Turn on all devices:**

   Turn on both laptops (1 for running the WoOZ, 1 for running the extra screen on for the math game),

   turn on the NAO (and possibly connect the NAO to the local power grid),

   turn on the router.

2) **Open all programs:**

   Experimenter laptop:

   Connect to the router.

   Open Microsoft Visual C# 2010 and load the WoOZ solution (WoOzExpress.sln),

   open Eclipse and load the Rating system project,

   open GOAL and open the files “GameMoveGenerator.mas2g” (the agent for the experimental condition) and “ControleGroep.mas2g” (the agent for the control condition),

   open the stopwatch program.

   Participant laptop:

   Connect to the router

   Start the extra screen by opening ExtraScreen.exe.

   Open the internet browser and preload the alternative games.

   Close the screen as much as possible, without the laptop going into sleep-mode.

3) **Set IP-addresses:**

   Check the NAO’s IP address by pressing it’s chest button once.

   Set the NAO’s IP-address in the WoOZ. Go to FrmNAOConnector -> App.config -> line 9.

   Check the “extra screen” laptop’s IP (press start, run: “cmd”, type: “ipconfig”).

   Set laptop’s IP address in the WoOz. Go to FrmMathGame -> App.config -> line 9.

4) **Run WoOZ & Rating system:**

   Run the WoOz by pressing “Start debugging” or F5.

   run the rating system.

   Load the following modules: “Game Manager”, “DialogComposer”, “GoalConnector”, “NAOClient”, “RatingConnector”, “UserModel”, “Imitation” and “Math”.

5) **Make the connection:**

   Press “Connect All” in the Game Manager.

6) **Run the GOAL agent:**

Select either of the GOAL agents, depending on what condition the participant is in. Launch the agent program in GOAL (click on run -> run, or the green play button), and press run again to unpause the agent.

3. Experiment procedure

3.1 Timetable of the experiment

0:00 – 2:00 The NAO introduces itself, the experiment and the first game.

2:00 – 7:00 Play the first game for ~5 minutes.

7:00 – 8:00 The NAO explains the second game.

8:00 – 13:00 Play the second game for ~5 minutes.

13:00 – 14:00 The NAO explains the free choice period.

14:00 – 19:00 Play the free choice period.

19:00 – 20:00 The NAO explains this is the end of the experiment and says goodbye.

20:00 – 23:00 Ask the participant to fill in the questionnaire.

23:00 – 25:00 Thank the participant and make a picture of the participant with the NAO.

3.2 The procedure

Pre-experiment

- Write down the participant's number on all forms.
- Check what condition the participant is in.

Experiment

1) Start recording with the video camera and start the stopwatch.
2) Explain to the participant:
    - That the experiment is to test the robot, not the participant's skill.
    - That the robot may sometimes show erratic movements, which they are to ignore.
    - That the experimenters are there only in case the robot fails and that they can act as if the experimenters are not in the room.
    - That the participant can decide at any point to stop participating in the experiment.
    - That the experiment is over after playing the two games, but that they could freely choose what activity to engage in, while the experimenters checked the data.
3) Ask whether or not the participant is ready to start. If so, continue.
4) Load the "Exp Bob.dialog" file in the dialog manager, and start the introduction
5) During the introduction fill in the user's name and id number in the User model.
6) If this is the first time for the user, then click on the tab "math" in the User model, to set the initial rating and initial deviation (for this experiment 2100 rating and 200 deviation). And do the same for the imitation game (1500 rating, 200 deviation is standard).

First game:

7) Introduce either of the games, using the dialog manager.
8) Select a game and click the “Set current game” button in the Game manager.
9) Open the game and press “select item”. And, in the case of the imitation game, press execute poselist.
10) When the child is done, click whether the answer was correct or incorrect.
11) Rinse and repeat until the time is up for that game.

Second game:

12) Let the NAO say that’s it’s time for another game (press the “next game” button) and introduce the next game using the dialog manager.
13) Select items just like in the first game.

Free choice period:

14) Let the NAO explain that the participant can continue playing games with the NAO, or read a comic / play a game on the laptop, using the dialog manager.
15) If the participant wants to continue playing a game with the NAO, select that game and play it until the time is up. Else, the experimenter hands the comics over to the participant, so he or she can make a choice on which comic to read. If the participant wanted to play a game on the laptop, open the laptop and show the games that can be played. Make sure the laptop is connected to the internet, instead of to the router used for the experiment.

End of the experiment:

16) Let the NAO say that this is the end of the experiment and say goodbye.
17) The experimenter asks if the participant could fill in the questionnaire. After doing so, the experimenter sits down again and waits for the participant to finish.
18) When the participant has filled in the questionnaire, check it to see if there are any missing values. If so, ask the participant to fill in the missing values. If not, thank the participant for participating in the experiment.

Post-experiment

1) Stop recording with the video camera and reset the stopwatch.
2) Exit the WoOZ by pressing the “exit” button in the Game Manger, or press stop in Windows 2010 Express.
3) Backup the log files in the /RatingSystem folder, the /RunWoOZ/UserModels folder, the /RunWoOZ/Memory folder and the /RunWoOZ/WoOzLog folder. Also, save the log files on a USB-stick.
4) Restart all programs.

4. Troubleshooting

Q: I am pressing the select item button, but no item is being selected.

A: Is the rating system running? Is the GOAL agent running? Did you update the user model with ratings and deviations higher than zero? Did you select a game in the Game Manager?

Q: I am getting the "Prolog.err" error. What am I doing wrong?

A: You have to wait until the NAO has finished speaking, before you can send another speech request to the NAO. The NAO is finished when you receive the message "FeedbackTalkFinished".

Q: The NAO is not performing anymore movements of the imitation game. Now what?

A: This is a bug. You have to restart the WoOZ (and GOAL).

Q: The NAO is making erratic movements during the imitation game. Can I prevent it from doing so?

A: No, this is also a bug in this version of the imitation game.

Q: The math calculation is not being shown on the extra screen. Now what?

A: Try reconnecting with the extra screen (press stop and then start, in the math screen). Also make sure that the laptop with which the extra screen is connected, is connected to the router and it's ip address is the same IP address as in the Math games' "App.config" file.

Q: The NAO is saying "name", rather than the user's name. What am I doing wrong?

A: Update the user model, including the user's name. The NAO will now say the user's name, rather than "name".

Q: GOAL keeps giving me advice to switch games. How can I continue playing?

A: Select (another) game and press "Set current game" in the Game Manager.

Q: The WoOZ cannot find the NAO. Now what?

A: Check if the IP-adress in the FrmNaoConnector's "App.config" contains the correct IP-adress. If so and it is still not connecting, try using a UTP cable rather than wireless.

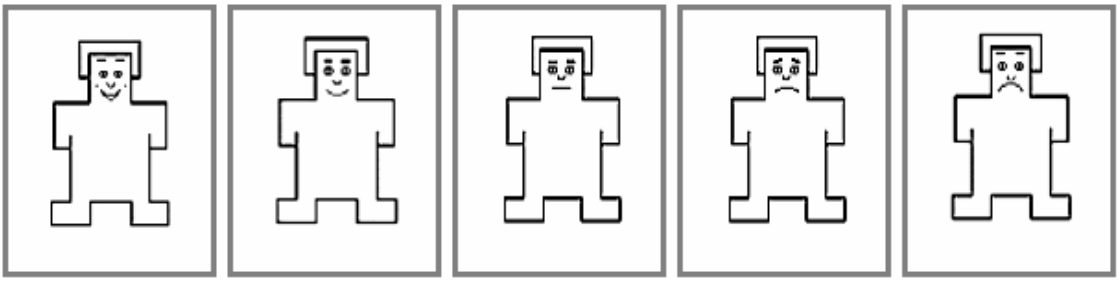
Appendix D

Questionnaire

Vragen over het onderzoek

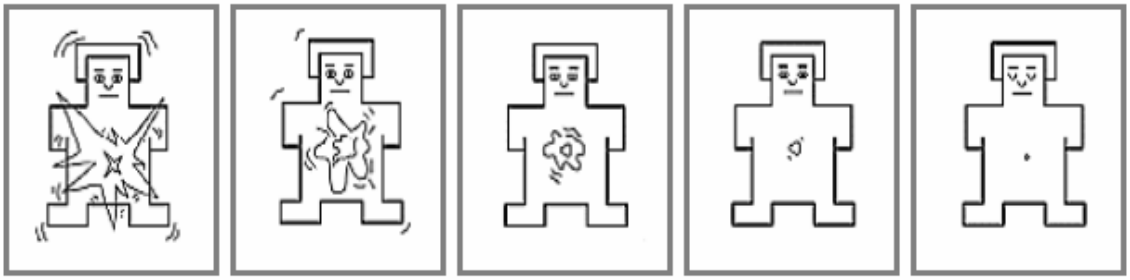
De eerste twee vragen gaan over hoe **jij** je voelt.

Op de eerste rij lopen de figuurtjes van HEEL VROLIJK tot HEEL DROEVIG.
1) Kruis het figuurtje aan wat het beste past bij hoe je je voelt:



O O O O O

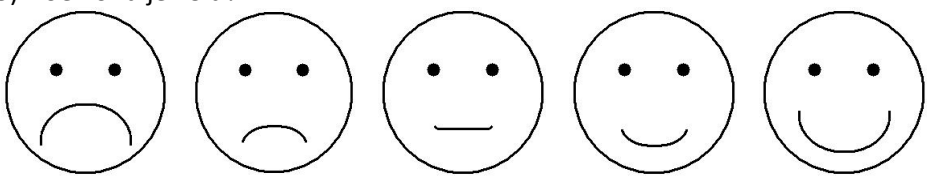
Op de volgende rij lopen de figuurtjes van HEEL OPGEWONDEN tot HEEL KALM.
2) Kruis het figuurtjes aan wat het beste past bij hoe je je voelt:



O O O O O

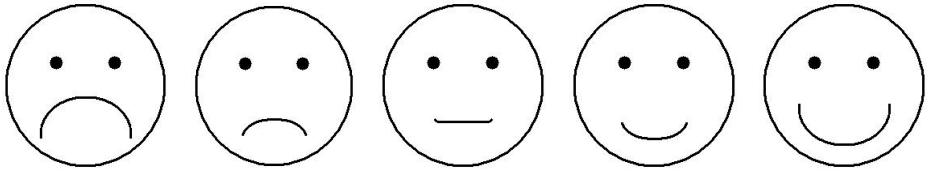
De volgende vragen gaan over **Lola**. Kruis bij deze vragen aan wat het best bij je past.

3) Hoe vond je **Lola**?



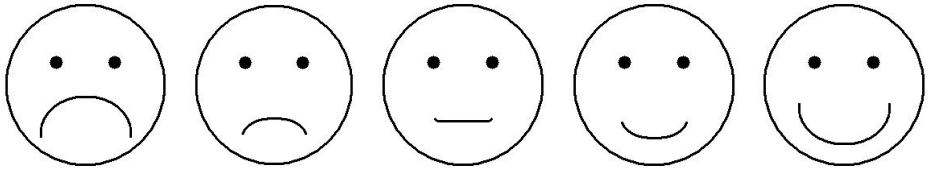
vreselijk niet zo leuk leuk heel leuk geweldig
O O O O O

4) Hoe vond je het **rekenspel**?



vreselijk	niet zo leuk	leuk	heel leuk	geweldig
O	O	O	O	O

5) Hoe vond je het **imitatiespel**?



vreselijk	niet zo leuk	leuk	heel leuk	geweldig
O	O	O	O	O

6) Welk spel vond je het leukst?

O Rekenspel
O Imitatiespel

7) Hoe vond je het **rekenspel**?

Erg makkelijk	makkelijk	Niet moeilijk, maar ook niet makkelijk	moeilijk	erg moeilijk
O	O	O	O	O

8) Hoe vond je het **imitatiespel**?

Erg makkelijk	makkelijk	Niet moeilijk, maar ook niet makkelijk	moeilijk	erg moeilijk
O	O	O	O	O

9) Wil je nog iets anders zeggen over Lola of de spelletjes? (Het hoeft niet)

……

……

……

……