Personalising game difficulty to keep children motivated to play with a social robot

Schadenberg, Bob; Neerincx, Mark A.; Cnossen, Fokeltje; Looije, Rosemarijn

Published in:
Cognitive Systems Research

DOI:
10.1016/j.cogsys.2016.08.003

Publication date:
2017

Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research 43 (2017) 222–231

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

Personalising game difficulty to keep children motivated to play with a social robot: A Bayesian approach

B.R. Schadenberg[a,b,*], M.A. Neerincx[a,c], F. Cnossen[b], R. Looije[a]

[a] Perceptual and Cognitive Systems, TNO, Postbus 23, 3769 ZG Soesterberg, The Netherlands
[b] Artificial Intelligence and Cognitive Engineering, University of Groningen, Nijenborgh 9, 9747 AG Groningen, The Netherlands
[c] Interactive Intelligence, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands

Received 31 March 2016; received in revised form 9 July 2016; accepted 16 August 2016
Available online 22 August 2016

Abstract

For effective child education, playing games with a social robot should be motivating for a longer period of time. One aspect that can affect the motivation of a child is the difficulty of a game. The game should be perceived as challenging, while at the same time, the child should be confident to meet the challenge. We designed a user modelling module that adapts the difficulty of a game to the child's skill level, in order to provide children with the optimal challenge. This module applies a Bayesian rating method that estimates the child's skill and game item's difficulty levels to personalise the game progress. In an experiment with 22 children (aged between 10 and 12 years old), we tested whether the personalisation leads to a higher motivation to play with the robot. Although the personalised system did not challenge the participants optimally, this study shows that the Bayesian rating system is in principle able to measure the skill and performance of children in playing a game with a robot (even without accurate estimates of the difficulty of items). We outline multiple ways in which the rating method and module can be used to further personalise and enhance the child-robot interaction, other than adapting the difficulty of games (e.g. by adapting the dialogue and feedback).

Keywords: Social robotics; User modeling; Rating system; Child-robot interaction; Motivation

1. Introduction

For children, playing educational games with a social robot can be a fun way to learn, to take their mind of their current situation, or simply for the sake of enjoyment. Within the ALIZ-E and PAL projects, educational games are used to teach children with diabetes how to manage their chronic illness. The children need to be able to calculate how much insulin they need to inject based on their food intake and the physical exercise. To this end, a math game and imitation game were developed for the social robot. With these two games the robot can provide the children with a fun way of learning how to calculate the insulin dosage and learn about the relation between physical exercise and food intake in regard to the insulin dosage. However, keeping children motivated to interact with a social robot can be difficult after the initial novelty has worn off (Gockley et al., 2005; Kanda, Hirano, Eaton, & Ishiguro, 2004; Leite, Martinho, & Paiva, 2013). One factor that can influence the motivation is the perceived difficulty (Csikszentmihalyi, 1990; Deci & Ryan, 1985). 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

[*] Corresponding author at: Human Media Interaction, University of Twente, The Netherlands.
E-mail address: b.r.schadenberg@utwente.nl (B.R. Schadenberg).

http://dx.doi.org/10.1016/j.cogsys.2016.08.003
1389-0417/© 2016 Elsevier B.V. All rights reserved.

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. To facilitate intrinsic motivation, a child should perceive the content of the game as challenging, but within his or her capability (Csikszentmihalyi, 1990). If the content is perceived as too easy, the child will become bored. And if the content is perceived as too challenging, the child can become discouraged and give up. Furthermore, according to Vygotsky's Zone of Proximal Development (Vygotsky, 1978), providing challenging content for children to learn, whilst providing guidance, will encourage learning.

Because the skill amongst children in any given group may vary significantly, there is a need to adapt the difficulty of the game to a personal level in order to keep children motivated (Janssen, van der Wal, Neerincx, & Looije, 2011). While adults can quickly gauge how skilled a child is in playing a game, a social robot will require a computer model to assess the skill of a child. Such a model has to be accurate quick, reliable, be applicable to different kinds of games, and should work without negatively influencing the interaction, for example by asking a series of skill-related questions. In this paper we discuss such a (Bayesian) model and present the results of a study using this model to adapt the difficulty of the two games to a personal level, in order to keep children motivated to interact with the social robot. The main question of this study is to determine the effectiveness of a Bayesian model to estimate a child's skill for playing a game with a robot and personalizing the game play to keep children motivated, whilst using a limited amount of data to calibrate the model.

2. Related work

Educational robots are robotic systems that can support children in a learning task by serving as tutor (Kennedy, Baxter, & Belpaeme, 2015; Saerbeck, Schut, Bartneck, & Janse, 2010), teaching assistant (Chang, Lee, Chao, Wang, & Chen, 2010), or as a peer (Kanda et al., 2004). Using the robot as a tutor or peer can be especially beneficial, as it can provide children with one-on-one tutoring which can increase learning gains (Bloom, 1984). For the robot to be effective in supporting children in learning a task, it will need to personalise its social behaviour and the learning content to the user. For example, social behaviours that can influence learning include the use of gestures (Szafir & Mutlu, 2012), socially supportive behaviours (Saerbeck et al., 2010), or personalised language (Henkemans et al., 2013). Personalising the tutoring strategies based on a user's skill has been shown to improve learning gains (Leyzberg, Spaulding, & Scassellati, 2014). In their study, Leyzberg et al. used two algorithms to model the user's skill: a simple additive model and a Bayesian network. The former model is susceptible to local maxima and minima, while the latter categorises a skill to be either learned or not learned. Gordon and Breazeal (2015) used a social robot to teach children word-reading skills. Words were selected by a specially designed Bayesian Active Learning model based on which word would lead to the largest gain in knowledge. The child's motivation was not taken into account.

In the related field of computer-based learning, educational computer programs called Intelligent Tutoring Systems (ITS) are developed to give individualised lessons and have shown to be effective tutors (VanLehn, 2011). ITS use student modelling techniques to measure and represent user characteristics (Polson & Richardson, 2013). Estimates of a user's skill can be obtained via models from the item response theory (IRT). This theory contains a number of statistical models which relate a user's response to items to a latent trait of the user (Lord, Novick, & Birnbaum, 1968). Typically, these models were developed with the assumption that skill does not change over time, which is a reasonable assumption for skills that are learned very slowly, or for short tests. In our context, the user's skill may change quickly, depending on the game, and is not an assumption we can make. More specialised models (for an overview see Chrysafiadi & Virvou (2013) and Desmarais & de Baker (2012)) like Bayesian Knowledge Tracing (Corbett & Anderson, 1994), do account for learning. These models are mainly used to model fine-grained tasks (e.g. specific operations to solve a complex equation). The implementation of these models is a time-consuming task, often requiring large samples of data for calibration, complex parameter fitting, or expert knowledge regarding the domain of the application, making these models not flexible and expensive to implement (Pelánek, 2016).

A 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. Rating systems are used in sports, such as football (Hvattum & Arntzen, 2010) or computer games, to pair players of equal skill to play with or against each other, resulting in a win, loss, or a draw. They are also used in an educational setting to match a user with an item, such as a mathematical assignment, of a certain difficulty (Klinkenberg, Straatemeier, & van der Maas, 2011).

There are three major benefits of using a rating system for a social robotic platform in an educational setting. First, rating systems are relatively easy to implement and adapt for different applications, and do not require input from experts on the implementation or on domain knowledge. Second, rating systems are still capable of achieving a high accuracy (Glickman, 1999; Klinkenberg et al., 2011), and therefore should be able to adapt the difficulty of the game in order for the children to answer 70% of the items correctly. And last, rating systems are non-domain specific and thus can be applied to any skill-based application.

Contemporary rating systems are based on the Elo rating system (Elo, 1978), which was designed to pair chess players to compete against each other. The Elo rating system is a paired-comparison model, closely related to the Rasch Model (Rasch, 1960) used in IRT, and works as follows. All users start out with a certain numerical rating θ_u, which represent the estimated skill level of user u. A similar rating θ_i is assigned to each item i and represents the level of difficulty. The higher the rating, the more skilled the user/more difficult the item is. When the initial ratings are set, the user is paired with an item, based on a selection algorithm that uses the ratings (e.g. minimising the difference between the user's rating and item's rating). When an item is completed by the user, the rating of both the user and item will be updated, based on the outcome of the instance. The user rating θ_u and item rating θ_i are updated as follows:

$$\theta_u = \theta_u + K(s - P(s = 1)) \tag{1}$$

$$\theta_i = \theta_i + K(P(s = 1) - s) \tag{2}$$

where K is a constant that governs how much a rating can change in one instance, s is the outcome of the instance – i.e., 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 $P(s = 1)$ is the probability of a correct outcome. For the item, a correct outcome is the user answering incorrectly (the item "wins" in this case). The expected probability for the user to answer correctly can be calculated using the following equation:

$$P(s = 1) = \frac{1}{1 + 10^{-(\theta_u - \theta_i/400)}} \tag{3}$$

The original Elo rating system was designed for chess and uses a specific rescaling of the standard logistic function, using base 10 instead of e and the constant 400. The same equation can be applied to calculating the expected probability for the item, when θ_u is substituted by θ_i and vice versa.

When the discrepancy between the user's rating and the item's rating is small $\theta_u \approx \theta_i$, the probability of the user answering correctly will be close to .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.

The estimated probability is taken into account by the rating update equation. 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 are against the user. As a result, the user will be rewarded with a greater increase in rating, when giving the correct answer. Also, the decrease in rating is 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 probability 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 and will generally become more reliable the more instances occur.

For example, user A starts out with a rating of 1500 and is going to answer an item with a difficulty of 1600. It is estimated that the user has a 36% chance to answer correctly. If the user answers the item correctly, the user's rating will increase by $K \times 0.64$, and the item rating will decrease by as much. When answered incorrectly, the user rating will decrease, and the item rating will increase, by $K \times 0.36$.

The Glicko rating system (Glickman, 1999) 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. 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. As a result, the Glicko rating system will approximate the true rating much quicker than the Elo rating system.

A disadvantage of using a rating system is that they are designed to provide users with a probability of 50% of answering correctly. Answering about 50% of the items correctly is often experienced as discouraging, as the game will be perceived as being too difficult. Based on other studies (Eggen & Verschoor, 2006; Klinkenberg et al., 2011), it can be concluded that a child user will be optimally challenged when they give the correct answer approximately 70% of the time. But in order to increase the percentage of correct answers from 50% to 70%, a high measurement precision is required (Eggen & Verschoor, 2006). In order for the rating system to optimally challenge the user, it can select items based on the current percentage that the user answered correctly. For example, when the user answered around 50% correctly, the rating system can select items that the user is more likely to answer correctly, and therefore increase the percentage of correct answers.

Alternatively, using a social robot as a platform grants the opportunity to capitalise on the properties of social robots, like our tendency to project social qualities to the behaviour of technology (Reeves & Nass, 1996) and to view social robots as social communication partners

(Powers, Kiesler, Fussell, & Torrey, 2007). Rather than increasing the motivation of the child, by increasing the percentage of correct answers, it may be better to utilise the social qualities attributed to the robot and make the user compete with the robot; the one with the most correct answers wins. While the child will answer only 50% of the items correctly on average, he or she can still win the game by outperforming the robot. This way, the optimal challenge is dependent on the skill of the robot, instead of the difficulty of the items.

Rating systems work best when a large amount of data is available to calibrate the item ratings. However, in practice it might not be feasible to gather enough data to accurately estimate the difficulty of the items. This leads to the question whether using a rating system in such cases is still viable for adapting the difficulty of a game. In this study we will explore the use of a rating system by implementing such a system in a social robot, using a limited amount of data to initialise the item ratings. Furthermore, we will explore whether it is possible to provide children with the optimal challenge by keeping the percentage of correctly answered items around 70% in order to keep the children motivated.

In our experiment, children will play two different games with the robot, namely a math game and an imitation game. We address the following research questions:

- How accurate is the Bayesian rating system in estimating the chance of the child answering correctly?
- To what extent does providing children with the optimal challenge affect the child's motivation to interact a social robot?
- How many items does a child need to answer before the user rating stabilises?

3. Implementation

For this study, we developed a module which is an implementation of the Glicko rating system and modified a math game and an imitation game to the use of this rating system. For the math game, the child has to solve arithmetic assignments, varying in complexity and operation. The imitation game involves memorising a sequence of arm movements and reproducing those movements at the end of the sequence.

We designed a GOAL agent (Hindriks, 2009) to model the decision making of the robot. The GOAL agent can take a number of actions, and bases its decisions on the goals and beliefs it holds. In our case, the goal of the GOAL agent was too keep the percentage of correctly answered items around 70%. At any point in time during the experiment, the GOAL agent would have a belief which reflects the current percentage of the items answered correctly by the participant for each game. In order to achieve its goal, the GOAL agent could take three different actions when it was tasked with selecting a new item. It could select either an easy, moderate, or difficult item, depending on its beliefs. An easy item is an item that on average will be answered correctly 70% of the time and is selected when the child answered less than 70% correct. A difficult item is selected when the child 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 child answered between 70% and 80% of the items correctly.

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

$$\prod_{m=1} P(s = 1 \,|\, \theta_i, \theta_j, RD_j)_m < .10 \tag{4}$$

where $P(s = 1 \,|\, \theta_i, \theta_j, RD_j)$ is the expected probability of a correct answer given the estimated user rating, the difficulty of the item, and the rating deviation of the item. Each time the child 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 correctly each item in the sequence. The cumulative probability was reset when the child answered incorrectly. When the cumulative probability was smaller than .10, the GOAL agent responded by complimenting the child on doing well. Eq. (4) was also used to estimate exceptionally poor performance, by storing the probability of an incorrect answer each time the answer was incorrect. When the cumulative probability was smaller than .05, the GOAL agent would change the game, as it was assumed the child was not motivated to play the current game.

For the math game, the initial item rating of each of the assignments was set using the levels of difficulty used in the study of Janssen and colleagues (Janssen et al., 2011). The levels of difficulty were based on two instruction books (Borghouts et al., 2005; Goffree & Oonk, 2004) 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. All the assignments were given an initial rating deviation of 150. The initial item ratings of the imitation game were set based on the length of the sequence and modified by the complexity of the movement(s), and the presence of similar subsequent movements. The initial user ratings were set based on the teacher's opinion on the skill of each of the participants on math. Concerning the imitation game, the initial user ratings were set to the middle of the scale. The rating deviations were set at 350 for both games.

4. Materials and methods

4.1. Participants

22 Dutch children (14 male and 8 female, age 10–12 years) from the elementary school ‘Griftschool’ (Woudenberg, the Netherlands) participated in the experiment. None of the participants had diabetes. The participants were randomly divided into two groups of 11 participants, balancing for gender. In return for participating, the school received toys (K’NEX) as a gift, a lecture on social robotics for the participating class, and the participants received a picture of themselves and the robot.

4.2. Experimental design

For the experiment, a between-subject design was used. For the experimental group, the percentage of correct answers was regulated to be between 60% and 80%, by asking items with an easy, moderate or hard difficulty, with respectively a chance of correctly answering the item of 70%, 50%, and 30%. Furthermore, the robot reacted when the performance of the participant was exceptionally poor or exceptionally well. 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, which were a week apart. The first session served three purposes, namely to get a reliable estimate of the user rating and to make the item ratings more reliable, to reduce the initial enthusiasm of interacting with a robot for the first time, and to ensure the participants know how to play the games.

4.3. Procedure

The experiment took place in one of the offices at the school. Because only one room was available, the experimenters had to be in the same room as the participants. The robot was placed on a desk, as can be seen in Fig. 1. The participant would sit or stand in front of the robot and the experimenters would sit behind the participant. In order to minimise the presence of the experimenters, a covering screen was placed between them and the participant. The experimenters observed the child via a live feed from the camera which was placed next to the desk.

Prior to the experiment, the experimenters introduced themselves and the robot to the participants during class. The experiment consisted of two sessions with identical procedures, and started with the experimenter explaining the course of the session to the participant. Next, the robot greeted the participant and would explain the first game. The participant then played both games for five minutes. The robot would ask the participant to provide an answer to one of the items (selected by the GOAL agent). When the child answered, the robot would say whether the answer was correct or incorrect. When the five-minute mark was reached for the first game, no new items were asked and the robot waited for the participant to answer the last 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 freely choose the next activity. At this point, the experimenter would tell the participant that the experiment was over, and that he or she could engage in an activity of their own choosing, while the experimenters checked to see if the data was in order. The participant could freely choose what to do. In order to provide the children with alternative activities which might interest them, we provided some options (cf. section 4.4.3 for the alternative activities). But they were free to engage in a different activity of their choosing. After five minutes, the robot would ask if the participant liked playing with the robot and say goodbye. Finally, the participant had to fill in a questionnaire that measures the participant’s pleasure and arousal (cf. section 4.5) and could then return to class. Overall, a session lasted approximately twenty minutes. Following the experiment, the experimenters gave a lecture on social robotics for the participants. This lecture also served as a debriefing.

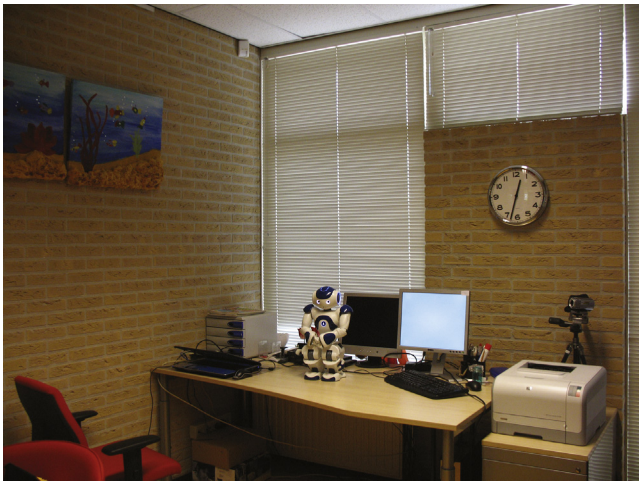

Fig. 1. The experimental set-up.

4.4. Materials

4.4.1. NAO Robot

We used SoftBank Robotics’ NAO robot, starting out with a blue NAO (v28), using Acapela Text-To-Speech (v7.0, using the mature woman’s voice Femke22Enhanced) to convert text to speech. It malfunctioned after having been used by four participants, thus for the remainder of the experiment the NAO v32 was used. This NAO was coloured red, and used Fluency Text-To-Speech speech editor professional (v4.0, using the child-like voice Fiona). Both robots were provided with the name Lola, because the first NAO was speaking with a mature woman’s voice.

4.4.2. Software

A Wizard of Oz set-up was used, preventing issues related to the quality of speech and motion recognition software. The experimenter interpreted the participant's movements and speech, and conveyed their response to the robot via the Wizard of Oz interface. The robot would then decide on how to respond. The participant was unaware of this, and from the outside it looked like as if the robot is fully autonomous.

4.4.3. Alternative activities

For the free-choice period, we provided some activities which the children could engage in. These included continuing playing with the robot, reading a comic or playing a game on the laptop. The children could choose from five comics (including two Donald Ducks, one Dirk Jan, one Asterix, and one Kid Paddle), or play Bubbles or Bejeweled on the laptop. All the comics and both games are popular and well known amongst 10 year old children.

4.5. Measures

The user and item ratings were stored after an item had been answered by a participant. The user ratings are taken as an estimate of the participant's skill, and the item ratings are an estimate of the difficulty of the item.

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) and Janssen et al. (2011)). 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 the participant spent playing with the robot was measured and functioned as a measure for the intrinsic motivation of the participant to interact with the robot.

After the free-choice period, the participant had to fill in a questionnaire. The Self-Assessment Manikin (Bradley & Lang, 1994) was used to measure the participant's pleasure and arousal. 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), each represented by a smiley. Finally, the participants had to indicate which game they enjoyed the most, and how difficult they considered the games.

4.6. Analysis

To assess the performance of the rating system, we calculated the difference between the actual percentage of correct answers and the percentage of correct answers predicted by the rating system. The item ratings were grouped for every 100 points by rounding them to hundreds. Additionally, for the math game, the item ratings are evaluated by comparing the response time for the different item ratings. We expected that the more difficult an item is, the more time a child will need to answer. For the imitation game, response times cannot be used to evaluate the initial item ratings, as more difficult sequences generally contained more movements and thus took longer to complete.

We analyse the user ratings to estimate whether the user rating approximated the child's true rating. This will be the case when the user rating fluctuates around a certain rating.

5. Results

5.1. Manipulation check

To check whether the GOAL agent was able to regulate the difference in percentage of correctly answered items between the experimental and control condition, we checked whether the means were significantly different in the two conditions for the second session. We found no difference in the percentage of correct answers on both the math ($F(1, 19) = 0.063$, $p = .805$) and imitation game ($F(1, 19) = 0.12$, $p = .912$) between the control and experimental condition (see Table 1). For the participants in the experimental condition, the GOAL agent could select easy and hard items, in addition to items with a moderate difficulty. In case of the math game, 41% of the items consisted of either easy or hard items. For the imitation game, this percentage was 52%.

5.2. User ratings

The relative change in user rating over time is a measure of the reliability of the user rating. Fig. 2 shows the user ratings on the math game during the first and second session. On average, the user rating increased by 447 during the first session, which can be attributed to the large increase during approximately the first ten items. Such a trend did not occur during the second session, where the user rating changed by 29 on average. For about half of the participants (e.g. participant number 6 and 19), the user ratings were relatively stable from the beginning of the second session. For the other half of the participants, the user rating showed significant increases or decreases.

The development of the user ratings on the imitation game can be seen in Fig. 3. Answering an item of the

Table 1
Percentage of correct answers for both conditions and games.

Game	Control ($n = 11$)		Experimental ($n = 11$)	
	M (%)	SD (%)	M (%)	SD (%)
Math Session 1	77	9	72	7
Math Session 2	64	20	65	15
Imitation Session 1	40	15	47	19
Imitation Session 2	52	13	53	12

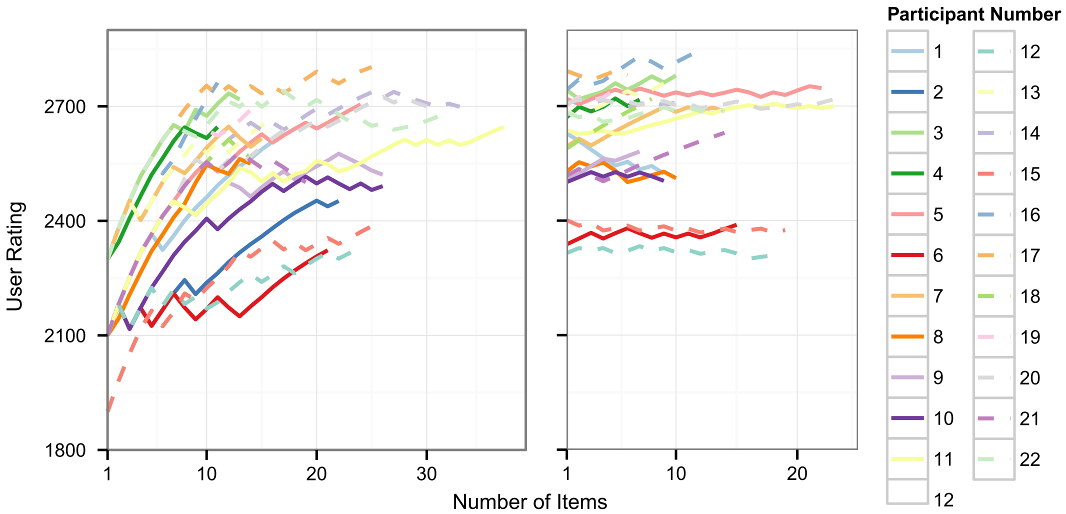

Fig. 2. The user ratings on the math game per participant, for session 1 (left) and session 2 (right).

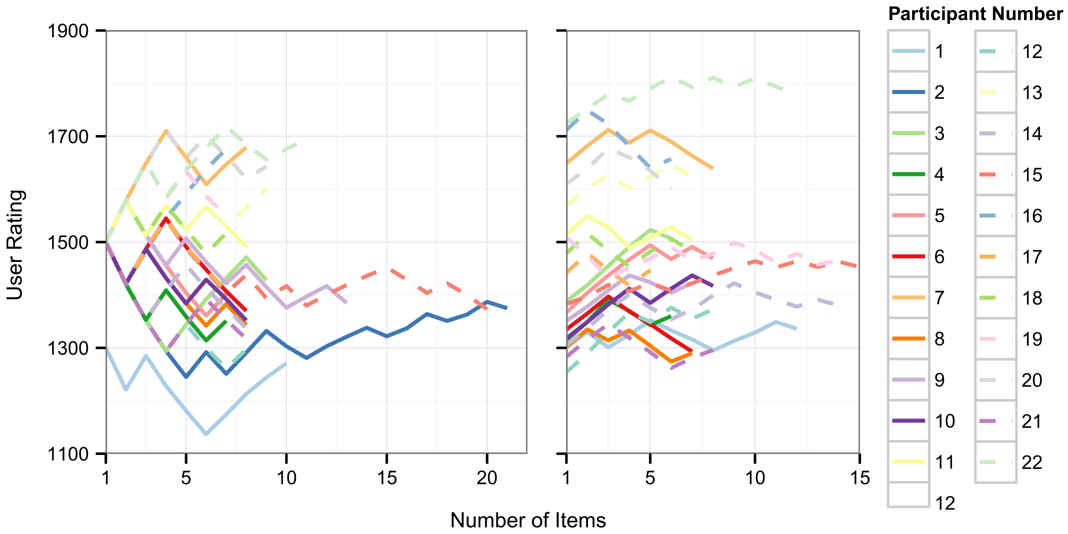

Fig. 3. The user ratings on the imitation game per participant, for session 1 (left) and session 2 (right).

imitation game takes more time than answering a math item. Therefore, fewer items have been answered for the imitation game, than for the math game. After the first session, too few items were answered during the session to state whether or not the user ratings fluctuate around a certain rating. In the second session, the user ratings appear to fluctuate around a certain rating. However, the user rating deviation was still too large, due to the limited number of items answered, to speak of a reliable estimate of the participant's true rating. On average, the user rating changed by 112 during the first session and by 37 in second session.

For the math game, the user ratings unanimously increased during the first 10 items. We take this as a measure of the validity of the ratings; a user rating is a measure of a child's skill and an item rating is a measure of the difficulty of the item. While no unanimous increase in user ratings can be seen for the imitation game, there is a correlation of.54 ($n = 21$, $p = .012$) between the median rating on the math game and the median rating on the imitation game for the second session.

5.3. Item ratings

Fig. 4 shows the average response times for a math item given the level of difficulty. As can be seen, the more difficult the item, the longer it takes for the child to answer the item. There is one exception to this trend, namely items which had an initial item rating of 2700. In the first session, the response times for these items were lower than items with an initial item rating of 2500 or 2600. In the second session, the update of the item ratings corrected the ratings of the items with an initial rating of 2700 by adjusting the ratings downwards to an item rating of approximately 2600. As a result, the response times were more in line with the item difficulty during the second session.

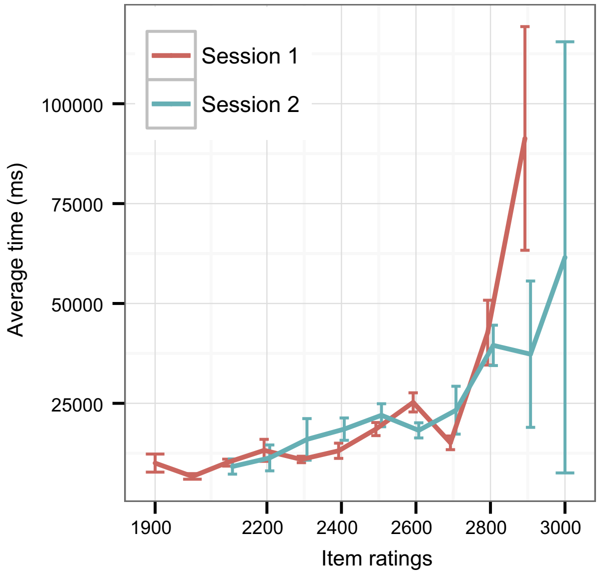

Fig. 4. The average response time (in milliseconds) for answering a math item. The bars show the 95% confidence interval.

In Fig. 5, the difference between the actual and predicted percentages of correct answers per level of difficulty are shown. For 9 out of 12 difficulty levels, items have been answered in both sessions. The difference between the predicted and actual percentage of correct answers decreased in the second session for 7 out of 9 difficulty levels. For difficulty level 2200, the difference remained the same between the two sessions. And for difficulty level 2600, the difference increased during the second session.

During the first session, 462 items were answered, divided amongst 78 unique items. On average, each of the 78 items was answered 5.9 times, with a standard deviation of 4.9.

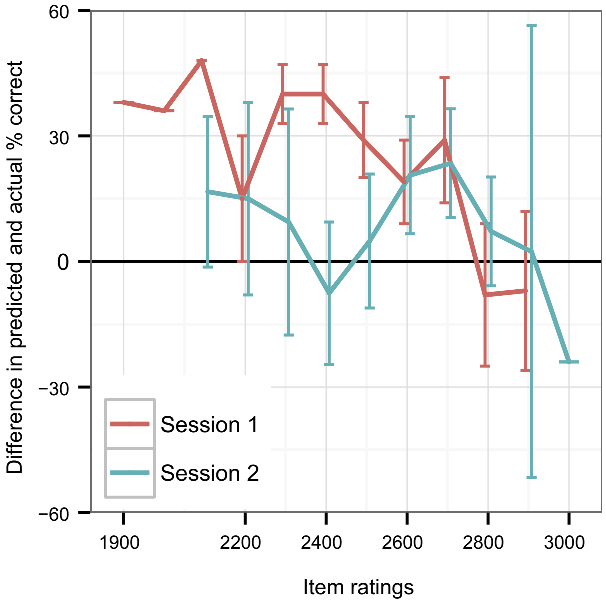

Fig. 5. Difference between the predicted percentage of correct answers and the actual percentage, for the math game.

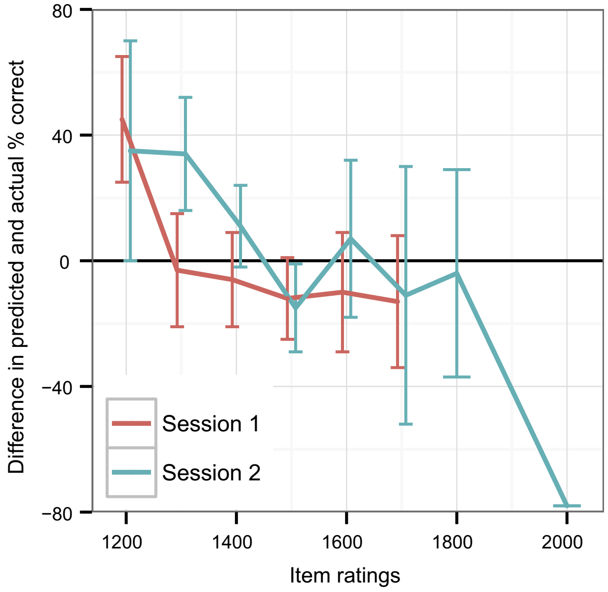

Fig. 6. Difference between the predicted percentage of correct answers and the actual percentage, for the imitation game.

Fig. 6 shows the difference between the actual and predicted percentage of correct answers. Two levels of difficulty deviate from what is expected, namely items with an item rating of 1300 and 1400. These items had a higher percentage correct than was predicted. Compared to the first session, the difference between the actual and predicted percentages of correct answers has become smaller after the adjustment of the item ratings. 208 imitation items were answered during the first session, divided amongst 37 unique items. On average, each of the 37 items was answered 5.6 times, with a standard deviation of 4.2.

5.4. Free-choice period

Of the 21 participants in the second session, 9 chose to continue playing with the robot during the free-choice period. 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 rated the robot with a 4.37 (standard deviation is 0.67).

6. Conclusion and discussion

6.1. Performance of the rating system

We expected that the user ratings would become relatively stable after the first session. The user rating should start to fluctuate around a particular rating, which we assume to be user's true rating. In the math game, the user ratings increased by 447 during the first session. This high increase indicates that the initial user ratings were too low. Contrary, during the second session, the user ratings were fluctuating around a particular rating after answering approximately ten items. The steady increase in user ratings during the first session, and the stability of the user ratings in the second session indicates that the item ratings have predictive value regarding the difficulty of items. For the imitation game, the user rating were less stable, because less items were answered than with the math game.

Interesting to note is that there is a strong correlation between the user ratings of the imitation and math game. Participants with a high user rating on the math game also had a high user rating on the imitation game. 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.

6.2. Providing the optimal challenge

For the participants in the experimental condition, items were selected based upon the current percentage of correct answers, in order to achieve approximately 70% correct. However, we were unable to achieve a difference in the percentage of correct answers between the experimental and control condition in this experiment.

The update of the item ratings between the first and second session shows a clear improvement in accuracy, increasing the predictive power of the rating system. However, during the second session, there was still a significant discrepancy between the predictions made by the rating system and the actual outcome. This can be attributed to the fact that the items were answered only a handful of times, and therefore not reliable enough to be used for the manipulation. Concluding, while the estimates of the rating system were accurate enough to provide the children with a personalised difficulty, they were not accurate enough to increase the percentage of correct answers to 70%.

Other studies have shown that a high measurement precision can be achieved (Glickman, 1999; Klinkenberg et al., 2011). In our case, the number of items answered was too small to achieve a high enough measurement precision. This illustrates the need for a large amount of data before the system can be fully functional. One way of collecting more data is by connecting the robot to a global platform where the item ratings are stored and updated using the data from all the connected social robots. This way, it should be relatively easy to gather reliable item ratings, and when shared, other robots can immediately start with these reliable ratings; the calibration of the item bank is a one-time activity.

The rating system itself can be improved for the math game, by incorporating response times. Maris and van der Maas (2012) propose a novel measurement model that incorporates response times, as well as the outcome of an instance. If the rating of the item is lower than the rating of the child, 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 child, 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. Klinkenberg et al. (2011) incorporated response times in their rating system allowing them to increase the number of correct answers from 50% to 75%, without a great loss of measurement precision.

6.3. Additional options for using the ratings system

The predictive power of the Bayesian rating system can also be used by a social robot to improve the social interaction with the child. The performance of the child may deviate between sessions for various reasons that may warrant the robot's attention. For example, a sudden drop in performance may indicate that the child is less motivated or attentive. The robot can then initiate a dialogue to find out whether this is the case and if it can assist the child. For example, the robot could change the game in case of boredom, or give comfort when the child is distracted by a bad experience. The predictions of a single item can also be used during dialogue, by giving feedback on the difficulty or performance when telling the correctness of the answer (i.e. reassuring the child that it is okay to answer a difficult item incorrectly).

6.4. General conclusion

The goal of this study was to determine the effectiveness of a Bayesian model to estimate child's skill level for playing a game with a robot and personalising the game play to keep children motivated. An important constraint was that only a small number of interactions was provided for this model. While the estimates of the Bayesian rating system were not accurate enough to increase the percentage of correct answers to 70%, the Bayesian rating system was still able to quickly assess the child's skill level and adapt the difficulty of the game accordingly. Providing the optimal challenge by actively increasing the percentage of correct answers to 70% is likely not feasible without access to enough data to reliably calibrate the item rating. Instead, it may be attractive to capitalise on the properties of social robots, like our tendency to project social qualities to the behaviour of technology (Reeves & Nass, 1996), and have the child compete with the robot. In this case, the child can answer only 50% of the items correctly, but still win the game by outperforming the robot. In this scenario, it may be that the child will not perceive the items as being too challenging.

In this paper, we have shown that the rating method and module can be effective for adapting the difficulty of a game to a personal level. Furthermore, we have outlined multiple ways in which the rating module can be used to further personalise and enhance child-robot interaction, other than adapting the difficulty of games.

Acknowledgements

This work is (partially) funded by the European Union FP7 ALIZ-E project (Grant No. 248116) and the Horizon2020 PAL project (Grant No. 643783-RIA).

Furthermore, the authors would like to thank the teachers and the children of the ‘Griftschool’ for their participation in this study.

References

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4–16.

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

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry, 25*(1), 49–59.

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 & Society, 13*(2), 13–24.

Chrysafiadi, K., & Virvou, M. (2013). Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications, 40*(11), 4715–4729.

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*(4), 253–278.

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

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

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

Desmarais, M. C., & de Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction, 22*(1-2), 9–38.

Eggen, T. J. H. M., & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement, 30*(5), 379–393.

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

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

Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M., Mundell, A., Rosenthal, S., et al. (2005). Designing robots for long-term social interaction. In *IEEE/RSJ international conference on intelligent robots and systems, IROS 2005* (pp. 1338–1343).

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

Gordon, G., & Breazeal, C. (2015). Bayesian active learning-based robot tutor for children’s word-reading skills. In *AAAI* (pp. 1343–1349).

Henkemans, O. A. B., Bierman, B. P., Janssen, J., Neerincx, M. A., Looije, R., van der Bosch, H., et al. (2013). Using a robot to personalise health education for children with diabetes type 1: A pilot study. *Patient Education and Counseling, 92*(2), 174–181.

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.

Hvattum, L. M., & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of forecasting, 26*(3), 460–470.

Janssen, J. B., van der Wal, C. C., Neerincx, M. A., & Looije, R. (2011). Motivating children to learn arithmetic with an adaptive robot game. In *Proceedings of the third international conference on Social Robotics 2011* (pp. 153–162).

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, 19*(1), 61–84.

Kennedy, J., Baxter, P., & Belpaeme, T. (2015). The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 67–74).

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, 57*(2), 1813–1824.

Leite, I., Martinho, C., & Paiva, A. (2013). Social robots for long-term interaction: A survey. *International Journal of Social Robotics, 5*(2), 291–308.

Leyzberg, D., Spaulding, S., & Scassellati, B. (2014). Personalizing robot tutors to individuals’ learning differences. In *Proceedings of the 2014 ACM/IEEE international conference on human-robot interaction* (pp. 423–430).

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). Statistical theories of mental test scores.

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

Pelánek, R. (2016). Applications of the ELO rating system in adaptive educational systems. *Computers & Education, 98*, 169–179.

Polson, M. C., & Richardson, J. J. (2013). *Foundations of intelligent tutoring systems*. Psychology Press.

Powers, A., Kiesler, S., Fussell, S., & Torrey, C. (2007). Comparing a computer agent with a humanoid robot. In *2007 2nd ACM/IEEE international conference on Human-Robot Interaction (HRI)* (pp. 145–152).

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute of Educational Research.

Reeves, B., & Nass, C. (1996). *The media equation*. Cambridge University Press.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25*(1), 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. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1613–1622).

Szafir, D., & Mutlu, B. (2012). Pay attention!: Designing adaptive agents that monitor and improve user engagement. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 11–20).

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

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge: Harvard University Press.