Perceived human-likeness of social robots

International Journal of Social Robotics (2019) 11:477–494
https://doi.org/10.1007/s12369-019-00516-z

Perceived Human-Likeness of Social Robots: Testing the Rasch Model as a Method for Measuring Anthropomorphism

Peter A. M. Ruijten[1] · **Antal Haans**[1] · **Jaap Ham**[1] · **Cees J. H. Midden**[1]

Accepted: 8 January 2019 / Published online: 19 January 2019

Abstract
Anthropomorphism is generally defined as the attribution of human-like characteristics to social robots and other non-human objects. We argue that different researchers have different interpretations of this concept, leading to measuring instruments that focus on different subsets of human-like characteristics. In the current paper, we discuss these different interpretations and explore a new method for measuring anthropomorphism, based on the Rasch model. The aim of the current work is to map anthropomorphism as a range of human-like characteristics on a one-dimensional scale. The scale's validity and sensitivity were tested by comparing it with two available measuring instruments and by comparing people's responses to different types of agents. In three studies, we explored whether the Rasch model is suitable for measuring anthropomorphism. Despite some limitations, results showed that the Rasch model can successfully be applied to the measurement of anthropomorphism. Implications for future work on anthropomorphism of social robots are discussed.

Keywords Anthropomorphism · Measurement scale · Rasch model

1 Introduction

During the past decades we have seen an increasing interest in research on social robots. They are being commercialized and becoming available to the general public. Because of this, understanding the effects of their appearance and behavior on people's interactions with them is gaining importance as well. In the near future, social robots may be provided more human-like features and are likely to be represented as social entities with faces, engaging in social conversation with humans. These developments could make people perceive those robots as more and more human-like. This perceived human-likeness is an important determinant for people's responses to social robots (see [1,36,38]). Moreover, human-likeness in social robots has been shown to positively affect people's engagement with those robots [4], people's expectations of the robot's navigation behavior [30], and the robot's persuasive power [7]. The term anthropomorphism refers to the extent to which people perceive robots and other non-human objects as human-like [2,8,9,12,20,32]. This conventional and rather general description of anthropomorphism has lead to a variety of interpretations of the concept. As a consequence, different groups of researchers focus on different subsets of human-like characteristics.

The research in the current paper was funded by the Netherlands Energy Agency.

✉ Peter A. M. Ruijten
p.a.m.ruijten@tue.nl

[1] Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

1.1 Subsets of Human-Like Characteristics

An examination of existing measurement instruments (e.g., [2,5,32]) revealed that human-like characteristics are generally categorized into appearances, thoughts and emotions. With *appearances* we refer to characteristics that reflect human form or behavior (i.e., how an object or robot looks and/or moves), including both physical shapes and physical abilities. With their measuring instrument for anthropomorphism, Bartneck and colleagues [2] focused on such appearances by asking people to indicate, among others, to what extent a robot looks human-like, looks life-like, and shows realistic movements. These items clearly focus on a robot's appearance and the extent to which this appearance resembles the human body.

With *thoughts* we refer to characteristics that reflect cognitive states and processes. According to Waytz and colleagues [32], anthropomorphism is a process of inductive inference

which most likely occurs by attributing cognitive states that are perceived to be uniquely human to other agents (for a review, see [9]). Hence, anthropomorphism was measured by asking people to indicate to what extent an agent has cognitive abilities like consciousness and free will [32]. These cognitive abilities cannot be included in the physical design of robots, because they can only be inferred from its behavior.

Finally, with *emotions* we refer to characteristics that indicate subjective conscious experiences which can be distinguished in primary and secondary ones (for an overview of the hierarchical organization of emotions, see [26]). Eyssel and colleagues [11] measured anthropomorphism by asking people to indicate to what extent robots can experience such primary and secondary emotions.

In their further work, Eyssel and colleagues [10,12] differentiated between personality characteristics that reflect human nature and human uniqueness. This distinction was adapted from earlier research on social perception in humans [16,17]. In this line of research, human nature characteristics are described as characteristics of the human species that are shared with other animals (e.g., innate and affective traits). Uniquely human characteristics, on the other hand, are considered to be exclusive to humans and not possessed by any other species (e.g., social learning and higher cognition, see [17]). Whether or not such a clear-cut distinction exists between humans and other entities is outside the scope of the current work, but it does explain why others have approached it as a 2-dimensional construct.

The aim of this paper is twofold. First, we conceptualize anthropomorphism and discuss why measurements obtained with existing instruments may have little correspondence. Second, we propose a method for measuring anthropomorphism based on the Rasch model (see for example [3]), and test its dimensionality and validity in three studies. We end with a discussion on the benefits and limitations of using the Rasch model for measuring anthropomorphism.

1.2 Conceptualizing Anthropomorphism

We argue that anthropomorphism is a single predisposition, meaning that whatever human-like characteristics an individual ascribes to a robot—be it human nature or human uniqueness characteristics, basic physical abilities or moral decision making, they all stem from that single predisposition to do so. This is in line with Waytz and colleagues [31], who stated that attributions of human-like characteristics stem from stable individual predispositions.

However, Waytz and colleagues [31] also referred to anthropomorphism as the attribution of characteristics that people regard as distinctively human, in particular mental capacities. This view is shared by Zawieska and colleagues [37], who argued that anthropomorphism only includes human-like characteristics that robots do not have. As a result of this ostensibly narrow focus, only mental capacities (i.e., having intentions, free will, a mind of your own, consciousness, and experiencing emotions) are included in the measurement of anthropomorphism (see [31]). More specifically, the Individual Differences in Anthropomorphism Questionnaire (IDAQ) consists of items that describe characteristics of specific agents (e.g., cars, cows, and mountains) within one of three categories (i.e., technologies, animals, and natural entities). We argue that this operationalization of anthropomorphism insufficiently measures the concept, because it focuses merely on a narrow subset of human-like characteristics (mental capacities), and it includes only very specific agents. In other words, we believe that asking a person to what extent an average fish has free will does not explain or predict that person's responses to a social robot. We also wonder whether such a scale can be used for comparing the perceived human-likeness of different types of agents.

It is evident that the different measuring instruments have been developed with a focus on different subsets of the construct. Consequently, the question arises as to what extent these measuring instruments of anthropomorphism measure the same concept, and thus ultimately to whether we can faithfully compare research findings. We view anthropomorphism as a one-dimensional construct, and argue that all human-like characteristics—no matter which subset they belong to—are ordered according to the probability with which they are ascribed to robots. Some human-like characteristics are expected to be more easily ascribed to robots than others. For example, human-like appearances are expected to be more easily ascribed to social robots than underlying cognitive states and processes, regardless of an individual's general tendency to anthropomorphize.

Finally, we argue that the ordering of human-like characteristics with respect to the difficulty to attribute them to robots is similar for all individuals. More specifically, people are expected to be more likely to attribute human-like appearances to robots than they are to attribute cognitive states. Such an invariant ordering also entails that if a person attributes the ability of moral reasoning to a robot, (s)he is also expected to attribute the ability of seeing to that robot. Another person who does not attribute the ability of seeing to the same robot is not expected to attribute the ability of moral reasoning to it.

If all human-like characteristics can be invariantly ordered across people, we can compare people's individual predispositions to anthropomorphize and the perceived human-likeness of a robot on a single scale. One model that is able to map a person's predisposition to anthropomorphize and the human-like characteristics (s)he is likely to attribute to a robot as locations on a single dimension, and thus seems highly suitable for measuring anthropomorphism, is the Rasch model [3].

1.3 The Rasch Model

The Rasch model (see Eq. 1) describes the odds of a certain response as a logistic function of person and item parameters. In our case, this relates to the probability of attributing a specific human-like characteristic i as an additive function of a person n's general predisposition to anthropomorphize (θ_n) and the difficulty to attribute that specific human-like characteristic to a robot (δ_i).

$$\ln\left(\frac{P(x_{ni}=1)}{1-P(x_{ni}=1)}\right)=\theta_n-\delta_i \tag{1}$$

Both parameters in this equation (an individual's predisposition to anthropomorphize θ, and the difficulty of attributing a specific characteristic to a robot δ) are estimated by means of maximum likelihood estimation. Predispositions of persons and the difficulty of the various self-report items (i.e., whether or not a robot is thought to possess a certain characteristic) are expressed in log odd units (also called logits). For a specific characteristic i to have a 50% chance to be attributed to a robot, the difficulty of that characteristic (e.g., $\delta_i = 1$) has to be matched numerically with an equivalent amount of a person n's predisposition to anthropomorphize ($\theta_n = 1$).

The Rasch model, however, is not only descriptive, but also prescriptive. It requires items (or human-like characteristics) to be ordered invariantly and transitively across persons, and persons to be ordered invariantly and transitively across items. Thus, if a person attributes, for example, four out of 10 characteristics to a social robot, the Rasch model also prescribes which four should be attributed: They should be amongst the four least difficult ones. For the most difficult to attribute characteristics, we expect to exclusively find such attributions amongst the individuals that have a high predisposition to anthropomorphize.

These formal Rasch model expectations can be tested empirically against the observed data. Provided the responses are ordered from the person with the highest to the lowest predisposition to anthropomorphize, the model anticipates the following response string for an averagely difficult to attribute characteristic: 111101010000. In this response string, a one indicates that a person agreed to the robot having that characteristics, and a zero that a person did not. The first four individuals all have high predisposition, so all four are likely to attribute this characteristic to the robot. The next four individuals have a predisposition that approaches numerically the difficulty of this characteristic or item. As a result, some of them will attribute this characteristic to the robot, and others will not. Finally the last four individuals have such a low predisposition to anthropomorphize that all are very unlikely to attribute this characteristic to the robot.

The mean square (MS) statistic is commonly used to test the match between model-predicted and observed response patterns. The MS infit-value is the weighted average of the squared standardized residuals, in which each residual is weighted by its variance [3]. The model-predicted response pattern (e.g., 111101010000), yields a MS infit-value of MS = 1.00. Excessively high MS-values can be expected when the observed response string opposes the Rasch model prediction, for example when the likelihood of reporting an experiential effect increases with diminishing susceptibility (e.g., 000000111111). In contrast, the MS-value would be smaller than 1.00 for an item with a deterministic response pattern (i.e., 111111000000). In such a case, the model prediction-to-data fit is better than what one would anticipate with a probabilistic model. MS-values below 1.00 do not really challenge the Rasch model prediction, but can be used to improve one's measurement instrument. Besides MS infit also MS outfit-values are often reported. MS outfit-values are unweighted fit statistics, and are more sensitive to unexpected responses on relatively easy or difficult items.

The invariance assumption should be sufficiently met in order to map both persons and human-like characteristics on a single scale (see 1), and thus to compare individuals and human-like characteristics against each other in a meaningful way. For assessing item fit, MS-values up to 1.20 are considered excellent, and MS-values below 1.50 are considered acceptable [35].

The additional advantage of the model is the invariance between items and persons. As a result, the assessment of a persons predisposition to antropomorphize and of the items with respect to their difficulty are independent of each other (so-called specific objectivity, see [34]). As results, and in contrast to many other measurement models, measurements of personal attributes are not defined by the specific set of items used (see [18]). It thus also allows for items being deleted and/or replaced. This enables the use of different sets of items (see [33]), making this method easily adjustable for measuring responses to different types of agents (e.g., animals or natural phenomena).

1.4 Research Aims

The current research was designed with the aim to explore a new method for measuring anthropomorphism. We hypothesized that anthropomorphism can be successfully mapped onto a one-dimensional scale and that human-like characteristics are ordered with respect to their likelihood of being attributed to robots in a way that is similar for all individuals in their encounter with different types of agents in different contexts. Data from three studies were used to test these hypotheses. Two of these studies were originally designed with different purposes, but we will only assess results on the included measuring instruments for anthropomorphism.

In the first study, we developed and tested the construct validity of a 37-item anthropomorphism scale. With construct validity we refer to the relation between the ordering of items on the scale according to their difficulty and their perceived human nature and human uniqueness. Construct validity is high when strong relations between the locations of items on the dimension and their perceived human nature and human uniqueness are found. Because of the expected unidimensionality, human nature and human uniqueness were also expected to be strongly correlated.

In the second study, a 25-item version of the anthropomorphism scale was tested on its convergent validity. With convergent validity we refer to the extent to which estimates of the scale are related to estimates obtained with two other measuring instruments for anthropomorphism: the questionnaire used by Waytz and colleagues [32, referred to as the Waytz-instrument, see "Appendix A"] and the anthropomorphism part of the Godspeed questionnaire developed by Bartneck and colleagues [2, referred to as the Godspeed-instrument, see "Appendix B"]. Because the focus of these instruments is on different subsets of human-like characteristics, we expected to find moderate correlations between the anthropomorphism scale and the Waytz- and Godspeed-instruments. The second study was originally designed to compare two different robots on their perceived human-likeness. Because of this, we could use this study for testing the invariant ordering of the items on the scale for those two robots.

In the third study, we extended our scope from robots to other types of agents. A 19-item version of the anthropomorphism scale was tested on its sensitivity for differentiating between humans and other types of agents. This study was originally designed to investigate people's responses to four types of players in a game (i.e., humans, robots, computers, and algorithms), enabling us to compare responses for other types of technologies than robots as well. We expected that the scale would successfully differentiate between humans and other types of agents.

Finally, we expected that the anthropomorphism scale would show an invariant ordering of human-like characteristics on a single dimension across all studies and all experimental conditions.

2 Study 1

In this study, a list with 37 human-like characteristics was created, largely based on earlier work on humanness and anthropomorphism [2,11,16,17,32], and tested on its construct validity. The human-like characteristics were modeled as a function of a person's predisposition to anthropomorphize and the difficulty to attribute that human-like characteristic to a robot. We hypothesized that items and persons could be mapped onto a single one-dimensional scale, and that the items would be invariantly ordered according to the difficulty with which they are attributed to a robot.

Additionally, for construct validity purposes, the extent to which the 37 human-like characteristics were perceived as being human nature and uniquely human was measured. We hypothesized that the estimated difficulties with which the 37 characteristics are attributed to a robot would be related to their perceived human nature and human uniqueness. Because of the proposed unidimensionality of anthropomorphism, human nature and human uniqueness were expected to be strongly correlated as well.

2.1 Method

2.1.1 Participants and Design

One hundred and sixty one participants sampled through social media participated in one of three groups in the current study. The first group consisted of 124 participants (53 males and 71 females; $M_{age} = 26.08$, $SD_{age} = 8.82$, Range = 15 to 59) who were given a description about a robot and completed 37 survey items. Another group of 20 participants (9 males and 9 females, $M_{age} = 19.94$, $SD_{age} = 1.98$, Range = 18 to 23; two participants did not indicate their age and gender) rated the 37 human-like characteristics on human nature. The remaining 17 participants (11 males and 6 females, $M_{age} = 21.12$, $SD_{age} = 1.80$, Range = 18 to 24) rated all characteristics on human uniqueness. Participants in all three groups participated voluntarily, gave informed consent, and were not compensated for participation.

2.1.2 Materials and Procedure

A set of 37 items describing human-like characteristics was constructed. For all three groups of participants, items were arranged in alphabetical order.

For the first group of 124 participants, items were formulated as a statement which could be answered with yes (coded with a 1) or no (coded with a 0). The items were presented through an online survey. After reading a short explanation about the study, participants were provided with a short description about a robot: 'The robot has eyes to perceive the environment, has arms and legs to move around in this environment, and today the robot is trying to solve a moral dilemma'. This description was followed by an instruction to not think elaborately about the statements and to give the answers that first came to mind. Finally, participants indicated their gender, age and education level, and they were thanked for their contribution. This study took approximately 5 min to complete.

The two other groups of participants were not given the description of the robot, but instead were asked to indicate to

what extent each of the 37 items on the scale was perceived as 'typically human' (i.e., human nature) or 'uniquely human' respectively. Human nature was measured with one question ('To what extent is this characteristic *typical* for humans?') on a 7-point response format ranging from 'not at all' (coded with a 1) to 'very much' (coded with a 7). Human uniqueness was measured with one question ('Is this characteristic *unique* for humans?') on a dichotomous response format with 'not unique' (coded with a 0) and 'unique' (coded with a 1) as options. The two concepts were not further introduced or explained to participants. We chose for a dichotomous scale for human uniqueness because a characteristic either is, or is not uniquely human (i.e., characteristics cannot be 'a little' unique for humans). Both these evaluations took approximately 5 min to complete.

2.2 Results and Discussion

2.2.1 Model Test

To test whether the data sufficiently fit the model, four tests were conducted. First, fit statistics were used to test whether items and persons fitted the Rasch model (for an overview, see [3]). For assessing item fit, both infit and outfit were used. Infit indicates unexpected observations on items that are close in difficulty to a person's predisposition, whereas outfit indicates unexpected observations on items that are relatively easy or difficult [23]. Infit and outfit mean square (MS) values ≤ 1.20 are considered excellent, and MS-values ≤ 1.50 are considered acceptable [35]. The second test determined whether the items were sufficiently spread over the perceived human-likeness dimension. The third one tested the hypothesis that items all belong to a single dimension. Finally, the fourth one tested the hypothesis that items would be invariantly ordered according to the difficulty of attributing them to robots.

Item fit Ideally, each of the items contributes in a meaningful way to the measuring instrument, indicated by a sufficient item fit. Considering the notion that the Rasch model is stochastic and that data depend on probability (and not on certainty), some misfit is to be expected [27]. An acceptable five of the 37 items had outfit MS values outside of the acceptable boundaries. These were items 1 ('experience pain', outfit MS = 2.09), 26 ('anticipate on surroundings', outfit MS = 1.62), 32 ('organized', outfit MS = 2.74), 33 ('estimate distances', outfit MS = 2.12), and 37 ('avoid objects', outfit MS = 1.97). Another 7 items had outfit MS values between the good and acceptable values (see Table 1 for estimated item difficulties).

Item difficulties were estimated with a reliability of .98, and the average item difficulty was anchored at $M = .00$ logits ($SD = 2.34$, Range = -4.60 to 4.08). Infit MS values of the 37 items ranged from 0.72 to 1.22 ($M = 0.98$, $SD = 0.12$), and outfit MS values ranged from 0.46 to 2.74 ($M = 1.10$, $SD = 0.50$). These findings together indicate that there was an acceptable item fit, meaning that there were not many observations that did not fit the overall structure of the data.

Person fit The purpose of person fit measurement is to detect response patterns that are unlikely given the model [24]. More specifically, person fit indicates whether a person responds as expected given his/her individual predisposition to anthropomorphize. Individual predispositions to anthropomorphize were estimated with a separation reliability of .80. The average predisposition was $M = -.21$ logits ($SD = 1.16$; Range = -4.13 to 2.74). Since items are responded to by individuals who can be tired or misread the statements, some misfit is to be expected. For an acceptable ten out of 124 participants (8.1%), the model prediction did not fit the data as indicated by a t-value of $t \geq 1.96$. These findings indicate that there was an acceptable person fit, meaning that there were not many observations that did not fit the overall structure of the data.

Item spread All items and persons are mapped onto a single scale in Fig. 1. As can be seen in this figure, the spread of items sufficiently covers the spread of persons. In other words, the current scale was able to reliably measure individual predispositions to anthropomorphize for all participants in the current sample. It also appeared that the top region of the scale comprised many items, but not so many persons. Thus, some items appeared to be too unlikely for persons in the present sample to attribute to the robot, and therefore did not contribute to the assessment of individual differences in people's predisposition to anthropomorphize. For this reason, some of these items will be omitted from the scale in the next studies.

Dimensionality Next, we tested the hypothesis that the items of the anthropomorphism scale would all belong to one dimension. Results showed that the Rasch model explained 52.8% of the variance in the data (for computational details, see [22]). A Principal Component Analysis was performed on the standardized residuals (i.e., the data not explained by the model), which checks whether multiple items share the same unexpected response pattern (for details, see [22,29]). If the model would fit perfectly, then 52.5% of the overall variance would be quantification variance, revealing a slight overfit (0.3%) to the model. Because the Rasch model estimates probabilities for discrete events (i.e., whether a person attributes a certain human-like characteristic to a robot or not), substantial quantification variance is to be expected (see also [15]). The empirical proportion of unexplained variance (i.e., 47.2%) was thus highly similar to the proportion of quantification variance one would expect with a perfect data-to-model fit (i.e., 47.5%).

An additional factor would result in an increase of 6.8% in the proportion of explained variance. The set of items thus largely tapped into a single dimension only. These findings supported the expected unidimensionality of the

Table 1 Item difficulties (δ), infit- and outfit mean squares, average values of human uniqueness and human nature (with the latter adjusted to the same scaling by subtracting 1 and dividing it by 7) of the anthropomorphism scale in Study 1

	Item	δ (SE)	Infit MS	Outfit MS	Human uniqueness	Human nature
1.	Experience pain	4.08 (.60)	1.04	2.04	0.00	0.54
2.	Unhappy about the dilemma	3.77 (.52)	0.91	0.69	0.88	0.75
3.	Imaginative	3.52 (.47)	0.86	0.51	0.71	0.78
4.	Angry	3.32 (.44)	0.87	0.71	0.06	0.78
5.	Empathize	2.84 (.36)	0.73	0.46	0.41	0.73
6.	Happy	2.71 (.35)	0.88	0.86	0.12	0.69
7.	Chose the dilemma	2.71 (.35)	1.03	1.13	0.82	0.75
8.	Satisfied	2.29 (.30)	0.79	0.58	0.12	0.54
9.	Responsible	2.29 (.30)	0.86	0.64	0.24	0.72
10.	Free will	2.20 (.30)	1.14	1.14	0.18	0.63
11.	Understand others' emotions	1.88(.27)	0.86	0.62	0.24	0.66
12.	Ambitious	1.68 (.26)	0.95	0.72	0.76	0.81
13.	Understands the dilemma	1.21 (.23)	1.04	0.92	0.94	0.67
14.	Recognize others' emotions	0.55 (.21)	0.89	0.85	0.06	0.60
15.	Intention not to harm others	0.47 (.21)	1.06	1.03	0.59	0.65
16.	Think about the dilemma	0.42 (.21)	0.94	0.85	0.88	0.81
17.	Self-conscious	0.42 (.21)	0.99	0.95	0.29	0.73
18.	Jump	0.26 (.20)	1.22	1.24	0.00	0.34
19.	Deliberate actions	− 0.15 (.20)	1.11	1.38	0.06	0.58
20.	Talk	− 0.19 (.20)	0.95	1.04	0.24	0.78
21.	Solve riddles	− 0.47 (.20)	1.01	0.99	0.47	0.67
22.	Recognize voices	− 0.72 (.21)	1.11	1.46	0.00	0.48
23.	Understand language	− 0.89 (.21)	0.86	0.75	0.24	0.75
24.	Rational	− 0.93 (.21)	1.05	1.36	0.71	0.76
25.	See depth	− 0.97 (.21)	1.07	1.21	0.00	0.55
26.	Anticipate on surroundings	− 1.45 (.23)	1.09	1.62	0.12	0.58
27.	Conscious about surroundings	− 1.66 (.24)	0.72	0.54	0.00	0.50
28.	Detect color	− 1.83 (.24)	0.92	0.87	0.00	0.54
29.	Purposeful	− 1.96 (.25)	0.92	1.50	0.18	0.58
30.	Calculate	− 2.16 (.26)	0.94	0.78	0.35	0.67
31.	See	− 2.54 (.29)	1.13	0.98	0.00	0.35
32.	Organized	− 2.73 (.31)	1.17	2.74	0.18	0.52
33.	Estimate distances	− 2.73 (.31)	1.12	2.12	0.06	0.48
34.	Pick up objects	− 3.05 (.34)	1.02	0.86	0.06	0.54
35.	Walk	− 3.61 (.42)	1.02	1.49	0.00	0.48
36.	Detect objects	− 3.02 (.48)	1.07	0.99	0.00	0.49
37.	Avoid objects	− 4.60 (.61)	0.99	1.97	0.00	0.39

scale, showing that individual differences in predispositions to anthropomorphize can be assessed on a single scale of equal additive units. In other words, all human-like characteristics included in the scale were successfully mapped onto a single dimension, ranging from low to high on perceived human-likeness.

Invariant ordering To test the hypothesis that items on the scale would be invariantly ordered according to the difficulty of attributing them to robots, the sample was split in half and item difficulties were estimated twice: once for participants with even and once for participants with odd identification numbers. Consistent with the hypothesis of person-independent item difficulties, the two estimates of the 37 items were highly similar, $r = .97$, $p < .001$. The item invariance plot is provided in Fig. 2a. As can be seen in this figure, the ordering of items on the scale by their difficulty

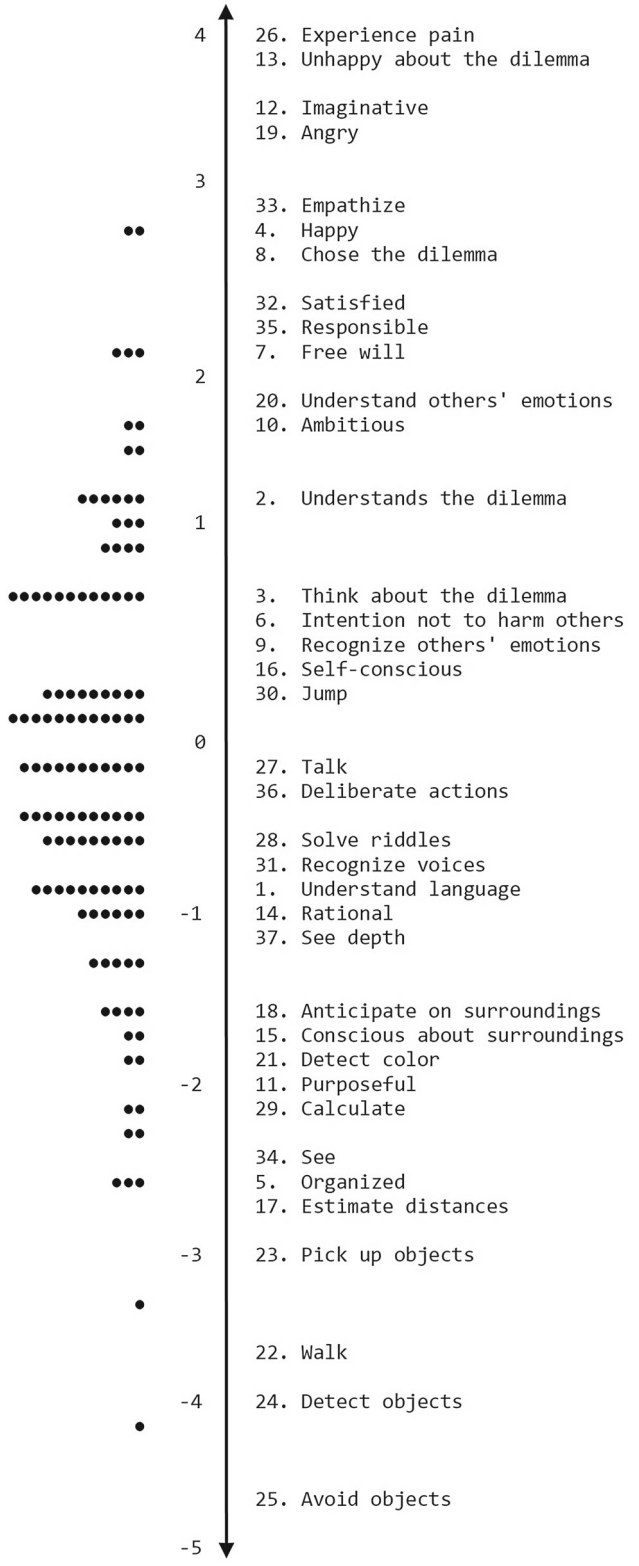


Fig. 1 Item-person map of Study 1, displaying the estimates of participant's predisposition to anthropomorphize and the item difficulty linked with each human-like characteristic mapped onto a single scale of equal additive units. Each number on the right represents an item. Each dot on the left represents a person

to ascribe them to the robot is similar across the samples of participants with even and odd identification numbers.

We also performed the 'Wright's challenge' (see [3]). For this, the sample was split in half once more, but this time according to the participants' estimated predispositions. More specifically, item difficulties were estimated for participants with high predispositions and for participants with low predispositions to anthropomorphize separately. The estimates of the 37 items were again highly similar, $r = .92$, $p < .001$. The item invariance plot is provided in Fig. 2b. As can be seen in this figure, the ordering of items on the scale is also similar across the samples of participants with high and low individual predispositions to anthropomorphize.

2.2.2 Construct Validity

To test the hypothesis that the ordering of items of the anthropomorphism scale according to their difficulty was related to perceived human nature and human uniqueness, item difficulties were compared with the mean scores on these concepts. Results showed moderate but significant correlations between item difficulties and human nature ($r = .60$, $p < .001$) and between item difficulties and human uniqueness ($r = .44$, $p < .01$). The higher a characteristic was rated on human nature and/or human uniqueness by participants in the content validity groups, the less likely participants in the survey were to attribute that specific human-like characteristic to a robot. These results support the expectation that the difficulty to attribute a specific characteristic to a robot is related to that characteristic's perceived human nature and human uniqueness. In line with previous literature [10], a significant correlation was found between human nature and human uniqueness ($r = .71$, $p < .001$), indicating that items that were rated high on human nature were more likely to be indicated as being uniquely human, whereas items that were rated low on human nature were less likely to be indicated as being uniquely human. This finding supports the expectation that human-like characteristics can be mapped onto a one-dimensional scale.

2.3 Conclusions

In the current study, a 37-item anthropomorphism scale was tested on its construct validity. Results showed that people's responses sufficiently fitted the Rasch model, indicated by an acceptable data-to-model fit. Also, all human-like characteristics included in the scale could be successfully mapped onto a single dimension, confirming the hypothesis that anthropomorphism can be measured on a single scale of equal additive units. Moreover, an invariant ordering was found when splitting the sample in half, supporting the expectation

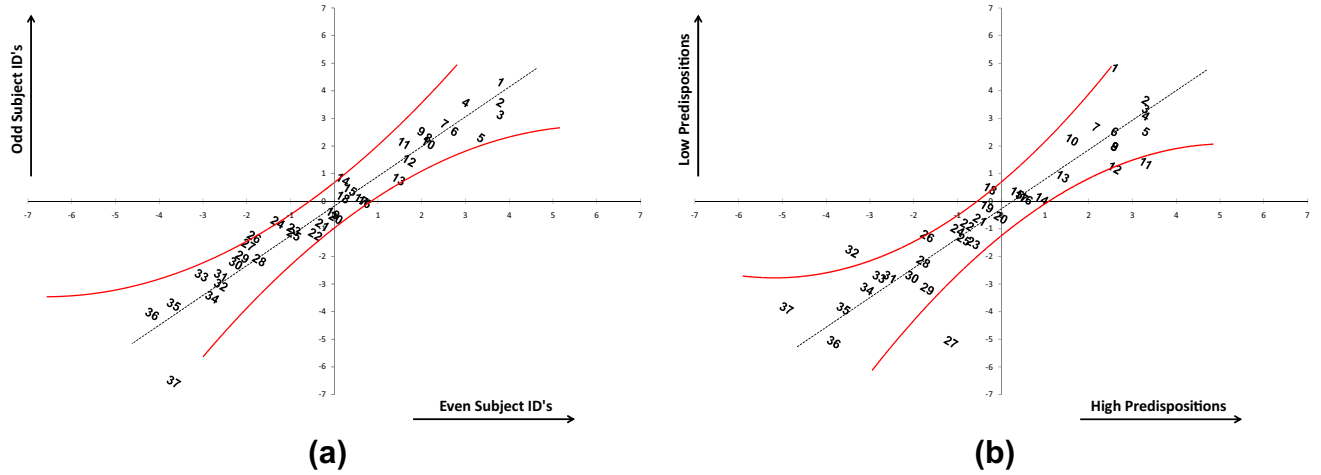


Fig. 2 Item invariance plots of the item difficulties of subjects with **a** even and odd identification numbers and **b** high and low predispositions to anthropomorphize in Study 1. Each number represents an item, corresponding with the numbers in Table 1. Red lines indicate 95% confidence intervals. (Color figure online)

that human-like characteristics can be invariantly ordered with respect to the probability of ascribing them to a robot.

As expected, items high in human nature were found to be more difficult to attribute to a robot than those low in human nature. Additionally, uniquely human characteristics were shown to be more difficult to attribute to robots than non-unique ones, as indicated by their locations on the scale. These results indicated that the difficulty of attributing a specific item of the scale to a robot was related to that item's perceived human-likeness. As such, item difficulties are a valuable indicator of how a specific characteristic perceived human-likeness, which is an important aspect in the measurement of anthropomorphism. In sum, the scale had high construct validity and the method so far seems to be suitable for measuring anthropomorphism. In the next study, the scale's convergent validity will be tested by comparing it with existing measuring instruments.

3 Study 2

The current study was designed to investigate the relation between different measuring instruments for anthropomorphism in people's evaluations of two different robots. Data of this study will be used to test the convergent validity of our anthropomorphism scale. With convergent validity we refer to the extent to which estimates obtained with the scale converged with those obtained with two commonly used measuring instruments for anthropomorphism: the Waytz-instrument and the Godspeed-instrument. We tested to which extent estimates made with the three different instruments would be related. In addition, two different robots were evaluated on their perceived human-likeness, and we expected an invariant ordering of the items on the scale for those two robots.

3.1 Method

3.1.1 Participants and Design

One hundred and thirty one participants sampled through social media participated in this study. Of these 131 participants, 48 were male and 83 female (M_{age} = 34.86, SD_{age} = 17.59, Range = 13 to 77). They were randomly assigned to one of two groups in which they watched a video of a robot that either resembled mostly human-like *physical* features ($n = 68$) or mostly human-like *cognitive* features ($n = 63$). The two robots did not differ on any of the three included anthropomorphism measuring instruments (all t's < 1.31, all p's $> .20$), allowing data of both experimental conditions to be combined into a single sample for the analyses. All participants participated voluntarily, gave informed consent, and were not compensated for participation.

3.1.2 Materials and Procedure

Participants performed the study online. On the welcome page, they could choose to complete the study in Dutch or in English, after which they were provided information about the procedure of the study in their preferred language. Next, they watched a short (about 1 min) video of one of the two robots, depending on the experimental condition they were in. The robot with human-like physical features was running around and pouring water in a cup (the video can be found at https://goo.gl/npyDfG), and the robot with human-like cognitive features appeared to become angry at a person who left dirt on the floor (the video can be found at https://goo.gl/i2Sqgg).

After participants watched the video of one of the two robots, they completed the three measuring instruments for

Table 2 Item difficulties (δ), infit- and outfit mean squares of the anthropomorphism scale in Study 2

	Item	δ (SE)	Infit MS	Outfit MS
13.	Understands the dilemma	5.76 (.76)	1.30	5.21
2.	Unhappy about the dilemma	5.76 (.76)	0.71	0.10
9.	Responsible	4.43 (.46)	0.76	0.22
5.	Empathize	4.22 (.43)	1.15	1.64
4.	Angry	3.88 (.39)	1.00	2.01
12.	Ambitious	3.05 (.31)	1.03	0.73
11.	Understand others' emotions	2.62 (.28)	0.79	0.81
17.	Self-conscious	2.46 (.27)	1.04	0.97
14.	Recognize others' emotions	1.29 (.22)	1.03	0.97
27.	Conscious about surroundings	− 0.23 (.21)	0.93	0.77
19.	Deliberate actions	− 0.58 (.21)	1.08	1.07
21.	Solve riddles	− 0.63 (.21)	1.20	1.17
23.	Understand language	− 0.72 (.22)	1.04	0.89
31.	See	− 1.01 (.22)	0.94	0.92
25.	See depth	− 1.01 (.22)	1.01	5.97
20.	Talk	− 1.27 (.24)	0.96	0.95
29.	Purposeful	− 1.39 (.24)	0.94	0.68
26.	Anticipate on surroundings	− 1.69 (.26)	0.92	0.72
18.	Jump	− 2.40 (.31)	1.01	0.92
30.	Calculate	− 2.84 (.36)	0.92	0.81
36.	Detect objects	− 3.28 (.41)	0.82	0.31
33.	Estimate distances	− 3.67 (.48)	0.78	0.51
22.	Recognize voices	− 3.67 (.48)	1.06	0.90
35.	Walk	− 3.92 (.52)	1.12	0.83
34.	Pick up objects	− 5.16 (.82)	1.59	0.70

anthropomorphism. The first one was a 25-item version of the anthropomorphism scale that was developed and tested in Study 1 and adjusted for the current study (see Table 2 for the items in this study). Some of the most difficult items, as well as the easiest one, were deleted because they were expected to contribute little to estimations of people's predisposition to anthropomorphize (i.e., items 1, 3, 6, 7, 8, 10, and 37 in Table 1). Three items were deleted because the construct validity test in Study 1 showed that they did not sufficiently relate to anthropomorphism (i.e., items 24, 28, and 32, in Table 1). Item 15 in Table 1 was deleted because it was phrased as a double negation. Items were formulated as a statement which could be answered with yes (coded with a 1) or no (coded with a 0).

The second questionnaire was the Godspeed-instrument (see "Appendix B"), which consisted of 5 items ($\alpha = .71$) with a 5-point response format. Five dummy items were included to make the goal of this questionnaire and of this study less obvious. Participants' averaged responses across the five target items were used in the analyses. The third questionnaire was the Waytz-instrument (see "Appendix A"), which consisted of 6 items ($\alpha = .78$) with a 5-point response format. Participants' averaged responses across the six items were used in the analyses.

After completing these questionnaires, participants completed several questionnaires that were related to the original purpose of the study. These questionnaires measured concepts such as desire for control, need to belong, trust, and predictability. Data on these questionnaires will not be used in our analysis. Finally, participants indicated their age and gender, were debriefed and thanked for their participation. The study took approximately 10 min to complete.

3.2 Results and Discussion

3.2.1 Model Test

In this section, the hypotheses that items and persons can be mapped onto a single one-dimensional scale, and that items are invariantly ordered according to the difficulty with which they are ascribed to a robot are tested. The section has the same structure as in Study 1.

Item fit As in Study 1, most items fitted the model sufficiently with infit and outfit MS values ≤ 1.50 (see Table 2 for estimated item difficulties), except for items 13 ('understands

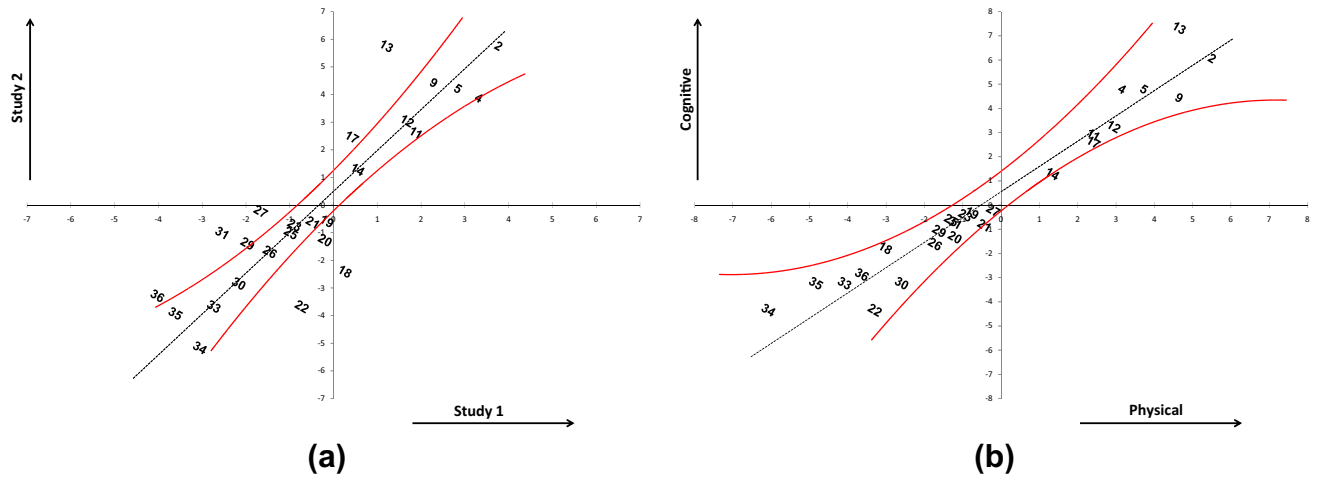


Fig. 3 Item invariance plot of the item difficulties of **a** Studies 1 and 2, and **b** robots with mostly physical and cognitive human-like features. Each number represents an item, corresponding with the numbers in Table 2. Red lines indicate 95% confidence intervals. (Color figure online)

moral dilemmas', outfit MS = 5.21), 5 ('empathize', outfit MS = 1.64), 4 ('angry', outfit MS = 2.01), 25 ('see depth', outfit MS = 5.97), and 34 ('pick up objects', infit MS = 1.59).

Item difficulties were estimated with a reliability of $\alpha = .98$. The average item difficulty was anchored at $M = .00$ logits ($SD = 3.14$, Range = -5.16 to 5.76). Infit MS values of the 25 items ranged from 0.71 to 1.59 ($M = 1.00$, $SD = 0.18$). Outfit MS values of the 25 items ranged from 0.10 to 5.97 ($M = 1.23$, $SD = 1.35$).

Person fit Individual predispositions to anthropomorphize were estimated with a reliability of $\alpha = .78$. The average predisposition was $M = .33$ logits ($SD = 1.59$, Range = -5.70 to 5.47). For a reasonable eight out of 131 participants (6.1%), the model prediction did not fit the data as indicated by a t-value of $t \geq 1.96$.

Dimensionality The Rasch model explained 63.7% of the variance in the data. If the model would fit perfectly, then 63.5% of the overall variance would be quantification variance. The empirical proportion of unexplained variance (i.e., 36.3%) was thus highly similar to the proportion of quantification variance expected with a perfect data-to-model fit (i.e., 36.5%). An additional factor would result in an increase of a trivial 3.6% in the proportion of explained variance. The set of items thus largely tapped into a single factor only.

Invariant ordering The ordering of the item difficulties was highly similar to that obtained in Study 1, as indicated by a strong positive correlation between the item difficulties estimated in Studies 1 and 2 ($r = .88$, $p < .001$, see Fig. 3a for the invariance plot). This result supports the expectation that the probability with which the various human-like characteristics are ascribed to robots is largely independent of the individual's predisposition to do so. In other words, the scale showed an ordering of human-like characteristics that is similar for different individuals in different samples.

To explore whether the expected invariance of item difficulties also holds across the two different robots that were evaluated, the sample was split in half with respect to the robot that was evaluated. Consistent with the hypothesis of robot-independent item difficulties, the two sets of estimates (one for the robot with mostly physical human-like features, the other for the robot with mostly cognitive human-like features) were highly similar ($r = .97$, $p < .001$, see Fig. 3b for the invariance plot). This finding again supports the expectation that human-like characteristics are invariantly ordered with respect to the difficulty to attribute them to robots.

3.2.2 Convergent Validity

To test to what extent estimates obtained with the anthropomorphism scale converged with the two commonly used measuring instruments for anthropomorphism, the three scales were compared. Results indicated a low, but statistically significant correlation between our scale and the Godspeed-instrument ($r = .22$, $p = .01$). After correcting for measurement error attenuation, the correlation remained rather low ($r = .30$, for computational details, see [6]). In addition, a moderate and statistically significant correlation was found between our scale and the Waytz-instrument ($r = .46$, $p < .001$). After correcting for measurement error attenuation, this correlation remained rather moderate ($r = .59$).

3.3 Conclusions

In this study, an adjusted 25-item version of the anthropomorphism scale was compared with two available measuring instruments for anthropomorphism to test for convergent validity. Results showed that, as in Study 1, people's responses sufficiently fitted the Rasch model, indicated by

an acceptable data-to-model fit. This result supported the expected invariant ordering of human-like characteristics with respect to their item difficulty. Additionally, the scale correlated with existing measures of anthropomorphism, but not to the extent to which we could claim convergence. This indicates that our understanding of anthropomorphism is still limited. We will elaborate more on this in the general discussion.

No differences between the robots were found on any of the measuring instruments for anthropomorphism. The ordering of items on our anthropomorphism scale was also highly similar for both robots, indicating that the two robots in this study were perceived as equally human-like. An interesting question thus is how well the scale differentiates between other types of agents. In the next study, people's predispositions to anthropomorphize humans, robots, computers, and algorithms will be compared.

4 Study 3

The current study was originally designed to investigate people's responses to different types of players in a social game: the Ultimatum Game (see [13]). Data of this study will be used to explore the anthropomorphism scale's sensitivity for differentiating between various types of technologies. We compared people's responses to humans, robots, computers, and algorithms. We hypothesized that the scale would successfully differentiate humans from robots. In addition, we explored whether a difference exists between the attribution of human-like characteristics to computers and algorithms.

4.1 Method

4.1.1 Participants and Design

Two hundred and two participants (89 males and 113 females; M_{age} = 34.69, SD_{age} = 11.12, Range = 18 to 76) were recruited via Amazon Mechanical Turk (MTurk) to participate in an online experiment. They were randomly assigned to one of four experimental conditions (Agent type: human vs. robot vs. computer vs. algorithm) of a between-subjects design. Participants were paid $1 for their participation.

4.1.2 Materials and Procedure

Participants performed the study online. To create a social interaction that provides opportunities to anthropomorphize the agents, participants played the Ultimatum Game (see [13]). In this game, two players divide a sum of credits. The first player (the agent) proposes a certain division and the second player (the participant) decides whether (s)he accepts or rejects the offer. Participants in the algorithm condition were told that there were no other players available, and that they would be connected to an algorithm that would 'randomly generate offers' during the game. Participants in the other three conditions were told to be playing the game with humans, robots, or computers. During the game, participants were shown pictures of the other players. An example of each of the agents is provided in Fig. 4.

After playing the Ultimatum Game, participants completed an adjusted 19-item version of the anthropomorphism scale. The main dependent variable in this study was participants' behavior in the ultimatum game, and the anthropomorphism scale was included to explore the possible mediating/moderating role of anthropomorphism. Since the latter was not the main goal, and due to time constraints, fewer indicators were used in this study. Items for this adjusted version were selected from the original set of 37 items in such a way that they would still cover a wide range of the anthropomorphism continuum (see Table 3 for the items in this study). Items were formulated as a statement which could be answered with yes (coded with a 1) or no (coded with a 0).

After completing the anthropomorphism questionnaire, participants indicated their age and gender, were debriefed, thanked for their participation, and paid through the MTurk system. The experiment took approximately 6 min to complete.

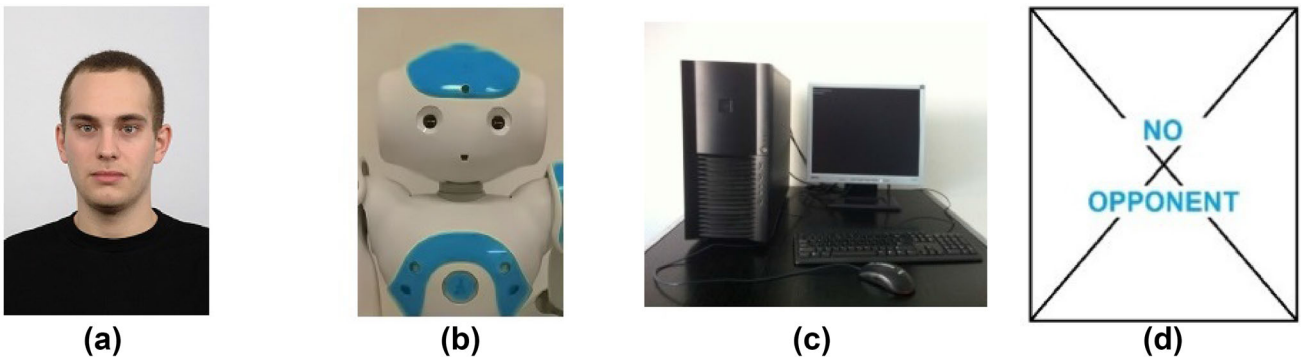


Fig. 4 Examples of pictures used in the **a** human, **b** robot, **c** computer, and **d** algorithm groups in Study 3

4.2 Results and Discussion

4.2.1 Model Test

In this section, the hypotheses that items and persons can be mapped onto a single one-dimensional scale and that items are invariantly ordered according to the difficulty with which they are ascribed to agents are tested. The section has the same structure as in studies 1 and 2.

Item fit Most items fitted the model sufficiently with infit and outfit MS values ≤ 1.50 (see Table 3 for estimated item difficulties), except for items 12 ('ambitious', outfit MS = 1.59), 33 ('estimate distances', outfit MS = 1.97), and 30 ('calculate', outfit MS = 5.59).

Item difficulties were estimated with a reliability of $\alpha = .98$. The average item difficulty was anchored at $M = .00$ logits ($SD = 1.83$, Range = -4.15 to 3.05). Infit MS values of the 19 items ranged from 0.76 to 1.29 ($M = 1.00$, $SD = 0.17$). Outfit MS values of the 19 items ranged from 0.46 to 5.59 ($M = 1.14$, $SD = 1.12$).

Person fit Individual predispositions to anthropomorphize were estimated with a reliability of $\alpha = .88$. The average predisposition was $M = -.15$ logits ($SD = 3.28$, Range = -5.79 to 5.31). For a reasonable ten out of 202 participants (5.0%), the model prediction did not fit the data as indicated by a t-value of $t \geq 1.96$.

Dimensionality The Rasch model explained 52.4% of the variance. If the model would fit perfectly, then 52.1% of the overall variance would be quantification variance. The proportion of unexplained variance (i.e., 47.6%) was highly similar to the proportion of quantification variance expected with a perfect data-to-model fit (i.e., 47.9%). An additional factor would result in an increase of 8.3% of the explained variance.

Next, the ordering of items in the current study was compared with that of Study 1. Results showed a moderate but significant correlation between the two estimates ($r = .58$, $p < .01$, see Fig. 5 for the invariance plot). Despite the significant correlation between the item difficulties, many of the estimated difficulties appeared outside of the 95% confidence interval, indicating substantial differences between the two studies. Some of these items (i.e., 'pick up objects', or 'walk') had lower item difficulties in Study 1 than in Study 3, and some (i.e., 'calculate', or 'understands language') had lower item difficulties in Study 3 than in Study 1. Some human-like characteristics were thus more easy or more difficult to attribute to a specific type of agent than other characteristics. More specifically, the expected invariant ordering of human-like characteristics with respect to their probability of being ascribed to non-human agents was not supported when other agents than robots were evaluated. In the following sections, we will investigate the invariant ordering of the items in Study 3 in more detail.

4.2.2 Invariant Ordering of Items

To explore which human-like characteristics differed in their probability of being attributed to humans, robots, computers, and algorithms, item difficulties were estimated separately

Table 3 Item difficulties (δ), infit- and outfit mean squares of the anthropomorphism scale in Study 3

	Item	δ (SE)	Infit MS	Outfit MS
14.	Recognize others' emotions	3.05 (.36)	1.29	1.03
2.	Unhappy about the dilemma	2.36 (.32)	1.10	0.70
4.	Angry	2.17 (.31)	0.83	0.60
11.	Understand others' emotions	1.90 (.29)	0.93	0.53
13.	Understands the dilemma	1.51 (.27)	0.99	0.79
27.	Conscious about surroundings	1.04 (.25)	0.78	0.46
18.	Jump	1.04 (.25)	0.86	0.73
35.	Walk	0.27 (.22)	0.87	0.64
17.	Self-conscious	0.17 (.22)	0.76	0.54
26.	Anticipate on surroundings	0.12 (.22)	0.96	0.64
34.	Pick up objects	− 0.02 (.22)	0.92	0.71
12.	Ambitious	− 0.30 (.21)	1.09	1.59
20.	Talk	− 0.30 (.21)	0.99	0.77
33.	Estimate distances	− 0.95 (.21)	1.05	1.97
36.	Detect objects	− 1.20 (.21)	0.80	0.71
21.	Solve riddles	− 1.54 (.21)	1.34	1.17
29.	Purposeful	− 2.12 (.22)	1.17	1.14
23.	Understand language	− 3.04 (.24)	1.03	1.31
30.	Calculate	− 4.15 (.29)	1.25	5.59

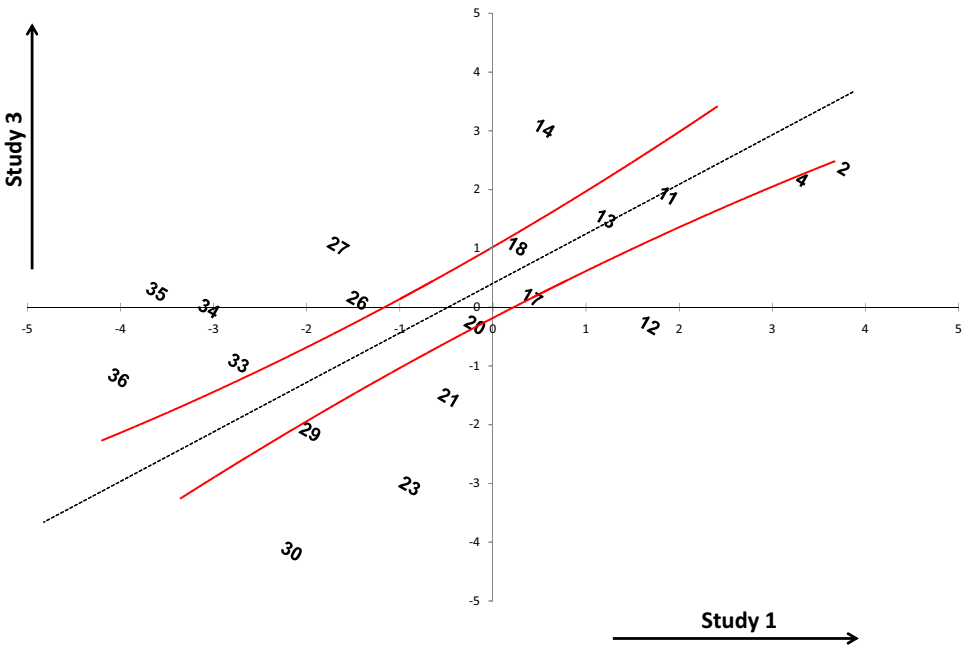


Fig. 5 Item invariance plot of the item difficulties in Studies 1 and 3. Each number represents an item. Red lines indicate 95% confidence intervals. (Color figure online)

for each of these four player types. Although all correlations between the four sets of item difficulties were significant (see Table 4), some of the characteristics were clearly different in their probability of being ascribed to specific agents.

Figure 6 displays the item invariance plots between each of the agent types. As can be seen in this Figure, three of the 19 items (i.e., items 18 'jump', 34, 'pick up objects', and 35 'walk') were consistently less likely to be ascribed to algorithms and computers than to robots and humans. This should not come as a surprise, as computers and algorithms lack the morphology that accommodates such physical activities. This finding revealed that a comparison between the agent types is unfair on specific attributes. Using a computer's ability to perform physical features as a measurement of its human-likeness is similar to judging people who are bound to a wheelchair as less human than their able counterparts for not being able to jump.

4.2.3 Sensitivity in Differentiating Between Agents

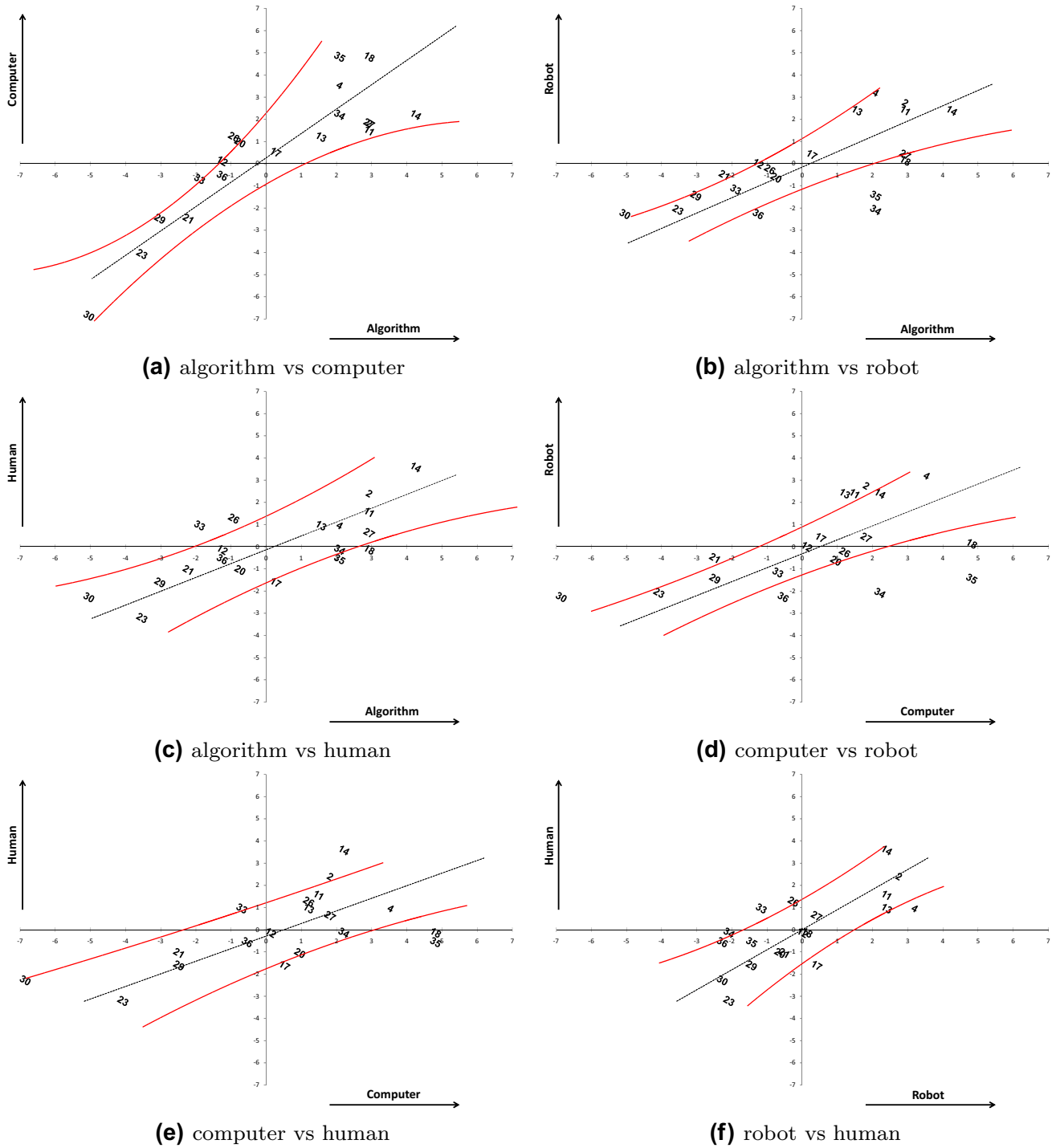
To test the extent to which the anthropomorphism scale differentiates humans from other player types, a one-way Analysis of Variance (ANOVA) was conducted with agent type as independent variable and the individual predisposition to anthropomorphize as dependent variable. Results indicated a statistically significant effect of agent type, $F(3, 201) = 42.04$, $p < .001$, $\eta^2 = .39$. More specifically, anthropomorphism was highest for humans ($M = 3.50$, $SD = 2.35$), followed by robots ($M = -0.63$, $SD = 2.81$), algorithms ($M = -1.53$, $SD = 2.82$) and computers ($M = -1.59$, $SD = 2.29$), see Fig. 7a.

Pairwise comparisons (LSD) showed a statistically significant difference between humans and all other agents, $t(198) = 11.03$, $p < .001$, $d = 1.89$. Post-hoc comparisons using Bonferroni correction indicated that differences between any two groups other than human were not significant.

Interestingly, after removing the three morphology-related items (i.e., items 18, 34, and 35) from the scale, the differences in anthropomorphism between non-human player types became smaller, with estimations being highest for humans ($M = 3.48$, $SS = 2.34$), followed by robots ($M = -0.83$, $SD = 2.90$), algorithms ($M = -1.32$, $SD = 2.83$) and computers ($M = -1.35$, $SD = 2.37$), see Fig. 7b. This could be caused by the nature of the experiment, because all players behaved in the exact same way during the Ultimatum Game. When those players were subsequently compared with respect to characteristics that they were equally likely to possess based on the behavior that they showed, it should not come as a surprise that no differences between them were found.

Table 4 Correlations between item difficulties in the four experimental conditions in Study 3

	Human		Robot		Computer	
Algorithm	$r = .74$	$p < .001$	$r = .68$	$p < .01$	$r = .88$	$p < .001$
Computer	$r = .59$	$p < .01$	$r = .50$	$p = .03$		
Robot	$r = .73$	$p < .001$				



(a) algorithm vs computer

(b) algorithm vs robot

(c) algorithm vs human

(d) computer vs robot

(e) computer vs human

(f) robot vs human

Fig. 6 Item invariance plots of the item difficulties of each of the four experimental conditions in Study 3: algorithm, computer, robot, and human. Each number represents an item. Red lines represent 95% confidence intervals. (Color figure online)

4.3 Conclusions

In the current study, an adjusted 19-item version of the anthropomorphism scale was used to investigate the scale's sensitivity. Results showed that, as in Studies 1 and 2, people's responses sufficiently fitted the Rasch model, indicated by an acceptable data-to-model fit. This result supported the expected invariant ordering of the human-like characteristics with respect to their item difficulty.

An adequate level of sensitivity was found. The scale was able to differentiate humans from different types of technological players, but it did not differentiate those players from

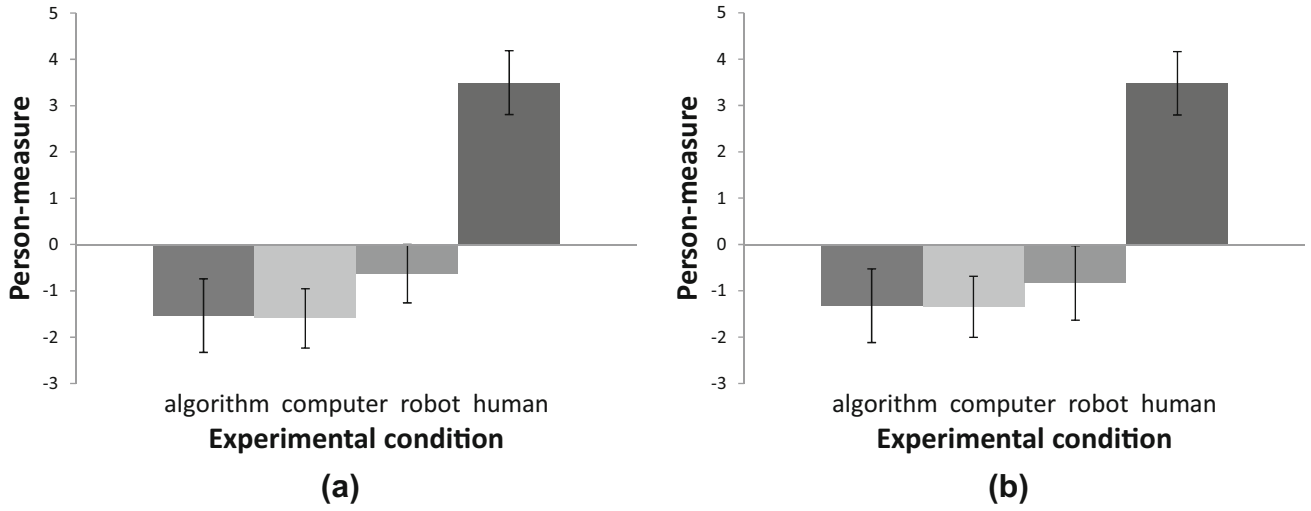


Fig. 7 Visualization of averaged person measures on the scale **a** before and **b** after removing the biased items in Study 3. Whiskers represent 95% confidence intervals

each other. Presumably the conceptual differences between those player types were too small to be detected by the current version of the scale. Additionally, the scale's sensitivity dropped when three morphology-related items were removed from the analysis.

5 General Discussion

The current research was designed to explore whether anthropomorphism can be successfully measured using the Rasch model, whether the concept can be mapped onto a one-dimensional scale, and whether human-like characteristics are ordered in a way that is similar for all individuals in their encounter with robots and other types of agents in different contexts. We argued that human-like characteristics can be ordered according to the probability with which they are ascribed to robots, and that this ordering of human-like characteristics on the range of perceived human-likeness is similar for all individuals in their encounter with different types of agents. We developed a set of human-like characteristics and used data from three studies to test the scale's psychometric qualities. These studies had designs with differing contexts and experimental conditions, and people's predispositions to anthropomorphize were compared in different samples with different types of agents.

In the first study, we hypothesized that items and persons could be mapped onto a single one-dimensional scale, and that items would be invariantly ordered according to the difficulty with which they are attributed to a robot. Additionally, the estimated difficulties with which the 37 characteristics are attributed to a robot were expected to be related to their perceived human nature and human uniqueness. In the second study, we tested the extent to which estimates made with three different measuring instruments for anthropomorphism would be related. In the third study, we hypothesized that the anthropomorphism scale would successfully differentiate humans from other types of agents.

All hypotheses were (at least partially) confirmed, and the next sections describe in more detail the dimensionality of anthropomorphism, implications for further research, and the comparison of the three different measuring instruments.

5.1 Dimensionality of Anthropomorphism

Across studies, an invariant ordering of human-like characteristics was found, indicating that this ordering was similar for different people in their encounter with different types of agents in different contexts. This finding was supported by dimensionality tests that consistently indicated that the data could be represented in a one-dimensional structure. More specifically, in each of the studies, an additional factor would result in only a small increase in the proportion of explained variance. This result supported the hypothesis that anthropomorphism can be represented as a one-dimensional construct.

Tests of construct validity showed significant correlations between item difficulties and their perceived human nature and human uniqueness, supporting the expectation that the scale measures anthropomorphism. Together, these findings indicate that human-like characteristics are ordered in such a way that they range from low to high on a single dimension.

5.2 Implications for Further Research

Findings on some of the characteristics have important implications for future research and thus need some further

consideration. For example, in Study 1 the item 'Experience pain' was rated as extremely low in human uniqueness and as medium in human nature, but it appeared to be the most difficult one to ascribe to a robot. One possible explanation for this unexpected finding could be that a nervous system is necessary for experiencing pain, which is not unique for humans, but is something that robots clearly do not have.

This result also raises the issue of physical versus cognitive capacities. For humans and other organisms, the ability to move around in the (physical) environment is a given, whereas for certain technological artifacts this may not be so obvious. Likewise, future artificial intelligence may create agents with high mental capacities without a body, which makes it easier for them to solve a moral dilemma than to pick up an object. This issue may become apparent in the near future, and more research is needed to further investigate this. A large benefit of the Rasch approach is the ability to select items for a specific design, and as such can cope with such technological developments.

Another important finding was that when morphology-related characteristics (i.e., the items about a robot being able to jump, walk, and pick up objects) were disregarded in Study 3, the differences on anthropomorphism between robots, computers, and algorithms decreased. Future research can be designed to explore whether we can create a measure for anthropomorphism that is universal for various types of agents (e.g., humans, animals, technologies, and deities).

5.3 Comparison of Measuring Instruments

The anthropomorphism scale was compared with two available measuring instruments for anthropomorphism to test for convergent validity. These available instruments were the Godspeed- and Waytz-instruments. We tested whether measurements obtained with the three instruments would be related. This hypothesis was confirmed by significant correlations between the scale and both other instruments, but these correlations were rather low.

Although these correlations were smaller than expected, this should not come as a surprise, given the nature of the different instruments. They were all developed with different views on the concept anthropomorphism. The Godspeed-instrument focuses on mostly appearance-related features, whereas the Waytz-instrument focuses mostly on cognition-related features.

The low correlation also clearly shows that we still have limited understanding of what anthropomorphism entails, and thus what indicators are best used in its measurement. It is only when multiple different measuring instruments converse that we can claim to fully grasp what the indicators of anthropomorphism are. We believe that the Rasch model is a promising tool to uncover such indicators and thus what anthropomorphism truly entails.

5.4 Limitations and Future Research

In all studies, the items of the anthropomorphism scale were ordered alphabetically, which may have influenced people's responses to certain items because of order effects. The items on the scale should be ordered randomly in future studies to prevent the occurrence of such ordering effects.

All studies were performed with mainly Dutch and American participants, so their cultural backgrounds and experiences with technology could have been quite similar. Earlier work has shown that people's tendencies to attribute human-likeness to non-humans can be related to religion [14], and that people from different countries (such as individualistic versus collectivistic ones) respond differently to computers [19] and evaluate robots differently [28]. It would be interesting to investigate whether cultural differences influence people's predispositions to anthropomorphize.

In addition, we used social media for sampling participants in studies 1 and 2, and thus relied on data gathered from experimenters' acquaintances. This may have led us to end up with homogeneous groups of participants, which could partly explain the high consistency between those studies. For increasing our understanding of the concept anthropomorphism, it is important to also collect data from more diverse groups and check whether the high consistency prevails.

In Study 2, each statement was presented as a general statement about 'a robot', and not specifically aimed at the robot that participants watched in the short movie clip. This could explain why no differences in anthropomorphism were found between the two robots. Future research that is designed to investigate evaluations of different robots should therefore not use the scale in a general way, but rather phrase each question to be specifically about a certain robot (or other non-human object).

We did not test whether people had previous knowledge about or experience with robots. This experience could influence people's responses to those robots (see e.g., [21,25]) and should therefore be taken into account in future studies. The Rasch model offers a promising approach to investigate how such experiences affect a person's predisposition to anthropomorphize, the difficulty of ascribing a specific characteristics to a robot, or both.

5.5 Conclusions

Despite some limitations, we have shown that anthropomorphism can be measured on a one-dimensional scale, and that items are ordered invariantly when this scale is applied to robots. The Rasch model thus provides a reliable way of measuring anthropomorphism. Because of the invariant ordering of the human-like characteristics, the scale provides opportunities for comparing various types of agents and robots with each other across studies. The method also provides the

possibility to select items based on the context of the study, making it a versatile tool for measuring anthropomorphism.

Findings of Study 3 showed that different types of agents (except humans) were difficult to distinguish, presumably because the behavior of these different agents was identical. It is therefore an interesting question how the method performs when *interactions* with different types of agents are evaluated. Ultimately, a scale can be developed that contributes to our understanding of what makes people attribute human-like characteristics to social robots and other non-humans. This understanding can help designers to create robots as social entities that are accepted as members of our society.

A Waytz-Instrument

Items of the anthropomorphism questionnaire adapted from [32], answered on a 5-point or a 7-point response format.

"To what extent does the robot have thoughts of its own?"
"To what extent does the robot have intentions?"
"To what extent does the robot have a free will?
"To what extent does the robot have a consciousness?"
"To what extent does the robot have desires?"
"To what extent does the robot have values and norms?"
"To what extent does the robot experience emotions?"

B Godspeed-Instrument

The items of the anthropomorphism part of the Godspeed questionnaire (adapted from [2]), answered on a 5-point or a 7-point response format.

"Fake - Natural"
"Machinelike - Humanlike"
"Unconscious - Conscious"
"Artificial - Lifelike"
"Moving rigidly - Moving elegantly"

References

1. Airenti G (2015) The cognitive bases of anthropomorphism: from relatedness to empathy. Int J Soc Robot 7(1):117–127
2. Bartneck C, Kulić D, Croft E, Zoghbi S (2009) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J Soc Robot 1(1):71–81
3. Bond TG, Fox CM (2013) Applying the Rasch model: fundamental measurement in the human sciences. Psychology Press, Portland
4. Bruce A, Nourbakhsh I, Simmons R (2002) The role of expressiveness and attention in human–robot interaction. In: Proceedings of IEEE international conference on 2002 robotics and automation. ICRA'02, vol 4. IEEE, pp 4138–4142
5. Carpinella CM, Wyman AB, Perez MA, Stroessner SJ (2017) The robotic social attributes scale (rosas): development and validation. In: Proceedings of the 2017 ACM/IEEE international conference on human–robot interaction. ACM, pp 254–262
6. Charles EP (2005) The correction for attenuation due to measurement error: clarifying concepts and creating confidence sets. Psychol Methods 10(2):206–226
7. Chidambaram V, Chiang YH, Mutlu B (2012) Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In: Proceedings of the seventh annual ACM/IEEE international conference on human–robot interaction. ACM, pp 293–300
8. Duffy BR (2003) Anthropomorphism and the social robot. Robot Auton Syst 42(3):177–190
9. Epley N, Waytz A, Cacioppo JT (2007) On seeing human: a three-factor theory of anthropomorphism. Psychol Rev 114(4):864–886
10. Eyssel F, Reich N (2013) Loneliness makes the heart grow fonder (of robots): on the effects of loneliness on psychological anthropomorphism. In: 8th ACM/IEEE international conference on human–robot interaction (HRI), 2013. IEEE, pp 121–122
11. Eyssel F, Hegel F, Horstmann G, Wagner C (2010) Anthropomorphic inferences from emotional nonverbal cues: a case study. In: RO-MAN, 2010 IEEE. IEEE, pp 646–651
12. Eyssel F, Kuchenbrandt D, Bobinger S (2011) Effects of anticipated human–robot interaction and predictability of robot behavior on perceptions of anthropomorphism. In: Proceedings of the 6th international conference on human–robot interaction. ACM, pp 61–68
13. Güth W, Schmittberger R, Schwarze B (1982) An experimental analysis of ultimatum bargaining. J Econ Behav Organ 3(4):367–388
14. Guthrie S (1993) Faces in the clouds: a new theory of religion. Oxford University Press, Oxford
15. Haans A, Kaiser FG, Bouwhuis DG, IJsselsteijn WA (2012) Individual differences in the rubber-hand illusion: predicting self-reports of people's personal experiences. Acta Psychol 141(2):169–177
16. Haslam N, Bain P, Douge L, Lee M, Bastian B (2005) More human than you: attributing humanness to self and others. J Personal Soc Psychol 89(6):937–950
17. Haslam N, Loughnan S, Kashima Y, Bain P (2008) Attributing and denying humanness to others. Eur Rev Soc Psychol 19(1):55–85
18. Kaiser FG, Merten M, Wetzel E (2018) How do we know we are measuring environmental attitude? Specific objectivity as the formal validation criterion for measures of latent attributes. J Environ Psychol 55:139–146
19. Katagiri Y, Nass C, Takeuchi Y (2001) Cross-cultural studies of the computers are social actors paradigm: the case of reciprocity. Usability evaluation and interface design: cognitive engineering, intelligent agents, and virtual reality, pp 1558–1562
20. Kennedy JS (1992) The new anthropomorphism. Cambridge University Press, Cambridge
21. Lemaignan S, Fink J, Dillenbourg P, Braboszcz C (2014) The cognitive correlates of anthropomorphism. In: Proceedings of the 2014 ACM/IEEE international conference on human–robot interaction
22. Linacre JM (2003) Data variance: explained, modeled and empirical. Rasch Meas Trans 17(3):942–943

23. Linacre JM (2006) Misfit diagnosis: infit outfit mean-square standardized. www.winsteps.com/winman/diagnosingmisfit. Accessed 1 Oct 2018
24. Meijer RR, Sijtsma K (2001) Methodology review: evaluating person fit. Appl Psychol Meas 25(2):107–135
25. Ruijten PAM, Cuijpers RH (2017) Dynamic perceptions of human-likeness while interacting with a social robot. In: Proceedings of the companion of the 2017 ACM/IEEE international conference on human–robot interaction. ACM, pp 273–274
26. Shaver P, Schwartz J, Kirson D, O'Connor C (1987) Emotion knowledge: further exploration of a prototype approach. J Person Soc Psychol 52(6):1061–1086
27. Shaw F (1991) Fits about "misfit". Rasch Meas Trans 5(1):132
28. Shibata T, Wada K, Ikeda Y, Sabanovic S (2009) Cross-cultural studies on subjective evaluation of a seal robot. Adv Robot 23(4):443–458
29. Smith EV Jr (2002) Understanding rasch measurement: detecting and evaluating the impact of multidimenstionality using item fit statistics and principal component analysis of residuals. J Appl Meas 3(2):205–231
30. Syrdal DS, Dautenhahn K, Walters ML, Koay KL (2008) Sharing spaces with robots in a home scenario-anthropomorphic attributions and their effect on proxemic expectations and evaluations in a live HRI trial. In: AAAI fall symposium: AI in Eldercare: new solutions to old problems, pp 116–123
31. Waytz A, Cacioppo J, Epley N (2010a) Who sees human? The stability and importance of individual differences in anthropomorphism. Perspect Psychol Sci 5(3):219–232
32. Waytz A, Morewedge CK, Epley N, Monteleone G, Gao JH, Cacioppo JT (2010b) Making sense by making sentient: effectance motivation increases anthropomorphism. J Person Soc Psychol 99(3):410–435
33. Wright BD (1977) Solving measurement problems with the rasch model. J Educ Meas 14(2):97–116
34. Wright BD, Linacre JM (1987) Dichotomous rasch model derived from specific objectivity. Rasch Meas Trans 1(1):5–6
35. Wright BD, Linacre JM, Gustafson JE, Martin-Lof P (1994) Reasonable mean-square fit values. Rasch Meas Trans 8(3)
36. Young JE, Sung J, Voida A, Sharlin E, Igarashi T, Christensen HI, Grinter RE (2011) Evaluating human–robot interaction. Int J Soc Robot 3(1):53–67
37. Zawieska K, Duffy BR, Strońska A (2012) Understanding anthropomorphisation in social robotics. Pomiary Automatyka Robotyka 16:78–82
38. Złotowski J, Strasser E, Bartneck C (2014) Dimensions of anthropomorphism: from humanness to humanlikeness. In: Proceedings of the 2014 ACM/IEEE international conference on human–robot interaction. ACM, pp 66–73

Peter A. M. Ruijten is Assistant Professor of the Human-Technology Interaction Group at Eindhoven University of Technology. He has a Bachelor in Electrical Engineering and a Master in Human-Technology Interaction. His research interests are Social Robots, Anthropomorphism, Behavior Change, and applications of this in Human-Inspired Machines.

Antal Haans is assistant professor in environmental psychology in the Human-Technology Interaction group at Eindhoven University of Technology. His research focuses on the interplay between humans and their surroundings—including built and media environments—in explaining human experience and performance. Research interests include environmental perception, presence in VR, mediated social touch, psychological measurement, and (smart) urban lighting applications.

Jaap Ham is associate professor at Eindhoven University of Technology. He studies theory-driven and problem-driven questions focusing on how technology can change human behavior and thinking. He acquired research grants and published papers in journals and conference proceedings on Persuasive Technology, Persuasive Social Robotics, Ambient Persuasion, and VR for changing sustainability behavior and health behavior.

Cees J. H. Midden is em. professor at Eindhoven University of technology specializing in the interaction between humans and technological systems. He received a degree in Psychology at the University of Leiden. His research interests concern the social and cognitive factors of human-technology interactions as these become apparent in the consumption and use of products and systems. He published on persuasive technology and behavior change, social robotics, environmental consumer behavior and the perception and communication of technological risks.