Apples and Oranges: Validity and Reliability of the Three Main Anthropomorphism Measures

Merel Keijsers[1](✉), Luisa Porzio[1], Anthony Tricarico[1], Daniel J. Rea[2], Stela H. Seo[3], and Takayuki Kanda[3]

[1] John Cabot University, Rome, Italy
mkeijsers@johncabot.edu
[2] University of New Brunswick, Fredericton, Canada
[3] Kyoto University, Kyoto, Japan

Abstract. Anthropomorphism is a key construct in social human-robot interaction, with a variety of measurements. Our experiment compares the three most used scales on construct and convergent validity as well as reliability. We used a mixed methods design where participants first viewed a video featuring either a Sota or a TIAGo robot, and then completed each scale with randomisation of the order of presentation. Results indicate that the Mind Attribution scale (Gray et al., 2007) and the Individual Differences in Anthropomorphism questionnaire (Waytz et al., 2010) experienced some order effects and appear to be founded on a different latent process than the revised Godspeed questionnaire (Ho & MacDorman, 2010; Bartneck et al., 2009), though the different scales yield mostly comparable results.

Keywords: Anthropomorphism · Validity · Reliability

1 Introduction

Mind attribution, humanlikeness, or anthropomorphism has been a point of focus within social human-robot interaction (HRI)[1], whether as a predictor, manipulation, control or dependent variable. And with reason: perceiving a robot as possessing (some extent of) humanness is considered the starting point to various outcomes, in terms of cognition – e.g. trust [9] and moral agency [16] – affect – e.g. empathy [13] – and behaviour – e.g. prosocial [23] and bullying behaviour [20]. Scientists have operationalised anthropomorphism in various ways, including measurements of social behaviour [41], analysis of participants' descriptions of robots [1], or the extent to which they apply (human) stereotypes to the robot [10]. The most common way, however, is explicit self-report through a scale or questionnaire, where participants indicate how much they see the robot

[1] [42] notes that these terms are often used interchangeably; this paper will treat them as such.

O. Palinko et al. (Eds.): ICSR + AI 2024, LNAI 15562, pp. 429–442, 2025.
https://doi.org/10.1007/978-981-96-3519-1_39

as possessing a variety of attributes that all constitute psychological anthropomorphism.

These self-report scales may differ in how sensitive they are to certain variables (e.g., embodiment), capture variants of the same concept (e.g., robot behaviour as voluntary vs programmed), or be more or less precise in their outcomes, both within the scale (e.g. standard error) and compared to other measures (e.g. shared variance with other scales). This kind of information is fundamental for study design as it informs researchers which questionnaire best fits their research objectives, and what order questions should be presented in; and also for interpreting the relation between multiple findings or performing meta-analyses. While conceptual comparisons have been made (such as discussing to what extent they aim to capture the same construct [42], contrasting explicit and implicit measures [37,46], or to validate new measures of the construct [32]), we do not know of any study that statistically compares existing anthropomorphism scales against another with psychometrics.

Thus, we compare the three most popular anthropomorphism scales, analysing convergent and content validity as well as reliability: the Mind Perception scale (MP; [15]), the humanlikeness subscale from the revised Godspeed questionnaire (rGS; [17]), and the adapted Individual Differences in Anthropomorphism Questionnaire (IDAQ; [44]). Specifically, we compare the impact of context and robot design on measurement means, standard errors, and internal consistency.

1.1 Measurement Reliability and Validity

As a latent variable, anthropomorphism can be operationalised in various ways (see 1.2). Each of these ways aim to measure the same construct, but since this construct is fundamentally unobservable there is always the question of to what extent these operationalisations cover anthropomorphism exactly and to its full extent, i.e. their validity. Questions of validity do not necessarily mean a measure is unusable, but could also suggest that for example two measures focus on different aspects of the concept, such as a robot having a humanlike appearance versus it having hopes and dreams, leading to (slightly) different anthropomorphism scores. Even if the scores are the same, one of the two may show more variability (e.g., because participants find it easier to decide if the robot looks humanlike, than whether it can hope and dream), or be affected by different cues (for example the ability to express emotion may increase anthropomorphism in both scales, but more so for the *hopes and dreams*-based measure). This kind of validity does not reject measures, but highlights the subtle yet important differences in how measures can or should be interpreted.

When comparing measures on validity a good starting point is reliability, i.e. the extent to which measures are consistent, as this is required (but not sufficient) for validity [14]. Within scales reliability is often measured through Cronbach's alpha, which ranges from 0 to 1 and indicates the co-variation between different items. Considering that each question targets the same latent construct but does so from its own unique perspective, a questionnaire should score between .7 and

.95 [40]. In addition to ascertaining that individual scales meet this requirement, one can consider how much the scales' reliability is affected by external factors; for example item randomisation or different kinds of robots. Ideally, a scale that is reliable for robot A will have similar reliability for robot B; and if item randomisation affects reliability levels, this suggests that item interpretation may be dependent on context [27]. Finally, a measure's variance can be compared to that of other measures; even if two means are similar, the extent to which all individual data points cluster around that mean may differ. With greater variance, the standard error goes up as well, which indicates greater measurement imprecision [12]. Thus, comparing variances of the different measures, as well as their susceptibility to randomisation and different robots, is useful as an indication of reliability.

If measures are reliable and have comparable variance, they may still not represent anthropomorphism in the same way. This issue touches upon convergent validity and content validity, which are concerned with the extent to which different measures of the same construct give comparable outcomes [4], and the extent to which the scales measure unique aspects of a construct in all their variety [36], respectively. If the same scaling is employed for two questionnaires, convergent validity can be tested by comparing the scores of different scales on the same robot. It may again be beneficial to include a variety of robots and randomisation to test the robustness of these measures. If the same robot is scored differently on two anthropomorphism measures in spite of a similar scaling, then clearly a 3 on measurement A does not mean the same thing as a 3 on measurement B. Note that this does not mean directly that the measure is invalid altogether (a size M t-shirt in the USA may be much larger than the same size t-shirt in Japan, even though the t-shirts in question are perfectly indicative of the medium size in their respective countries), but at the very least it warrants caution when comparing measurements (an American tourist in Japan may want to go a size up from their usual when buying a t-shirt) and it is reason for further investigation.

To further interpret any differences, the shared variance between measurements can be taken into account: how much of the variability in measure A can be explained through measure B? This approach involves content validity, as two measures that consider the same aspects of anthropomorphism will share more variance with one another. A more detailed investigation is to check for similarity in latent (sub)variables. If the scales base themselves on similar aspects of anthropomorphism, this would show in an overlap of the factors identified between the measurements. For both content and convergent validity, finding that measurements do not match does not directly mean that either or both should not be considered an appropriate measure of anthropomorphism. It does however indicate that they target different aspects of the construct, and that depending on the research question one may be more appropriate than the other.

1.2 Anthropomorphism and Its Operationalisations

Anthropomorphism is believed to be the result of a (series of) latent process(es), which form the starting point to a range of social perceptions and behaviours such as activation of social schemas [10], morality [16], and trust [9]. Theoretical models differ on whether they conceptualise anthropomorphism along a single or multiple dimensions; as the result of a single or multiple processes; and in the case of multiple, whether these are in- or interdependent.

For example, [8] proposes a single dimension of psychological anthropomorphism as the result of three separate processes, one cognitive (activation of schemas) and two motivational (the need to be social and the need for effectance). Dispositional, contextual, and agent qualities influence each of the three processes. Cues from the robot's appearance, name, or behaviour may activate schemas associated with humans (cognitive component; [30,33]); the (un)predictability of their behaviour triggers the user's need to interact with the world in an efficient manner, which requires predictability of and control over the course of events (i.e. effectance; [34]); and their presentation as social agents may tap into the user's need to have meaningful social interactions (i.e. sociability; [26]). Thus, all processes are set in motion through robot's appearance and behaviour. While the model explicitly acknowledges the influence of dispositional and experiential qualities of the user as well as circumstantial factors [8], we also focus on robot appearance.

A similar model [22] proposes how visual aspects of the robot (e.g., physical features and movements) are precursors, and attribution of personality and morality are the consequences of anthropomorphism, with the subjective attribution of mental qualities (i.e. cognition and emotion) as the construct itself. This process of "mind perception" may be applied to human as well as non-human agents [15], and could be the foundation for cognitive and affective responses such as empathy, morality, and trust [16]. This model is two-dimensional, with attributions of affective states ("*Experience*") independent of attributions of cognitive abilities ("*Agency*") [15]. With appearance being the precursor rather than part of the construct, it is not included in their scale.

The Role of Robot Appearance. While robots can activate specific (social) schema through their presentation as for example feminine or masculine [10,29], embodiment seems to have surprisingly little influence on their perceived status as social agents [28,38], nor does their appearance seem to influence whether humans follow general social norms around them [24]. This is remarkable as robot appearance is theorised to be either an important starting point or a facet of anthropomorphism, and it suggests that social cues may be mostly inferred from aspects such as verbal and non-verbal behaviour, movement, and backstory.

Not all anthropomorphism measurements take into account a robot's physical attributes, which means that if robot appearance indeed plays at most a minor role, scales that focus strongly on humanlikeness may show a larger measurement error. If robot appearance does play a role after all, two different looking robots should get different anthropomorphism scores on the same scale. Thus, the current research included a manipulation of appearance through the use of two different of robots: the small humanoid Sota and the larger, more industrial looking TIAGo (Fig. 1).

The Role of Measurement and Item Order. Item order can influence participant's answers, as previous questions may set an expectation or mental framework that influences the answers to the later questions [21,35]. This is especially the case if the questionnaire contains items that require more careful consideration due to, e.g., ambiguity or complexity [21], or if questions build on another [27]. Thus, randomisation of items can lead to lower scores and greater measurement error [45]. This same reasoning can be extended to the multiple entire measures; if the score depends on its order relative to another measure, that implies suggestibility in the scale.

1.3 The Current Study

We selected the three most commonly used anthropomorphism scales in social HRI [42], which differ in terms of focus (humanlike appearance versus the process itself, i.e. ascription of human traits) and in dimensionality. The two-dimensional Mind Perception (MP; [15]) scale uses 18 items to measure the robot's cognitive and emotional capabilities, but omits measures of animism such as how "alive" or "mechanic" an agent is. The Individual Differences in Anthropomorphism Questionnaire (IDAQ; [44]) in principle measures dispositional anthropomorphism, but is often modified to a specific agent (e.g., [39,43]). Like MP, the IDAQ measures an agent's cognitive and emotional capabilities, but at 5 unique items (when applied to a specific target) it is considerably shorter. Finally, the "humanlikeness" subscale of the revised Godspeed questionnaire (rGS; [17]), based on the Godspeed questionnaire [2], asks in 6 items directly about the robot's qualities in terms of appearing living, organic, and humanlike (as opposed to inanimate, mechanical, human-made).

The current study aims to compare the MP, IDAQ, and rGS on their convergent and content validity as well as their sensitivity to contextual influences (specifically appearance and randomisation of items and scales). Participants watched a video where one of two different robot types was introduced, and afterwards completed the three anthropomorphism scales, with randomsiation of the order in which the scales were presented, and whether the item order within each scale was randomised. To avoid a study design with 54 conditions, we applied full randomization (order randomization of both surveys and items within them) to only two of the surveys: MP and rGS, thus limiting the design to 24 conditions (two robot embodiments, randomisation within either MP or rGS, first presentation either MP or rGS, and three different measures of anthropomorphism). We selected these two measurements since, at face value, they were the most different and because MP is the longest questionnaire and thus has the largest chance of order effects.

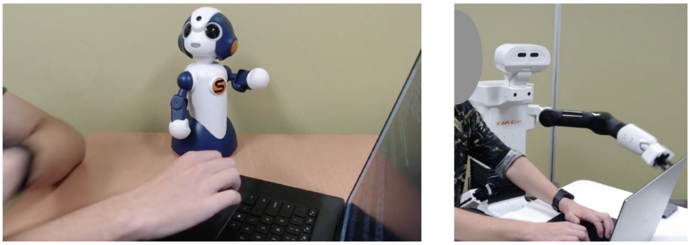

Fig. 1. Images of human-robot collaboration from our study videos. Left: Sota, Right: TIAGo.

2 Methods

2.1 Participants

We recruited 254 participants from English-speaking countries from Amazon MTurk's Masters, an elite subsection of Amazon Workers who maintain a high quality of work. Two participants were removed due to failing the attention checks, resulting in a sample of 252. This sample size has a power of 0.95 to detect differences of effect size $R^2 = .014$ or above at $\alpha < .05$ [11].

The sample had a mean age of 46.4 years ($s = 11.26$, range = 26–76), with similar proportions of men ($N = 124$, 49.2%) and women ($N = 123$, 48.8%) and a minority of nonbinary ($N = 2$, 0.8%) and non-disclosing participants (1.1%). Most participants ($N = 239$, 94.8%) identified as American, with only a small proportion specifying Asian American or African American (each $N = 4$, 1.6%).

2.2 Materials

Measurements. We administered three anthropomorphism measurements in addition to attention checks and participant demographics (i.e. age, gender, and nationality). See also Table 1.

The *Mind Perception scale (MP)* [15] is a measure of psychological anthropomorphism consisting of 18 items scored on a 7-point Likert scale. It measures to what extent an agent is perceived to be able to *experience* the world and its own feelings (e.g. "pleasure", "pain", "desire") , as well as its capacity for *agency* (e.g. "thought", "rationality", "self-awareness").

The *revised Godspeed questionnaire (rGS)* [17] assesses an agent's perceived attractiveness, eeriness, and humanness. For this study, only the humanness subscale was used, which consists of 6 contrapositions to be rated on a 7-point Likert scale. For example, participants are asked to indicate whether they consider the robot to be more *inanimate* or *living*, *synthetic* or *real*, *artificial* or *natural*, with the scores of 1 and 7 indicating either extreme and 4 at the midpoint.

The original *Individual Differences in Anthropomorphism Questionnaire (IDAQ)* consists of 15 items, and asks participants to indicate to what extent a variety of non-human entities (e.g. a cheetah, a mountain, a car) is capable of a variety of cognitive and affective responses (i.e. having consciousness, free will, and intentions) [44]. To measure a specific agent's anthropomorphism the entities are commonly replaced by the agent in question, and the recurring items removed, resulting in a questionnaire of five items measured on a 7-point Likert scale.

Two *attention checks* were included: one asking participants to identify the robot from the video, and one (embedded within the IDAQ) requiring the participant to respond with a specific value.

Video Stimuli. To ensure that any effects found could be attributed only to robot appearance, the videos were identical in length, text-to-voice narration, robot behaviours, background, and human collaborator (shown peripherally). Thus, the only difference was that either the TIAGo or the Sota robot was used (see Fig. 1). The robot behaviours included the robot waving at the camera, the robot picking up a water bottle and holding it up for someone to take, and the robot pointing out something on a laptop while the human was working on it. The narration was done from the robot's perspective, who introduced itself, and shared some general facts about robots and the specific tasks it helped out with (e.g. "I help out with coding, and bring water bottles to those who could use a refreshment"). The videos were 1 min and 10 s long.

2.3 Procedure

Ethics approval was obtained through the John Cabot University review board (007–23). The experiment advertisement on MTurk offered participants 2 US$ for completing an 8 min survey on Qualtrics. On average, participants took 6.2 min (s = 3.2). After being given an information sheet and providing consent, the participant was shown a one-minute video introducing either the TIAGo or Sota robot (see Fig. 1) before filling out three measures of mind attribution (MP, rGS, and IDAQ). Both the order in which these questionnaires were presented (MP followed by rGS, or rGS followed by MP) and which of the questionnaires had its items randomised (MP or rGS), were randomised between participants. Full randomisation including the IDAQ was not realised due to the consequences for study design complexity and the required number of participants. Finally, participants completed demographics, the second attention check and were given the opportunity to leave comments on the study before receiving the MTurk compensation code.

Table 1. Descriptives of the anthropomorphism measures

	mean (*sd*)	median	min	max	se	# items	Cronb. α
Mind Perception	2.43 (*0.90*)	2.39	1.00	5.72	0.06	18	.93
Revised Godspeed	2.25 (*1.08*)	2.00	1.00	5.67	0.07	6	.88
IDAQ	1.95 (*1.10*)	1.60	1.00	5.80	0.07	5	.87

3 Results

3.1 Preliminary Checks

The conditions did not differ in terms of number of respondents, $\chi(7\ ) = 8.83$, $p =$ .265, or gender distribution, $\chi(21\ ) = 20.32$, $p = .501$, or age, $F(7, 243) = 1.21$, $p = .296$. Thus, randomisation was considered successful. Internal reliability of the measurements was excellent for Mind Perception (MP; $\alpha = .93$) and good for the revised Godspeed (rGS; $\alpha = .88$) and the Individual Differences in Anthropomorphism Questionnaire (IDAQ; $\alpha = .87$) [40].

3.2 Measurement Differences in Variance and Reliability

To test the effects of our conditions on the internal reliability of the scales, we used the procedure as proposed by [7] – a test statistic for comparison of multiple alpha coefficients based on the chi-distribution. We found no influence of randomisation or robot appearance on reliability: MP $\chi(7\ ) = 10.59$, $p = .158$; rGS $\chi(7\ ) = 13.71$, $p = .057$; and IDAQ $\chi(7\ ) = 10.58$, $p = .158$. In addition, Levene's test was used to test whether the measures' variance differed between the conditions. This was not the case. For the MP, $F(7, 243) = 0.80$, $p = .587$; for the rGS $F(7, 244) = 1.31$, $p = .245$; and for the IDAQ $F(7{,}244) = 1.23$, $p =$ .288. Finally, the variances also did not differ significantly between the different measurements, $F(23, 729) = 1.27$, $p = .182$. Together, these results indicates that the measurement error in different scales is stable across conditions; in other words, the measurement's precision was unaffected.

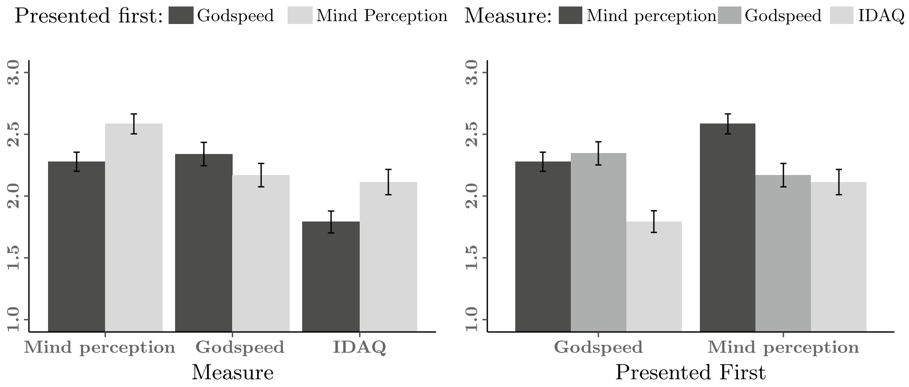

Fig. 2. Means and standard errors of the three different measurements, split out against either measurement (a) or which measure was presented first (b).

3.3 Measurement Differences in Scores

Following the method as proposed by [12], who uses multilevel modeling for mixed designs, a 3 × 2 × 2 × 2 mixed ANOVA was defined as a multilevel model with measurement type (MP, rGS, IDAQ) nested under the participant level; and robot type (Sota or TIAGo, see Fig. 1), item randomisation (items randomised within a measurement for MP or rGS), and measurement randomisation (MP or rGS presented first) as between-participant variables. Starting from a baseline model with no predictors, the model was built up one variable at a time; the resulting series of models were compared to another through chi-square tests to see which variables significantly improved predictions.

The model improved significantly with the addition of a main effect for measurement type, $\chi(2) = 69.69$, $p < .001$, and an interaction effect between the measurement type and measurement randomisation, $\chi(2) = 24.69$, $p < .001$, meaning we should not directly interpret the main effects. A post hoc with Benjamini-Hochberg correction [3] found MP scores were significantly higher if the MP scale was presented before all other conditions, all ts > 2.74, all p_{bh} $< .031$, mean Cohen's $d = .515$; with exception of the rGS if it was presented first, $t(244) = 1.93$, $p_{bh} = .093$. Moreover, IDAQ anthropomorphism was significantly lower than all conditions if the rGS was presented first, all ts > 2.37, all $p_{bh} < .024$, mean Cohen's $d = .436$; see Fig. 2. Remarkably, robot type was non-significant as main effect $t(247) = 0.55$, $p_{bh} = .584$, or as interaction, ts $<$ $.56$, $p_{bh} > .579$.

3.4 Factor Analysis on the Full Set of Items and Covariance

To test to what extent the three different measurements base themselves off similar latent variables, we conducted an exploratory factor analysis using OLS to find the minimum residual solution. All items from all measurements were included. We set the number of factors to three, so if all three measures target distinctly different aspects of anthropomorphism, the items of each measure should load on their own factor (i.e. the factors would correspond to the three different measurements).

Indeed, the rGS questionnaire neatly loaded on a single factor, which shared no items with the other measurements. The MP and IDAQ scales however each loaded on both of the remaining factors. These two factors lined up with the Agency and Experience dimensions from the MP [15]: MP items like *Pain, Desire, Pride,* and *Consciousness* all loaded on one factor along with the IDAQ item *Emotions.* On the other hand, MP items such as *Personality, Memory, Emotion recognition,* and *Thought* all loaded on a different factor along with the IDAQ's *Intentions, Mind of its own.* Finally, each of the scales had one item that loaded on both factors: *Morality* and *Free will* (from MP and IDAQ respectively) could not be assigned to one factor alone. This finding is echoed in the correlations between the different measurements; while the IDAQ scores can be used to explain 66.31% of the variance on the MA, it shares only 27.62% of the variance with the rGS. Shared variance between MA and rGS similarly is 30.66%.

4 Discussion

Anthropomorphism is a central concept in the field of social robotics and human-robot interaction. While there are a wide variety of measurements available, to our knowledge our online video study is the first to attempt to benchmark three of the most popular scales against another psychometrically.

Interestingly, robot embodiment did not influence anthropomorphism scores, and neither did item randomisation within scales. However, the order of questionnaire presentation did: the Mind Attribution (MA) scale yielded higher scores if it was presented first, and the humanlikeness subscale of the Individual Differences in Anthropomorphism Questionnaire (IDAQ) had lower scores if it was directly preceded by the MA rather than the revised Godspeed questionnaire (rGS). Reliability and variance were similar between the measurements, and robot embodiment nor randomisation influenced these. However, differences were found in factor analysis. The MA and the IDAQ shared considerable variance and showed two similar underlying dimensions. The rGS on the other hand shared less than a third of its variance with the other two measurements, and loaded on a separate factor from the other two.

These findings suggest that the measurements arrive at their answer in different ways, and that this may affect the eventual score. Specifically, the IDAQ and MA scales are sensitive to preceding questions, while the rGS seems unaffected by where in the set of measurements it appears. These order effects are not the consequence of participants getting more careless as the survey continues (i.e. participant exhaustion, [18,31]) as this would have shown in larger measurement error for the later assessments (i.e. unequal variance). Thus, they have to be interpreted as the influence of context; apparently the interpretation of the items on the IDAQ and MP is affected by any questions that were asked previously. Since this influence is absent for the rGS, it stands to reason that this measure of anthropomorphism bases itself on either different aspects of anthropomorphism, or takes into account additional aspects, or both of the above. This conclusion is also supported by the factor analysis, which showed two shared latent subscales for the IDAQ and the MA, with items related to affective capabilities loading on one factor, cognitive ability items on another, and two items that load on both and appear to combine both abilities (i.e. free will and morality); while the rGS loaded on a separate dimension. It is further supported by the relatively low amount of shared variance of the rGS with the other two scales, and matches the literature and the theoretical models of anthropomorphism which suggest that while robot appearance may prompt anthropomorphism to some extent (hence the similar scores for the three measures), other cues and motivational processes exert their own, independent influence.

The findings from this experiment should not be taken as an indication that either of these measures is invalid per se, but rather as an consideration for researchers trying to either design an experiment or interpret research. As far as tested here, the scales are fairly comparable in spite of them operationalising anthropomorphism in quite different ways. The lack of differences in reliability and variance under the different conditions is particularly good news, as it supports the quality of performance under different circumstances. At the same time, the interaction effect between measurement presentation order and anthropomorphism scores warns researchers to be careful of context when designing a study, as the questions asked previously to measure may affect scores. Researchers should consider methods such as counterbalancing the order in which different measurements are taken, or adding filler tasks between the measurements to avoid order effects.

In addition to these subtle differences in outcome, the results confirmed that the rGS bases itself on a fundamentally different aspect of anthropomorphism than either the IDAQ or the MP scales. While this may not come as much as a surprise when looking at the items on the scales, having this difference confirmed may help explain why e.g. results do not replicate with different measures of anthropomorphism (e.g. [46]). Researchers interested in anthropomorphism may want to include both rGS and either IDAQ or MA, for a more complete measure of its different facets.

4.1 Limitations and Future Directions

With certain precautions, online studies such as the one in question can yield data of similar quality as in-person ones [5,19] in terms of reliability and participant attention. However, a more serious limitation of the online setup is that participants did not interact, nor were they in the presence of, an embodied robot. Robot presence and embodiment have been shown to influence robot perception and engagement [6,25], and our results suggest that certain measurements may be more or less sensitive to environmental cues besides the robot. Future studies could consider making comparisons with embodied and virtual robots.

Our experiment was not fully randomized as the IDAQ was always presented last, and thus the order effects have to be interpreted with caution. For example, it is unclear if the difference in scores should be interpreted as "being third in line lowers anthropomorphism ratings" or as "presenting the revised Godspeed immediately prior to the IDAQ heightens the IDAQ scores".

Our experiment only investigated explicit measures of anthropomorphism, thus not addressing e.g. predictive validity, i.e. whether the scales differ in the strength of their relationship to related measures, such as predicting how much empathy someone will have for a robot.

Finally, our current work approached the validity question from the classical test theory perspective and considered the measurements as a whole, comparing the mean score across items for each scale. A follow-up on these findings with item-response theory (IRT) methods would be worthwhile. Under IRT discrimination and difficulty parameters are estimated for individual items, which can be used to infer how well the scale functions in different ranges of the latent trait; as well as help identify items that are redundant; and finally identify issues with the response options for specific items.

Thus, our study should be considered a starting point and a call to other researchers to help empirically validate the measures we rely on in HRI.

Anthropomorphism is complicated and multidimensional construct, and in order to study it well we need a more nuanced catalogue of how various measurements cover specific aspects of the construct.

References

1. Banks, J.: Theory of mind in social robots: replication of five established human tests. Int. J. Social Rob. **12**(2), 403–414 (2020)
2. Bartneck, C., Kulić, D., Croft, E., Zoghbi, S.: Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int. J. Soc. Rob. **1** (2009)
3. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Royal Statis, Soc. Series B (Methodological) **57**(1) (1995)
4. Carlson, K.D., Herdman, A.O.: Understanding the impact of convergent validity on research results. Organ. Res. Methods **15**(1), 17–32 (2012)
5. Chmielewski, M., Kucker, S.C.: An MTurk crisis? Shifts in data quality and the impact on study results. Soc, Psychol. Pers. Sci. **11**(4), 464–473 (2020)
6. Deng, E., Mutlu, B., Mataric, M.J., et al.: Embodiment in socially interactive robots. Found. Trends Rob. **7**(4), 251–356 (2019)
7. Diedenhofen, B., Musch, J.: cocron: a web interface and R package for the statistical comparison of Cronbach's alpha coefficients. Int. J. Internet Sci. **11**(1) (2016)
8. Epley, N., Waytz, A., Cacioppo, J.T.: On seeing human: a three-factor theory of anthropomorphism. Psychol. Rev. **114**(4), 864 (2007)
9. Esterwood, C., Robert, L.P.: The theory of mind and human-robot trust repair. Sci. Rep. **13**(1), 9877 (2023)
10. Eyssel, F., Hegel, F.: (S)he's got the look: gender stereotyping of robots. J. Appl. Soc. Psychol. **42**(9), 2213–2230 (2012)
11. Faul, F., Erdfelder, E., Buchner, A., Lang, A.G.: Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Beh. Res. Methods **41**(4), 1149–1160 (2009)
12. Field, A., Field, Z., Miles, J.: Discovering statistics using R. Sage (2012)
13. Gena, C., Manini, F., Lieto, A., Lillo, A., Vernero, F.: Can empathy affect the attribution of mental states to robots? In: International Conference on Multimodal Interaction (2023)
14. Golafshani, N.: Understanding reliability and validity in qualitative research. Qual. Rep. **8**(4), 597–607 (2003)
15. Gray, H.M., Gray, K., Wegner, D.M.: Dimensions of mind perception. Science **315**(5812), 619 (2007)
16. Gray, K., Young, L., Waytz, A.: Mind perception is the essence of morality. Psych. Inqu. **23**(2) (2012)
17. Ho, C.C., MacDorman, K.F.: Revisiting the uncanny valley theory: developing and validating an alternative to the Godspeed indices. Comput. Hum. Behav. **26**(6), 1508–1518 (2010)
18. Jeong, D., Kumar, N., Aggarwal, S., Robinson, J., Spearot, A., Park, D.S.: Exhaustive or exhausting? Evidence on respondent fatigue in long surveys (2022)
19. Kees, J., Berry, C., Burton, S., Sheehan, K.: An analysis of data quality: professional panels, student subject pools, and Amazon's Mechanical Turk. J. Advertising **46**(1), 141–155 (2017)

20. Keijsers, M., Bartneck, C.: Mindless robots get bullied. In: ACM/IEEE HRI. ACM/IEEE (2018)
21. Krosnick, J.A., Alwin, D.F.: An evaluation of a cognitive theory of response-order effects in survey measurement. Public Opin. Q. **51**(2), 201–219 (1987)
22. Kühne, R., Peter, J.: Anthropomorphism in human-robot interactions: a multidimensional conceptualization. Commun. Theory **33**(1), 42–52 (2023)
23. Lee, M., Lucas, G., Gratch, J.: Comparing mind perception in strategic exchanges: human-agent negotiation, dictator and ultimatum games. J. Multimodal User Interfaces **15**, 201–214 (2021)
24. Leichtmann, B., Nitsch, V.: How much distance do humans keep toward robots? Literature review, meta-analysis, and theoretical considerations on personal space in human-robot interaction. J. Environ. Psychol. **68**, 101386 (2020)
25. Li, J.: The benefit of being physically present: a survey of experimental works comparing copresent robots, telepresent robots and virtual agents. Int. J. Hum. Comput. Stud. **77** (2015)
26. Li, S., Yu, F., Peng, K.: Effect of state loneliness on robot anthropomorphism: potential edge of social robots compared to common nonhumans. In: Journal of Physics: Conference Series. IOP (2020)
27. Marks, A.M., Cronje, J.C.: Randomised items in computer-based tests: Russian roulette in assessment? J. Educ. Technol. Soc. **11**(4), 41–50 (2008)
28. Mumm, J., Mutlu, B.: Human-robot proxemics: physical and psychological distancing in human-robot interaction. In: ACM/IEEE HRI. ACM/IEEE (2011)
29. Nass, C., Steuer, J., Tauber, E.R.: Computers are social actors. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 72–78 (1994)
30. Pister, H.L., Kondrad, R., Kwong, J., Smith, A., Vahlbusch, J.: What's in a name? Preschoolers treat a bug as moral agent when it has a proper name. Ph.D. thesis, Appalachian State University (2017)
31. Porter, S.R., Whitcomb, M.E., Weitzer, W.H.: Multiple surveys of students and survey fatigue. New Dir. Inst. Res. **2004**(121), 63–73 (2004)
32. Ruijten, P.A., Haans, A., Ham, J., Midden, C.J.: Perceived human-likeness of social robots: testing the Rasch model as a method for measuring anthropomorphism. Int. J. Soc. Rob. **11** (2019)
33. Sacino, A., et al.: Human-or object-like? Cognitive anthropomorphism of humanoid robots. PLoS ONE **17**(7), e0270787 (2022)
34. Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., Joublin, F.: To Err is human (-like): effects of robot gesture on perceived anthropomorphism and likability. Int. J. Soc. Rob. **5** (2013)
35. Sanjeev, M., Balyan, P.: Response order effects in online surveys: an empirical investigation. Int. J. Online Mark. **4**(2), 28–44 (2014)
36. Sireci, S.G.: The construct of content validity. Soc. Indic. Res. **45**, 83–117 (1998)
37. Tahiroglu, D., Taylor, M.: Anthropomorphism, social understanding, and imaginary companions. Br. J. Dev. Psychol. **37**(2), 284–299 (2019)
38. Takayama, L., Pantofaru, C.: Influences on proxemic behaviors in human-robot interaction. In: 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 5495–5502. IEEE (2009)
39. Tan, H., Wang, D., Sabanovic, S.: Projecting life onto robots: the effects of cultural factors and design type on multi-level evaluations of robot anthropomorphism. In: 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 129–136. IEEE (2018)

40. Tavakol, M., Dennick, R.: Making sense of Cronbach's alpha. Int. J. Med. Educ. **2** (2011)
41. Thellman, S., Giagtzidou, A., Silvervarg, A., Ziemke, T.: An implicit, non-verbal measure of belief attribution to robots. In: Companion of ACM/IEEE HRI (2020)
42. Thellman, S., de Graaf, M., Ziemke, T.: Mental state attribution to robots: a systematic review of conceptions, methods, and findings. ACM Trans. HRI **11**(4) (2022)
43. Van Straten, C.L., Peter, J., Kühne, R., Barco, A.: The wizard and I: how transparent teleoperation and self-description (do not) affect children's robot perceptions and child-robot relationship formation. AI and Society, pp. 1–17 (2022)
44. Waytz, A., Cacioppo, J., Epley, N.: Who sees human? The stability and importance of individual differences in anthropomorphism. Perspect. Psychol. Sci. **5**(3) (2010)
45. Weinberg, M.K., Seton, C., Cameron, N.: The measurement of subjective wellbeing: item-order effects in the personal wellbeing index–adult. J. Happiness Stud. **19**, 315–332 (2018)
46. Złotowski, J., Sumioka, H., Eyssel, F., Nishio, S., Bartneck, C., Ishiguro, H.: Model of dual anthropomorphism: the relationship between the Media Equation effect and implicit anthropomorphism. Int. J. Soc. Rob. **10**, 701–714 (2018)