The Perception of Agency: Scale Reduction and Construct Validity*

J. Gregory Trafton[1] and Chelsea R. Frazier[2] and Kevin Zish[3] and Branden J. Bio[4] and J. Malcolm McCurry[5]

***Abstract*— The perception of agency in robots and AI characters has become increasingly important as different agents increase their capabilities. Experiment 1 took an existing measure of perceived agency and created a reduced version by using existing Rasch item reduction measures; Eight and five item scales were created. Experiment 2 showed that all three scales (PA, PA8, PA5) were able to capture differences in perceived agency between a cheating robot (higher PA) and a non-cheating robot (lower PA). Experiment 3 showed that all three scales were able to show the predicted positive relationship between perceived agency and perceived moral agency. All three scales also showed high internal validity. Suggestions for the usage of the scales was also discussed.**

I. INTRODUCTION

How long should a measurement scale be? Scale designers typically need to balance theory (scales with many items are typically able to cover the full range of the relevant dimension) and practice (people don't want to fill out long surveys and in fact are more likely to drop out the longer the survey is [1], [2]). The practical aspect of scale length – completion rate – is important for a number of reasons, including representativeness of the sample, money, and time.

Within the field of Human Robot Interaction, social robotics, cognitive robotics, and conversational AI agents, these issues are especially relevant because researchers are attempting to explore and understand multiple measures concurrently. In fact, a great deal of these fields are concerned with exploring and understanding how and why a design or algorithm or behavior impacts people. Is the robot or AI social [3]? Is it trustworthy [4]? Is it perceived as a moral agent [5] or as a moral patient [6] or anthropomorphic [7]? It is very difficult from a practical perspective to run more than 2 or 3 surveys if each survey has many items or takes a long time to complete. Of course researchers also need to have confidence that the instrument is reliably measuring what it is intended to measure.

Recently, we developed a new scale that measures how much an individual thinks another entity has agency: this is the Perceived Agency (PA) scale [8]. The scale was developed using a Rasch methodology [9]. Multi-faceted Rasch measurement was used in part because it is particularly well-suited for measuring an external entity by modeling the variance associated with each entity. The scale has 13 items that provide excellent coverage across a wide range of robots, AI characters, and humans that exhibit different levels of perceived agency. The original scale-creation [8] paper focused on how the scale was designed and showed that it could capture differences in PA across a wide range of entities better than other existing measures of perceived agency [8].

While 13 items is not excessively long for a single scale, it may be possible to reduce the number of items. Thus, this paper will use standard Rasch methodologies to attempt to reduce the size of the scale. After we show scale reduction, we will describe two experiments that show construct validity of the full scale and examine how well the reduced scales perform.

Our contributions in this paper are:

- To reduce the size of the perceived agency scale that can be applied to robots, AIs, conversational agents, and organic entities;
- To provide construct validity of the reduced scale;
- To demonstrate the theorized relationship between perceived agency and cheating; and
- To demonstrate the theorized relationship between perceived agency and perceived moral agency.

II. EXPERIMENT 1: SCALE REDUCTION

The goal of experiment 1 was to reduce the number of items on the PA scale; the original 13 items are shown in Table I. This data analysis is a re-analysis of the original scale creation from [8]; hence, the method will be summarized since the full method is available in the source article.

186 participants viewed 7 videos and answered a series of Likert items (five points with values of Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree) after each video. The videos consisted primarily of robots, but also contained an AI character and a human; each video lasted between 30 seconds and 3 minutes. The full 13 item PA survey was analyzed using the Rasch method [10].[1]

The Rasch model is an additive linear model based on a logistic transformation of ratings to a logit scale. Rasch models can have multiple facets (in our case, entities, raters, and items) that are all on the same logit scale and all can influence the final rating. Conceptually, this suggests that the logit scale represents the latent value (dimension) – the amount of perceived agency [9], [11], [12], [13].

*This work was supported by the Office of Naval Research

[1]Greg Trafton is with the Naval Research Laboratory, Washington, DC 20375, USA greg.trafton@nrl.navy.mil

[2]Chelsea Frazier is with United States Military Academy, West Point, chelsea.frazier@westpoint.edu

[3]Kevin Zish is with Global Systems Technologies, kevin.zish@associates.tsa.dhs.gov

[4]Branden Bio is with the National Research Council branden.bio.ctr@nrl.navy.mil

[5]Malcolm McCurry is with Arcfield john.mccurry@arcfield.com

[1]All three experiments had IRB approval of Naval Research Laboratory

Like other approaches to survey creation, Rasch has methods to identify items that are poor fits [14], whether the scale is unidimensional [15], [16], and the overall reliability of the scale. In the original analysis, these analyses showed that the scale had high reliability, was unidimensional, and the items had acceptable fits for scale use [8].

A. Perceived Agency

We followed best-in-practice suggestions for scale construction [17], [18] and began with a definition:

People perceive agency in another entity when the entity's actions may be assumed by an outside observer to be driven primarily by its internal thoughts and feelings and less by the external environment.

A large number of items were generated from that definition; the final items covered thoughts (acts with purpose; has goals; can create new goals; can communicate with people; treats others as if they had a mind), feelings (wanted to perform these actions; can show emotions to other people; can change their behavior based on how people treat them), and environmental impacts on behavior (can adapt to different situations, would do well in other environments, can perform many different types of tasks). There were also two integrative items. The actor scenario was "Imagine the robot/character/person was asked to be an actor in a local theater production. How well do you think they would do?" The dinner scenario was "Imagine the robot/character/person was asked to host a dinner party for your friends next weekend. This includes coming up with a menu, cooking, and hosting. How well do you think they will do?" This definition allowed us to select items that differentiate perceived agency from anthropomorphism, reality interaction, consiousness, and other related but distinct concepts.

TABLE I

PERCEIVED AGENCY ITEMS FOR THE ORIGINAL PA SCALE, AND THE REDUCED SCALES PA8 AND PA5.

Order	Item	PA	PA8	PA5
1	acts with purpose	✓		
2	has goals	✓	✓	
3	can create new goals	✓	✓	✓
4	can communicate with people	✓	✓	✓
5	treats others as if they had a mind	✓		
6	wanted to perform these actions	✓	✓	
7	can show emotions to other people	✓		
8	can change their behavior based on how people treat them	✓	✓	✓
9	can adapt to different situations	✓	✓	✓
10	would do well in other environments	✓		
11	has a face (attention check)			
12	can perform many different types of tasks	✓		
13	actor scenario	✓	✓	✓
14	dinner scenario	✓	✓	

B. Item Reduction: High Skewness and Kurtosis

Previous research has suggested that items can be removed if they have skewness and kurtosis outside the range of ± 2 [19], [20], [21].

Items with high skewness or kurtosis can suggest that the items may be mis-targeting some of the sample and not provide enough information to be as useful in the scale. These items with high skewness or kurtosis can therefore be removed.

The item "acts with purpose" had a kurtosis of 3.01, so was removed.

C. Item Reduction: Local Dependency

Researchers have also suggested that items can be removed if there is local dependency between items [19], [20], [21].

Local dependency occurs when 2 or more items should not be correlated with each other after the underlying dimension is conditioned out. Items can be dependent on each other if a higher (or lower) response on one item leads to a higher (or lower) response on another item. Thus, items should only be correlated through the dimension that we are measuring [22]. Local dependency can be measured using Yen's Q3 statistic. If a pair of items had residual correlations exceeding the mean of all the residual correlation by 0.20, they are considered to have local dependence and can be removed [23].

After "acts with purpose" was removed, the local dependency test was performed to determine if more items could be removed.

The average residual correlation of all pairs of items was 0.224, so pairs of items that had residual correlation values > 0.424 (0.224 + 0.2) suggested dependency. Table II shows the residual correlation values greater than the value of 0.424 for all pairs of items over this threshold, which indicated local dependency. In order to determine which of the pair to remove, the item with a MNSQ outfit closer to 1 was kept. In Rasch, MNSQ outfit is the most common method for evaluating how well the item fits the model. High values of outfit suggest there are many outliers (i.e., the model expected a low score based on the entity and the rater, but there were many high scores instead) while low values of outfit suggest the item is providing little additional data from the other items. The mean of outfit is always 1, so items that are further away from 1 are considered worse.

TABLE II

LOCAL DEPENDENCY ANALYSIS USING YEN'S Q3 STATISTIC. PAIRS OF ITEMS WITH A RESIDUAL CORRELATION $> .2$ ABOVE THE AVERAGE CORRELATION (.224) SUGGEST DEPENDENCY. ITEMS WITH AN OUTFIT MNSQ CLOSER TO 1 (STARRED) WERE RETAINED.

Resid Corr	Item	Outfit	Item	Outfit
.62	*Can communicate...	.87	Treats others...	.80
.54	Can show emotion...	.77	*Can change beh...	.86
.43	Can perform many...	1.19	*Can adapt ...	1.06
.54	Can perform many...	1.19	*Would do well...	1.13
.57	*Can adapt...	1.06	Would do well...	1.13

The skewness / kurtosis test removed one item and the local dependency test removed another four items, reducing the scale from 13 to 8 items, shown in Table I.

D. Item Reduction: Iterative Precision

Another approach to further reduce the number of items from a scale developed using Rasch is to iteratively remove items based on the reliability of the scale [24], [25]. Because our goal was to measure the perceived agency of external entities, we focused on the reliability of the entity (video) facet rather than the more traditional person reliability [24].

Rasch analysis provides two different measures for reliability. The first, separation, indicates how many different levels of the facet can be distinguished. A small separation value suggests that different levels can not be distinguished while a larger value is more desired for measurement. The second reliability measure, separation reliability, is equivalent to cronbach α reliability and is a measure of internal consistency. Separation reliability ranges from 0 to 1; over .8 is considered acceptable for scale creation.

The item that decreased reliability or separation the least was removed and the process was repeated. We stopped when separation reliability became ≤ 0.8 or when five items remained, whichever happened first.

This process allowed the removal of three items "has goals", "wanted to perform these actions", and the "dinner scenario." Fit statistics for all items were deemed to be excellent for scale creation [14]: all items had a MNSQ outfit statistic < 1.5, reliability for all three facets was ≥ 0.89, separation for the entities was over 22, person separation was 2.8, and item separation was 16.5.

E. Reliability of each scale

Cronbach's α, a standard measure of internal consistency, was calculated for each scale. The PA scale had an α of 0.94; The PA8 scale had an α of 0.90; and the PA5 scale had an α of 0.92. These are all excellent αs for a unidemsional scale.

F. Discussion

Experiment 1 re-analyzed the data from [8] in order to reduce the scale size from the original perceived agency scale, which had 13 items. Several methods were used to reduce the number of items; all methods have been used by previous researchers. Items from PA8 were selected by removing a single item with high kurtosis, and then another four items that had local dependency. An iterative reduction approach based on the removing items that reduced the precision the least was then applied to the items from PA8. An additional three items were removed, making the PA5 scale. Cronbach's α showed all three scales to have high internal consistency.

It is interesting to note that all three scales have at least one item that is core to our definition of perceived agency (thoughts, feelings, environment), suggesting that these concepts are important to the accurate measurement of perceived agency.

One of the fundamental questions about new scales is how much construct validity they have. We next examine convergent validity of all three scales by exploring how well they can measure increased perceived agency of a cheating robot.

III. Experiment 2: Cheating Construct validity

One of the most influential studies on perceived agency showed that a cheating robot had more perceived agency than a robot that did not cheat or that made a mistake [26]. Short et al. measured perceived agency by performing qualitative coding on the language the participants used. Other researchers have used a similar methodology and had similar findings [27].

Construct validity is "extent to which an instrument assesses a construct of concern and is associated with evidence that measures other constructs in that domain" [28]. Convergent validity is a subset of construct validity and can be shown if our new measure reflects previous findings on perceived agency [28], [29]. Thus, if we can show that our PA measures detect differences in the expected direction (cheating robots have more perceived agency than non-cheating robots), our PA scales will have shown convergent construct validity.

A. Method

1) Participants: The number of participants needed to detect a medium effect size with 80% power and α = 0.05 was n = 126. 135 participants were recruited through Cloud Research and paid $3.00 for participation in the study. 21 participants were removed because they missed an attention check ("has a face"), 1 participant was removed for missing an item, leaving 113 participants. The average age of participants was 39 years old. 43 participants were women, 68 participants were men, and 2 participants were unreported[1].

2) Materials (Videos): In the original cheating robot study [26] as well as a more recent study [27], a robot played "rock paper scissors" against a human opponent. A video of a robot in the cheating condition from [27] was used. In that experiment, there were 30 in-person rounds of the game "rock paper scissors"; we only used three rounds in our online experiment.

There were two conditions in our experiment: a cheating condition and a non-cheating condition. In both conditions, the robot taught the person how to play the game and how the robot made each different "throw" (rock/paper/scissors). The non-cheating condition had three distinct rounds of a person playing a game of rock paper scissors with the robot: in all three rounds the robot played the game normally (without cheating). In the cheating condition, the first two rounds were identical to the non-cheating condition, but in the third round the robot cheated by changing from a losing throw to a winning throw and claiming they had won the round.

3) Materials (Survey Items): The full set of items from the three PA scales (PA, PA8, and PA5) were used (see Table I).

4) Procedure: After answering a series of demographic questions, participants were given a brief description of the task and told they would answer a series of questions after watching a video. Participants were randomly put into either the cheating or the non-cheating condition. At the end of the video, they were taken to a single page with the same video that they could watch again if desired. They were first asked

to describe the video in at least one sentence. Next they were asked to answer the Perceived Agency survey items in Table I. Finally, at the end of the session, participants were invited to provide experimental feedback.

B. Results

1) Calculating scale values: For the PA, PA8, and PA5 scales, the respective items were averaged to give a single score for each rater for each entity.

2) Perceived Agency: Recall that in previous studies using a qualitative method for determining perceived agency, the cheating robot had more perceived agency than the non-cheating robot. Thus, our goal here is to examine whether our PA, PA8, and PA5 surveys are able to detect a difference between conditions.

All three scales showed high internal reliability as measured by Cronbach's α: PA had an α of 0.92; PA8 had an α of 0.87; and PA5 had an α of 0.87.

Consistent with our hypothesis and showing convergent validity, all three surveys showed that the cheating robot had more perceived agency than the non-cheating robot (see Table III).

TABLE III
DIFFERENCES BETWEEN CHEATING AND NON-CHEATING CONDITIONS FOR PA, PA8, AND PA5.

Scale	Cheating mean	Non-Cheating mean	t	df	prob	Cohen's d
PA	3.6	3.3	2.3	111	< 0.05	.42
PA8	3.6	3.3	2.7	111	< 0.05	.49
PA5	3.6	3.3	2.3	111	< 0.05	.43

While there are slight differences in the three scales, all three show enough sensitivity to detect a difference between the cheating and non-cheating conditions. This finding shows construct validity for all three scales and also suggests that there is a reliable, replicable impact of cheating on perceived agency.

C. Discussion

All three PA scales showed that the cheating robot had more perceived agency than the non-cheating robot. This provides initial construct validity for the PA scales and also provides additional support that cheating seems to increase the amount of perceived agency in a robot.

IV. Experiment 3: Perceived Moral Agency Construct validity

Previous researchers have suggested that there is a positive relationship between agency and morality [30], [31], [32]. None of those studies, however, have used validated measures of both perceived agency and perceived morality. For example, [30] used agency measures that corresponded to "humanness" (i.e., culturally refined or rational/logical) and moral agency was measured by holding people "morally responsible." Similarly, [32] measured moral patiency by the amount of pain someone felt and they measured agency by the amount of intentionality someone exhibited. While all of these measures are suggestive of perceived agency and perceived morality, we believe that there are better methods for measuring both. To this end, the present study explores the relationship between these two constructs. This experiment explicitly examined the hypothesis that perceived agency and perceived moral agency are positively correlated by examining the relationship between our three perceived agency scales and the morality sub-scale of the perceived moral agency scale [5].

A. Method

1) Participants: 99 participants were recruited through Cloud Research and paid $11.50 for participation in the study.[2] 9 participants were removed because they missed an attention check ("has a face") leaving 90 participants. The average age of participants was 39 years old. 50 participants were women, 40 participants were men[1].

2) Materials (Videos): 7 Videos were selected and collected from a wide range of sources, including YouTube, academic conference proceedings, and personal communication with leaders of the field in robotics. The majority of our stimuli were robots (5 instances), but also included an AI agent (1 instance) and a human (1 instance). The entities portrayed a range of engagement with people and the non-human entities had different morphologies, differed in their sensing, and had different perceptual, navigation, mobility, cognitive, and social capabilities. According to pilot testing, they also differed in term of their perceived moral agency and perceived agency.

Table IV provides a label, a brief description, the morphology of the entity, and a citation of the source. The citation of each video is either a YouTube location or a paper or website describing video.

Our goal was to keep the robot videos between 30 seconds and 3 minutes. In some cases the video was trimmed or cut. In all cases, we attempted to show the core aspects of the target and their activity while making sure that participants would not become bored watching the video.

3) Materials (Survey Items): This study reports material collected from a larger data collection effort; for this study we measured perceived agency by using the three PA surveys as shown in Table I.

To measure perceived moral agency (PMA), the PMA survey that was developed and validated by Banks was used [5]; items are shown in Table V. The PA items used a Likert scale of 1-5 while the PMA items used a Likert scale range of 1-7.

4) Procedure: All participants viewed all seven videos. The procedure for each video was identical to that used in experiment 2 except that after viewing each video, participants saw both the PA and PMA surveys. The two different sets of survey items were kept together but the order of each block was randomly determined.

[2]An earlier version of this experiment appeared at the (non-archival) Perspectives on Moral Agency in Human–Robot Interaction at HRI 2023.

TABLE IV

DESCRIPTION OF VIDEOS USED IN EXPERIMENT.

Label	Robot actions	Morphology	Source
Welding	Welding metal	industrial arm	[33]
TaiChi	Balancing and movement	Humanoid	[34]
Pouring	Pushes cart, unscrews thermos, pours juice and gives it to human	Humanoid	[35]
Robot Secrets Revealed '09	Magician tricking robot	Humanoid	[36]
Bargaining3	Human bargaining with AI agent	Humanoid character	[37]
Punished	Robot put in closet unwillingly	Humanoid	[38], [39]
Professor	Teaching computer science	Human	[40]

TABLE V

PMA ITEMS FROM BANKS (MORALITY DIMENSION) (2019). ROBOT/CHARACTER/HUMAN WAS USED AS APPROPRIATE FOR THE ENTITY.

Number	Survey Item
1	has a sense for what is right and wrong
2	can think through whether an action is moral
3	is capable of being rational about good and evil
4	behaves according to moral rules
5	might feel obligated to behave in a moral way
6	would refrain from doing things that have painful repercussions

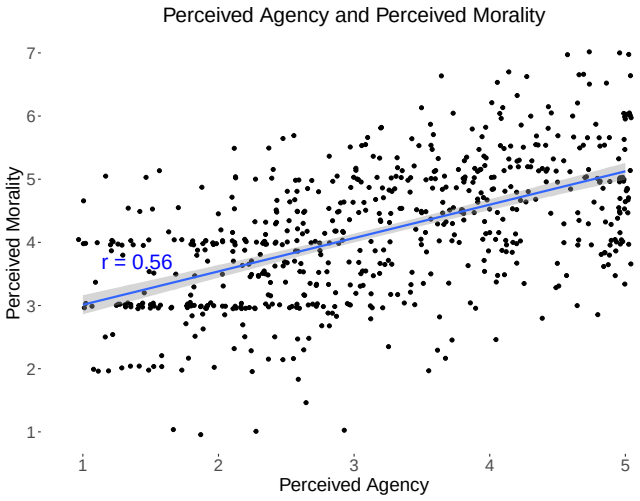

Fig. 1. Correlation between Perceived Moral Agency scale (morality dimension) [5] and Perceived Agency scale [8]. Values have been slightly jittered to show the density.

B. Results

1) Calculating scale values: For the PA and PMA scales, the respective items were averaged to give a single score for each rater for each entity. This resulted in a total of 1260 data points.

2) Comparing Perceived Agency and Perceived Moral Agency: Our primary goal was to determine whether there was a positive relationship between perceived moral agency and perceived agency. A Pearson's product-moment correlation based on Fisher's z transformation showed there was a large correlation between the morality scale of PMA [5] and the perceived agency scale [8], $r = 0.56, p < 0.0001$; this relationship is shown in Figure 1. The PA8 scale also showed a positive correlation with the PMA scale, $r = 0.54, p < 0.0001$. Finally, the PA5 scale showed a strong positive correlation with the PMA scale, $r = 0.54, p < 0.0001$.

Our second goal was to examine the range of both perceived agency and perceived moral agency. Figure 2 highlights the overall difference in range for both surveys. To make the patterns clear, Figure 2's x-axis is ordered by the mean perceived agency value for each entity. Both PA and PMA have a positive slope and they both capture differences in entity-type, PA $F(6, 89) = 294.8, MSE = 74.5, p < 0.0001$ and Morality $F(6, 89) = 23.4, MSE = 14.6, p < 0.0001$. Interestingly, both PA and Morality capture differences in entity-type for only non-humans (i.e., after removing the single human entity and re-running the ANOVA), PA, $F(5, 89) = 183.7, MSE = 44.1, p < 0.0001$ and Morality $F(5, 89) = 31.0, MSE = 13.7, p < 0.0001$. The mean PA scale ranges from a bit under 2 for the Welding robot to almost 5 for the human professor. In contrast, the Morality scale is much more narrow on average, ranging from 2.7 to 3.4[3]. The PA8 and PA5 scales show nearly identical results.

C. Discussion

We have shown that perceived agency and perceived moral agency are possitively correlated. While previous research has suggested there is a relationship, this is the first study we know of that can quantify the degree of relationship – approximately 32%. This is quite a substantial relationship between perceived agency and perceived moral agency.

This study also provides construct validity support for all three of the perceived agency scales as well as the perceived moral agency scale (morality dimension), providing converging support for both scales.

While the range is quite different for both PA and Morality across different entities, we can cautiously interpret what both scales mean across the entities. Perceived Agency provides a rather large and reasonable ordering across 7 different types of entities, ranging from a highly repetitive robot (welding) to a robot that complains when it is asked to go into a dark closet (closeted) to a human professor teaching algorithms (professor). People seem to be able to attribute

[3]For these numbers and Figure 2, the Morality numbers were converted to be comparable with the PA scale (i.e., we converted from 1-7 to 1-5).

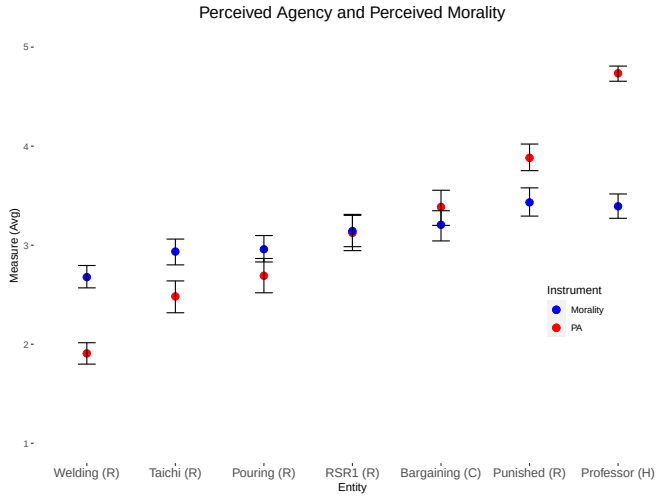

Fig. 2. Average scores for the Perceived Moral Agency scale (morality dimension) [5] and Perceived Agency scale [8] for each entity. Error bars are 95% CI.

perceived agency for these different entities in a strong and robust manner.

V. General Discussion and Future Work

We began with a perceived agency scale of 13 items. We then used three techniques from the Rasch method to reduce the number of items to 8 and then 5. Experiment 1 showed that all three scales (PA, PA8, and PA5) had excellent internal validity and reliability.

Experiment 2 showed that one of the most well known findings in HRI – that a robot that cheated was perceived to have more agency than a robot that did not cheat – was replicated using all three PA scales. This result is important because it took an important finding that had been time-consuming to measure and showed that it can now be measured in a straightforward manner.

Experiment 3 examined the hypothesized relationship between perceived agency and perceived moral agency. Experiment 3 showed that there was an extremely strong relationship between the two constructs. Experiments 2 and 3 showed construct validity for all three of the PA scales.

This paper did not explore how these measures would work in other cultures, other contexts (i.e., in person with a live robot), or with other stimuli (e.g., vignettes). In the future, these issues will be explored.

We also note that with the prevalence of large language models like ChatGPT, people seem to attribute more agency to these models than they actually have (*Washington Post*, June 11, 2022). Our PA scale may provide a theoretical reason why – people believe these models have high level of agency because they act as if they are motivated by their thoughts and feelings. Our PA scale also allows empirical examination of the phenomena.

We began this paper suggesting that researchers needed to balance the theory of having a scale with many items and broad coverage of the construct with the practical concern of losing participants to long scales. For the most complete coverage, we recommend using the PA-R (Perceived Agency-Rasch) scale developed in [8]; this scale uses the same 13 items described here, but also uses three calibration videos. We do not discuss this scale in this paper but it is theoretically the strongest scale of PA we know of. The averaged 13 item PA scale reported here also provides excellent coverage of the perceived agency dimension. For researchers that need shorter scales, both the averaged PA8 and PA5 seem to provide excellent coverage (experiment 1) and track other measures in predicted ways (experiments 2 and 3).

References

[1] M. Hoerger, "Participant dropout as a function of survey length in internet-mediated university studies: Implications for study design and voluntary participation in psychological research," *Cyberpsychology, behavior, and social networking*, vol. 13, no. 6, pp. 697–700, 2010.

[2] M. Liu and L. Wronski, "Examining completion rates in web surveys via over 25,000 real-world surveys," *Social Science Computer Review*, vol. 36, no. 1, pp. 116–124, 2018.

[3] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The robotic social attributes scale (rosas) development and validation," in *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*, 2017, pp. 254–262.

[4] B. F. Malle and D. Ullman, "A multidimensional conception and measure of human-robot trust," in *Trust in human-robot interaction*. Elsevier, 2021, pp. 3–25.

[5] J. Banks, "A perceived moral agency scale: development and validation of a metric for humans and social machines," *Computers in Human Behavior*, vol. 90, pp. 363–371, 2019.

[6] J. Banks and N. D. Bowman, "Perceived moral patiency of social robots: Explication and scale development," *International Journal of Social Robotics*, vol. 15, no. 1, pp. 101–113, 2023.

[7] N. Epley, A. Waytz, and J. T. Cacioppo, "On seeing human: a three-factor theory of anthropomorphism." *Psychological review*, vol. 114, no. 4, p. 864, 2007.

[8] J. Trafton, C. Young, J. McCurry, and K. Zish, "The perception of agency," *Transactions on Human Robot Interaction*, under review.

[9] T. Bond and C. Fox, "Applying the rasch model. mahwah, nj: L," 2001.

[10] J. M. Linacre, "Facets," *Computer Program for Many-faceted Rasch Measurement*, 2022.

[11] K. J. Conrad, B. D. Wright, P. McKnight, M. McFall, A. Fontana, and R. Rosenheck, "Comparing traditional and rasch analyses of the mississippi ptsd scale: Revealing limitations of reverse-scored items," *Journal of Applied Measurement*, vol. 5, no. 1, pp. 15–30, 2004.

[12] P. A. Ruijten, A. Haans, J. Ham, and C. J. Midden, "Perceived human-likeness of social robots: testing the rasch model as a method for measuring anthropomorphism," *International Journal of Social Robotics*, vol. 11, no. 3, pp. 477–494, 2019.

[13] B. D. Wright and M. H. Stone, "Best test design," 1979.

[14] J. M. Linacre, M. Stone, J. William, P. Fisher, and L. Tesio, "Rasch measurement," *Rasch Measurement Transactions*, vol. 16, 2002.

[15] Y.-T. Chou and W.-C. Wang, "Checking dimensionality in item response models with principal component analysis on standardized residuals," *Educational and Psychological Measurement*, vol. 70, no. 5, pp. 717–731, 2010.

[16] G. Raîche, "Critical eigenvalue sizes in standardized residual principal components analysis," *Rasch measurement transactions*, vol. 19, no. 1, p. 1012, 2005.

[17] G. O. Boateng, T. B. Neilands, E. A. Frongillo, H. R. Melgar-Quiñonez, and S. L. Young, "Best practices for developing and validating scales for health, social, and behavioral research: a primer," *Frontiers in public health*, vol. 6, p. 149, 2018.

[18] J. M. Cortina, Z. Sheng, S. K. Keener, K. R. Keeler, L. K. Grubb, N. Schmitt, S. Tonidandel, K. M. Summerville, E. D. Heggestad, and G. C. Banks, "From alpha to omega and beyond! a look at the past, present, and (possible) future of psychometric soundness in the journal of applied psychology." *Journal of Applied Psychology*, vol. 105, no. 12, p. 1351, 2020.

[19] M. Cantó-Cerdán, P. Cacho-Martínez, F. Lara-Lacárcel, and Á. García-Muñoz, "Rasch analysis for development and reduction of symptom questionnaire for visual dysfunctions (sqvd)," *Scientific Reports*, vol. 11, no. 1, pp. 1–10, 2021.

[20] K. Pesudovs, J. M. Burr, C. Harley, and D. B. Elliott, "The development, assessment, and selection of questionnaires," *Optometry and Vision Science*, vol. 84, no. 8, pp. 663–674, 2007.

[21] K. Pesudovs, E. Garamendi, J. P. Keeves, and D. B. Elliott, "The activities of daily vision scale for cataract surgery outcomes: re-evaluating validity with rasch analysis," *Investigative ophthalmology & visual science*, vol. 44, no. 7, pp. 2892–2899, 2003.

[22] F. M. Lord and M. R. Novick, *Statistical theories of mental test scores*. Reading, Mass: Addison-Wesley, 1968.

[23] K. B. Christensen, G. Makransky, and M. Horton, "Critical values for yen's q 3: Identification of local dependence in the rasch model using residual correlations," *Applied psychological measurement*, vol. 41, no. 3, pp. 178–194, 2017.

[24] J. H. Hibbard, E. R. Mahoney, J. Stockard, and M. Tusler, "Development and testing of a short form of the patient activation measure," *Health services research*, vol. 40, no. 6p1, pp. 1918–1930, 2005.

[25] J. A. Weller, N. F. Dieckmann, M. Tusler, C. Mertz, W. J. Burns, and E. Peters, "Development and testing of an abbreviated numeracy scale: A rasch analysis approach," *Journal of Behavioral Decision Making*, vol. 26, no. 2, pp. 198–212, 2013.

[26] E. Short, J. Hart, M. Vu, and B. Scassellati, "No fair!! an interaction with a cheating robot," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2010, pp. 219–226.

[27] S. Yasuda, D. Doheny, N. Salomons, S. S. Sebo, and B. Scassellati, "Perceived agency of a social norm violating robot," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2020.

[28] T. Raykov and G. A. Marcoulides, *Introduction to psychometric theory*. Routledge, 2011.

[29] V. Rousson, T. Gasser, and B. Seifert, "Assessing intrarater, interrater and test–retest reliability of continuous measurements," *Statistics in medicine*, vol. 21, no. 22, pp. 3431–3446, 2002.

[30] B. Bastian, S. M. Laham, S. Wilson, N. Haslam, and P. Koval, "Blaming, praising, and protecting our humanity: The implications of everyday dehumanization for judgments of moral status," *British Journal of Social Psychology*, vol. 50, no. 3, pp. 469–483, 2011.

[31] H. M. Gray, K. Gray, and D. M. Wegner, "Dimensions of mind perception," *science*, vol. 315, no. 5812, pp. 619–619, 2007.

[32] K. Gray, L. Young, and A. Waytz, "Mind perception is the essence of morality," *Psychological inquiry*, vol. 23, no. 2, pp. 101–124, 2012.

[33] O. Technologies, Nov 2017. [Online]. Available: https://www.youtube.com/watch?v=Oz7TE1Q1rhw

[34] Motorward, Nov 2017. [Online]. Available: https://youtu.be/jJYsOsoBIZU?t=243

[35] "Youtube/honda unveils all-new asimo humanoid robot," Nov 2011. [Online]. Available: https://www.youtube.com/watch?v=1V9XUMCPGF8

[36] A. M. Harrison, B. R. Fransen, M. Bugajska, and J. G. Trafton, 2009. [Online]. Available: https://www.youtube.com/watch?v=XsubQhtD6S0

[37] J. Gratch, D. DeVault, G. M. Lucas, and S. Marsella, "Negotiation as a challenge problem for virtual humans," in *International Conference on Intelligent Virtual Agents*. Springer, 2015, pp. 201–215.

[38] I. Spectrum, Apr 2012. [Online]. Available: https://www.youtube.com/watch?v=DAiWZO0dz8M

[39] P. H. Kahn Jr, T. Kanda, H. Ishiguro, N. G. Freier, R. L. Severson, B. T. Gill, J. H. Ruckert, and S. Shen, ""robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot." *Developmental psychology*, vol. 48, no. 2, p. 303, 2012.

[40] H. University, Jul 2016. [Online]. Available: https://www.youtube.com/watch?v=0JUN9aDxVmI&t=4273s