METHODOLOGICAL ARTICLE

WILEY

Preliminary validation of the European Portuguese version of the Robotic Social Attributes Scale (RoSAS)

Raquel Oliveira[1,2] | Patrícia Arriaga[1] | Steven J. Stroessner[3] | Ana Paiva[2]

[1]Department of Social and Organizational Psychology (CIS-IUL), ISCTE-University Institute of Lisbon, Lisbon, Portugal

[2]Department of Computer Science, INESC-ID (GAIPS) & Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

[3]Barnard College, Columbia University, New York City, New York, USA

Correspondence
Raquel Oliveira, Department of Social and Organizational Psychology (CIS-IUL), ISCTE-University Institute of Lisbon, Lisbon, Portugal.
Email: rsaoa@iscte-iul.pt

Funding information
Fundação para a Ciência e Tecnologia, Grant/Award Number: PD/BD/150570/2020

Abstract

People's perception of social robots is essential in determining their responses and acceptance of this type of agent. Currently, there are few instruments validated for the European Portuguese population that measure the perception of social robots. Our goal was to translate, validate, and evaluate the psychometric properties of the Robotic Social Attributes Scale (RoSAS) to European Portuguese. To achieve this goal, we conducted a validation study using a sample of 185 participants. We measured the temporal validity of the scale (over a 2-week interval) and its divergent and convergent validity using the Portuguese Negative Attitudes toward Robots Scale and the Godspeed scales. Our data analysis resulted in a shortened version of the Portuguese RoSAS with 11 items while retaining the original three-factor structure. The scale presented poor to acceptable levels of temporal reliability. We found a positive correlation between the warmth and competence dimensions. Further validation studies are needed to investigate the psychometric properties of this scale.

KEYWORDS
perception, Portuguese, RoSAS, social robots, validation

1 | INTRODUCTION

In the past few decades, many advancements in technology have led to drastic changes in the way people communicate and interact with one another. In particular, the creation and introduction of social robots in many social human environments has opened the door to a new type of techno-socialization that encompasses the relations and communication dynamics between humans and technological social artifacts (Gutierrez et al., 2019; Katz, 2017; Nikitina, 2007). In this context, these technological artifacts take on the role of interactive social agents that can help and collaborate with their human counterparts in a wide range of tasks (Lee & Nass, 2010; Nass et al., 1994). The emergence of these computerized agents as social actors has, thus, resulted in increased concern about the role of different robot-related characteristics in the users' perceptions, attitudes, and level of trust toward robots (Anzalone et al., 2015; Edwards et al., 2019; Höflich & El Bayed, 2015). These factors are believed to predict several important variables, such as user engagement, which are central to those developing and studying social robots.

1.1 | Assessing the human perception of robots

The Robotic Social Attributes Scale (RoSAS) was designed to assess the central attributes implicated in human perception of social robots (Carpinella et al., 2017). Since its creation, it has been used in multiple studies in the area of human–robot interaction (HRI; e.g., Bonani et al., 2018; Pan et al., 2018; Strohkorb Sebo et al., 2018), presenting good psychometric properties (Carpinella et al., 2017; Pan et al., 2017). This scale includes three main dimensions around which perception of social robots is organized: *warmth*, *competence*, and *discomfort*, and it is based on psychological research on human social perception that also encompasses warmth and competence attributes (Cuddy et al., 2008). In addition, it draws from and attempts to overcome the shortcomings of another widespread scale used in HRI research, the Godspeed series (Ho & MacDorman, 2010; Weiss & Bartneck, 2015).

The Godspeed series is a set of questionnaires meant to provide a reliable, validated measure of the dimensions relevant to the evaluation of social robots (Bartneck et al., 2009). The identification of these

main dimensions was conducted in a set of prior studies (Bartneck & Forlizzi, 2004; Dautenhahn, 2007; Fong et al., 2003) by extracting the main dimensions usually considered in previous research on social robotics and then evaluated by the authors in terms of its theoretical and practical relevance to the field of HRI (Bartneck et al., 2009). Despite its creation not having followed the traditional standards for scale development and validation, since its inception, the Godspeed series quickly became one of the most widely cited (and hence, used) (Weiss & Bartneck, 2015) instruments to measure HRI in a wide range of contexts (Ho & MacDorman, 2010), with small regard to its quality in terms of validity and reliability. Later analysis of the psychometric properties of this scale revealed a suboptimal level of quality with concerns about a lack of systematic approach to the original scale development and validation process, the high correlation between different scale dimensions, and nonreplication of the supposed factor structure (Ho & MacDorman, 2010; Weiss & Bartneck, 2015).

The RoSAS proposed to build on the perceptual dimensions advanced by the Godspeed series by incorporating insights from human social perception. In particular, the RoSAS builds on the premise that people make automatic social inferences about other social actors and that the process of social categorization can be extended to robots and other technological artifacts (Carpinella et al., 2017; Lang et al., 2013; Lee & Nass, 2010; Nass et al., 1994; Stroessner & Koya, 2021). This latter premise has been widely adopted and served as a background to much recent research in social robotics and computing (Lang et al., 2013). In this context, much of the research has been oriented toward the measurement of the social outcomes of the implementation of social robots in human social environments, such as effects on human engagement in prosocial behaviors (Correia et al., 2019), human disclosure of information to social robots (Noguchi et al., 2018), and the ability of robots to nudge human decision-making toward a predefined intended goal (Hashemian et al., 2019).

In developing the RoSAS, the authors considered research involving the stereotype content model (Cuddy et al., 2008). This model, prominent in social psychology, suggests that human perception is mainly anchored around two main dimensions: *warmth* and *competence*. In addition, initial factor analyses prompted the recognition of a third dimension that measures feelings of discomfort toward robots. These three dimensions (warmth, competence, and discomfort) parallel the dimensions of likeability, perceived intelligence, and security concerns (respectively) included in the Godspeed questionnaire. However, the RoSAS possessed stronger psychometric qualities (i.e., higher eigenvalues and higher levels of reliability) compared with the Godspeed questionnaire, thus offering a more parsimonious manner to evaluate individuals' perceptions of social robots.

In the original validation study, participants were asked to evaluate how closely various terms (e.g., *social*, *strange*, *reliable*) were associated with their perception of robots, using a Likert-type format scale ranging from 1 (definitely not associated) to 9 (definitely associated; see Study 2 in Carpinella et al., 2017). Factor analyses revealed the existence of three factors, each of which was composed of six items. The dimension of warmth is measured by collecting judgments of a robot on the terms (in the original version) *happy, feeling, social,*

organic, compassionate, and *emotional*. The competence dimension includes *capable, responsive, interactive, reliable, competent,* and *knowledgeable*. Finally, perceptions of discomfort are captured by the items *scary, strange, awkward, dangerous, awful,* and *aggressive*.

The original authors underline the practical potential of this scale in three main areas: (1) as a tool to evaluate preexisting robots; (2) as a way to inform the development of new robots, especially human-like robots designed to mimic human appearance and behavior, and (3) to serve as a standardized metric for those conducting HRI research (Carpinella et al., 2017). The scale was not intended to replace other metrics employed in social HRI such as the Godspeed questionnaire or measures of specific attributes. However, it presents a parsimonious and validated general scale that might be of value to researchers and developers of social robots (Carpinella et al., 2017). It has been used in over 200 studies in various countries in its brief existence.

Despite widespread international use of the RoSAS, no validated version of this scale exists in any language other than the original English. This instrument has been adapted for use in several other languages, including Portuguese (Bonani et al., 2018; Correia et al., 2019), French (Spatola et al., 2018), Danish (Suvei et al., 2018) to study how participants create and change their perception of social robots after interacting with them. However, translating and validating questionnaires in different languages is critical for ensuring that results are reliable, valid, and methodologically sound (Colina et al., 2017; Tsang et al., 2017).

2 | GOALS AND HYPOTHESIS

Our goal was to validate and evaluate the psychometric properties of the Portuguese translated version of the RoSAS scale. The scale validation was preregistered using Open Science Framework preregistration services (see Anonymous). In particular, we evaluated the convergent validity using the Godspeed scale and the divergent validity using the Negative Attitudes toward Robots Scale (NARS; Nomura et al., 2006; Piçarra et al., 2015). We also evaluated test–retest validity, construct validity, and conducted a confirmatory factor analysis (CFA) to validate the arrangement and item allocation of the original scale and the model fit of the translated scale. In this context, the following hypothesis were formulated and tested:

1. The CFA of the Portuguese version of RoSAS would replicate the three-dimensional structure (i.e., warmth, competence, and discomfort) observed in the original English version of the scale.
2. The Portuguese version of RoSAS would exhibit a good level of test–retest reliability over a 2-week interval between applications.
3. Warmth and competence would be positively correlated with the dimensions of likeability, perceived intelligence, and perceived safety of the Godspeed questionnaire. Discomfort was expected to negatively correlate with all of the dimensions of the Godspeed questionnaire.
4. The warmth and competence dimensions of RoSAS was expected to negatively correlate to the items of NARS. In contrast, the discomfort dimension was predicted to positively correlate to the items of the NARS.

3 | METHOD

An Ethical Research Committee approved the study before data collection and all relevant ethical guidelines were followed (Anonymous).

This study was conducted in three phases. First, we translated the original English scale to Portuguese and examined its translation equivalence using the back-translation method proposed by Guillemin et al. (1993). Then, we collected data from a sample of Portuguese native speakers to evaluate the scales' psychometric properties. Finally, we collected data from the initial set of respondents to evaluate test–retest reliability.

3.1 | Sample

Using the G*Power software (version 3.1.9.6 for macOS; Faul et al., 2007, 2009), we estimated a necessary sample size of 195 participants. We included only young adults (18–35 years old) living in Portugal and excluded participants who were native speakers of other variations of Portuguese (in particular, Brazilian Portuguese).

3.2 | Procedure for validation

3.2.1 | Translation and back-translation

The original version of the RoSAS scale was translated into Portuguese (European) by two bilingual Portuguese native speakers. A third individual (one of the authors, also bilingual and native Portuguese) checked the translation and coordinated with the other two translators to solve disagreements. The back-translation was conducted by a native Portuguese individual with academic training in English translation and later checked by a bilingual native English speaker and by one of the authors (native Portuguese speaker).

A committee composed of the individual with training in English translation, one of the authors (Portuguese native speaker), and a bilingual English native speaker reviewed the back-translated version and agreed upon a final version. Disagreements were resolved using a decentering technique and adding multiple synonym alternatives to accommodate all different possible translations, thus increasing redundancy and improving the likelihood of better comprehension (Brislin, 1973; Feinstein, 1987). The original study instructions for participants were also analyzed, and changes were implemented to improve clarity before the start of the data collection.

3.2.2 | Pilot test

To ensure that the questionnaire was easy to understand, we asked a convenience sample of 10 native Portuguese speakers to answer the survey. Participants were provided a printed version of the questionnaire and told to pay attention to the phrasing of the questions. They were asked to indicate if there was any instruction or sentence that was difficult to understand. Small changes were made to the questionnaire at this stage based on participants' comments.

3.2.3 | Main data collection

A convenience sample of participants was recruited using a snowball strategy for disseminating the survey in social media and through a voluntary pool of participants. The latter group of participants was compensated for their participation with course credits. The data collection was completed using the Qualtrics platform for online surveys.

Participants were told that the study's main aim was to investigate people's perceptions of certain social groups, so as not to influence their responses or willingness to participate in the study. After providing informed consent, participants responded to the following instruments:

- The Portuguese version of the RoSAS scale asking participants to indicate the extent to which they associated various words with social robots. The words presented correspond to the 18 translated items of the original RoSAS. Participants responded using a 9-point scale, with lower values indicating weaker associations and higher values reflecting stronger associations. In addition, one attention-check item requesting that participants select a particular response value was also included.
- The Portuguese Negative Attitudes toward Robots Scale (PNARS; initially developed by Nomura et al., 2006 and validated for the Portuguese population by Piçarra et al., 2015) composed of 12 items. NARS is divided into three subscales: one relative to negative attitudes toward robots with human traits (NARHT), one relative to negative attitudes toward interactions with robots (NATIR), and a third one concerning negative attitudes toward the social influence of robots. This scale includes items like "*I would feel very nervous just standing in front of a robot*" and "*I feel that if I depend on robots too much, something bad might happen*." Higher scores indicate more negative attitudes toward robots. The Portuguese NARS (henceforth, PNARS) has only two factors: NATIR and NARHT, with six items each. Participants were asked to indicate the extent they agreed or disagreed with each statement, using a 7-point scale. Lower values represent lesser agreement, and higher values represent stronger agreement.
- A translated version of the Godspeed questionnaire composed of 13 items. These 13 items correspond to the subscales of likeability (5 items), perceived intelligence (5 items), and perceived safety (3 items). Participants were asked to indicate to which extent each item corresponded well to their impressions about robots, using a 5-point semantic differential scale. Lower values indicated more negative evaluations, and higher values reflected more positive evaluations.
- A set of sociodemographic questions for sample characterization. This set of questions inquired about participants' sex, age, level of education, native tongue, frequency and familiarity with psychology studies, and previous interaction with robots.

The order of presentation of all scale items was randomized. After responding to these questionnaires, participants were presented with a debriefing statement, which included an explanation of the study goals. They were also invited to participate in the second part of the study, to take place 15 weekdays after their initial participation.

Participants who agreed to participate in the second stage were contacted via email 15 days later. This message provided a link for the second part of the study, and participants were given 3 days to respond. Email information was stored separately from their responses to the first part of the study and deleted immediately after the conclusion of data collection to avoid any risks to participants' right to anonymity and confidentiality.

The second part of the study was composed of the same Portuguese version of the RoSAS scale. Participants' responses to both parts of the questionnaire were linked using a personal unique identification code.

After completing the second part of the study, participants were again presented with a debriefing statement explaining the goal of this study. In addition, they were provided the first author's email contact to use in case of questions, suggestions, or requests. Participation in the first part of the study took approximately 15 min, and the second part took approximately 5 min to complete.

4 | RESULTS

4.1 | Data analysis procedure

The data were analyzed using IBM SPSS and AMOS software (both version 26). The scales presented some missing values, the distribution of which was analyzed using Little's MCAR test. This test revealed that a few existing missing values were missing at random ($p > .005$) and were replaced using the expectation–maximization method.

To analyze the structural properties of the scales, we conducted CFA. The best factor organization was achieved iteratively by examining item loadings onto each dimension (removing items with loadings <0.40), analyzing the consistency of the items of each subscale, and assessing the impact of (removing or maintaining) each item on the subscale consistency. Moreover, the suitability of the data for structure detection was evaluated using the Kaiser–Meyer–Olkin and Bartlett's test of sphericity.

Next we conducted structural equation modeling (SEM), and according to recommendations (Boateng et al., 2018; Ximénez, 2006), we report the following indicators for goodness-of-fit: the χ^2/df, the square root mean residual (SRMR), the comparative fit index (CFI), and the root mean square error of approximation (RMSEA). The following thresholds for each indicator were considered acceptable: $\chi^2/df < 3$, SRMR < 0.08, CFI > 0.09 and RMSEA < 0.08. In terms of validity and reliability, we sought average variance extracted (AVE) values higher than 0.5 and squared AVE values superior to maximum shared variance values (for convergent and discriminant validity, respectively); and composite reliability values superior to 0.7.

Following recommendations to improve goodness-of-fit (Hu & Bentler, 1999; Piçarra et al., 2015; Schumacker & Lomax, 2010), we considered the factor loadings of each item of the scale, the modification indices, the presence of cross-loadings, and interitem correlations.

Temporal reliability was assessed by analyzing Cronbach's alpha (α). In this context, if $.06 > \alpha \geq .05$, we considered the reliability to be poor; if $.07 > \alpha \geq .06$, we considered the reliability to be questionable; if $08 > \alpha \geq .07$, we considered the reliability to be acceptable; if $.09 > \alpha \geq .08$, we considered the reliability to be good. Cronbach's α below .05 and superior to .09 were considered unacceptable and excellent reliability indicators, respectively.

Bivariate correlations were calculated between all items of the RoSAS scale and between its subdimensions and the subdimensions of the other scales included in this study, using both the entire sample and separately for each sex.

Independent t-tests were conducted to examine potential mean differences in the participant's scores for each dimension of the RoSAS according to previous interaction with robots (coded in binary terms: with and without) and according to sex.

Across all analyses, we used the standard $\alpha = .05$ cut-off criterion. Thus, the null hypothesis was rejected when $p < .05$ and not rejected when $p > .05$.

4.2 | Descriptive statistics

We collected data from 205 participants. Of those 205, 20 were excluded for failing to answer correctly to the attention check item. A total valid sample of 185 participants was used for the preliminary validation analysis. Most of the participants in this study were female (~55%), with ages ranging between 18 and 35 years ($M = 23.40$; $SD = 5.21$). Approximately 43% of the participants reported having a university degree (bachelor's degree: 28.10%; master's degree: 14.60%; doctoral degree: 0.50%). The participants reported being somewhat familiar with research in psychology ($M = 3.56$; $SD = 1.76$) and participated in psychology studies frequently ($M = 4.15$; $SD = 1.97$). Most participants reported either never interacting with a robot ($n = 84$) or interacting with a robot at least once ($n = 82$). The remaining 13 participants reported that interacting with robots is something they consider a normal activity. All participants were native speakers of Portuguese, and almost all participants identified themselves as being Caucasian Europeans.

Of the original 185 participants, 85 participated in the second part of the study after the 15 weekday interval.

4.3 | Structure of the RoSAS

We initially conducted a CFA using maximum likelihood estimation to evaluate the goodness-of-fit, allowing each item to load only on one factor and the factors to correlate. In the initial CFA, we observed that the original scale structure did not present a good fit to our data, with several items not loading onto the expected dimensions (see Table S1).

After iteratively making several adjustments to the model (see details of the 7 models that were tested in Table 1), we achieved a solution that retained 11 items organized in the three original proposed dimensions: warmth (emotional, feeling, happy, and compassionate), discomfort (dangerous, scary, strange, and aggressive), and competence (competent, interactive, and responsive). The final solution obtained also indicated both good levels of convergent and discriminant validity and good levels of reliability (see Table 1).

In a second CFA of the final solution, all these items presented loadings higher than 0.40 to their respective categories (loadings ranging between 0.98 and 0.58; see Table S1). The interitem correlations are presented in Table S3

All subdimensions presented good levels of internal consistency (ranging between 0.74 and 0.84; see Table S5); together, these three dimensions explain 58.17% of the variance observed (see Table S1).

In terms of the data collected, we found that men and women differ in their perceptions of competence associated with social robots, but not in their perceptions of warmth or discomfort (see Table 2).

Interestingly, in our sample, previous interaction did not appear to offer an advantage in terms of the perception of social robots, with participants expressing similar views on social robots in all dimensions regardless of experience (all $p > .05$).

Analyzing the overall averages for each dimension (see Table 2), we see that, regardless of sex, participants reported low perceptions of warmth, suggesting that they typically do not associate warmth attributes with social robots.

4.3.1 | Temporal reliability

The pre–posttest correlations for each item are presented in Tables S4 and S5 provide the interitem correlations according to sex.

The warmth dimension presented a good level of correlation between the two applications (0.83, $p < .001$). However, both the discomfort (0.61, $p < .001$) and the competence dimension (0.41, $p < .001$) presented poor and unacceptable levels of correlation, respectively, between measurements. Despite this low level of correlation for the competence and discomfort dimensions, none of the subdimensions presented significant mean differences when comparing the initial full-sample average and the subsample that responded to the questionnaire a second time 15 days later (all $p > .05$).

When analyzing the separate correlations between the two applications of RoSAS according to sex, we observed that the majority of items presented good levels of correlation (i.e., >0.8) for men, but that the correlation levels for the same variables in women broadly fell below the minimum threshold for acceptability (i.e., 0.50).

4.4 | Convergent and divergent validity

Before analyzing the correlations between the RoSAS and the PNARS and Godspeed scales, we examined the psychometric properties of the latter two questionnaires.

TABLE 1 Model fit indicators for the models estimated through the iterative process

Model	χ^2	df	χ^2/df	p	CFI	RMSEA	SRMSR	AIC	AVE 1	AVE 2	AVE 3	CR 1	CR 2	CR 3	MSV 1	MSV 2	MSV 3
1	317.90	132	2.41	.000	0.91	0.08	0.12	395.90	0.47	0.48	0.46	0.83	0.82	0.80	0.008	0.008	0.000
2	252.20	116	2.17	.000	0.91	0.08	0.10	326.18	0.48	0.44	0.52	0.84	0.82	0.81	0.03	0.009	0.03
3	205.90	101	2.04	.000	0.92	0.07	0.09	275.92	0.47	0.45	0.57	0.83	0.83	0.81	0.07	0.009	0.07
4	164.00	87	1.89	.000	0.94	0.07	0.09	230.04	0.48	0.47	0.56	0.84	0.82	0.81	0.07	0.005	0.07
5	146.10	74	1.97	.000	0.92	0.07	0.08	208.13	0.47	0.47	0.31	0.83	0.82	0.57	0.02	0.005	0.02
6	127.30	62	2.05	.000	0.93	0.07	0.08	185.28	0.55	0.47	0.31	0.86	0.82	0.57	0.02	0.004	0.02
7	58.2	41		.04	0.98	0.05	0.06	108.24	0.59	.51	.61	.85	.81	.81	.04	.004	.04

Note: Model 1 includes all the items of the original scale. Model 2 excludes the item "reliable." Model 3 excludes the items "reliable" and "knowledgeable." Model 4 excludes the items "reliable," "knowledgeable," and "awkward." Model 5 excludes the items "reliable," "knowledgeable," "awkward," and "competent." Model 6 excludes the items "reliable," "knowledgeable," "awkward," "social," "capable," and "organic." Model 7 excludes the items "reliable," "knowledgeable," "awkward," "awful," "social," "capable," and "organic." Factor 1 is warmth, Factor 2 is discomfort, and Factor 3 is competence.

Abbreviations: AVE, average variance extracted; CFI, comparative fit index; CR, composite reliability; MSV, maximum shared variance; RMSEA, root mean square error of approximation.

TABLE 2 Average scores for the warmth, discomfort, and competence dimensions of the Portuguese RoSAS

| | M ± SD | | | | | |
Dimension	Overall	Women	Men	t	p	d
Warmth	2.32 ± 1.56	2.14 ± 1.46	2.53 ± 1.65	1.69	.09	
Discomfort	4.15 ± 1.70	4.26 ± 1.56	4.02 ± 1.85	−0.97	.34	
Competence	5.78 ± 1.67	5.47 ± 1.63	6.16 ± 1.65	0.79	.005	

Note: N = 185. The *t* and *p* values presented refer to comparing the mean scores for each category between men and women.
Abbreviation: RoSAS, Robotic Social Attributes Scale.

The PNARS presented good levels of internal consistency both for the NATIR dimension (0.83) and for the NARHT dimension (0.80), superior to those reported in the validation study (0.75 and 0.73, respectively, Piçarra et al., 2015; see Table S6 for consistency values according to sex).

The Godspeed items also presented good levels of consistency for perceived likeability and perceived intelligence; however, the perceived safety dimension presented unacceptable consistency (see Table S6). Furthermore, upon inspecting the factorial structure of the Godspeed scales, we observed that the items did not present the expected structure (see Table S2). In particular, the item "quiescent" did not present acceptable loadings (i.e., >0.40) in any of the dimensions. As such, we removed this item and estimated a second model.

The removal of this item generated a different model that presented a conceptual structure distinct from the original scale and of the organization of the first model estimated. Because the resulting factors could not be aggregated coherently in terms of their meaning (see Table S2), we opted to examine each of these three models separately (see Table 3). The interdimension correlations, however, were calculated only for the first and second model, given that the lack of coherent organization of the items of Model 3 made it difficult to interpret constructively (see Table S2).

As expected, we found the RoSAS' competence dimension to be negatively correlated with both the NARHT and NATIR dimensions of PNARS; and the discomfort dimension to be positively correlated with both. The warmth dimension was negatively correlated with NARHT but was not significantly correlated with NATIR (see Tables S7 and S8).

Regarding the Godspeed scales, we found that warmth was not significantly correlated with perceived likeability in neither of the models we estimated. However, warmth was negatively correlated with perceived intelligence in Model 1 (corresponding to the scale's original configuration) and with perceived intelligence and safety in Model 2 (estimated based on the organization of the scale's items in the CFA).

In addition, as expected, we also found that the discomfort dimension of RoSAS was negatively correlated with all of the Godspeed dimensions.

When considering the original configuration of the Godspeed scales, we found competence to be positively correlated to perceived intelligence and likeability. When considering the alternative configuration, we found that competence was positively correlated with perceived likeability but not with other dimensions.

As expected, all of the dimensions of the Godspeed scales were negatively associated with both dimensions of the PNARS.

5 | DISCUSSION, LIMITATIONS, AND CONCLUSION

The use of subjective measures is common in many areas of research, including HRI. In this area, subjective measures can determine how users perceive certain aspects of social robots or their interaction with them, influencing how future social robots are designed and developed. Thus, it follows from the important practical implications of the results that access to validated subjective measures that present both good levels of reliability and validity is a fundamental necessity for researchers in this area.

RoSAS is a scale developed by Carpinella et al. (2017) that has been used to evaluate users' perceptions of social robots in multiple studies. In this paper, we sought to validate the RoSAS for the European Portuguese-speaking population. To investigate the psychometric properties of this scale, we collected responses from a sample of 185 native Portuguese individuals. We examined the scale's reliability (both across time and items) and criterion validity (in relation to the PNARS and Godspeed scales).

Based on SEM, we propose a shortened version of the original scale in Portuguese, with 11 items instead of the original 18, while maintaining the original three-factor structure. The Portuguese version of RoSAS also presents good psychometric properties, satisfying all the relevant recommended criteria to evaluate model fit, composite reliability, and convergent and divergent validity (see Table 1).

In terms of criterion validity and following our hypotheses, we found that competence is negatively associated with both dimensions of PNARS; and that comfort is positively associated with them. These results were expected as PNARS items measure general negative attitudes. In other words, participants who associated robots with warmth and perceived them to be competent tended to have fewer negative attitudes toward them. In contrast, participants who report higher levels of discomfort with social robots tended to have more negative attitudes toward them.

We found only partial support regarding our hypothesis about the warmth dimension, which we expected to be negatively

TABLE 3 Scale configurations of the RoSAS and Godspeed scales and translated items

RoSAS			Godspeed (Model 1)			Godspeed (Model 2)		
Warmth	Discomfort	Competence	Perceived likeability	Perceived safety	Perceived intelligence	Perceived likeability	Perceived safety	Perceived intelligence
Emotional (Emocional)	Dangerous (Perigoso)	Competent (Competente)	(Dis)Like/ (Gosto/ Não gosto)	Anxious/Relaxed (Ansioso/Relaxado)	(In)Competent (Competent/ Incompetente)	(Dis)Like/ (Gosto/ Não gosto)	Anxious/Relaxed (Ansioso/ Relaxado)	(In)Competent (Competent/ Incompetente)
Feeling (Sentimental)	Strange (Estranho)	Interactive (Interativo)	(Un)Friendly (Amigável/Pouco Amigável)	Agitated/Calm (Agitado/Calmo)	Ignorant/ Knowledgeable (Ignorante/Culto)	(Un)Friendly (Amigável/Pouco Amigável)	Agitated/Calm (Agitado/Calmo)	Ignorant/ Knowledgeable (Ignorante/Culto)
Happy (Feliz)	Aggressive (Agressivo)	Responsive (Responsivo)	(Un)kind (Simpático/ Pouco simpático)	Quiescent/Surprised (Passivo/ Surpreendido)	(I)Responsible (Responsável/ Irresponsável)	(Un)kind (Simpático/ Pouco simpático)		(I)Responsible (Responsável/ Irresponsável)
Compassionate (Empático)	Scary (Assustador)		(Un)Pleasant (Agravável/Pouco agradável)		(Un)Intelligent (Inteligente/Pouco Inteligente)	(Un)Pleasant (Agravável/Pouco agradável)		(Un)Intelligent (Inteligente/Pouco Inteligente)
			Awful/Nice (Péssimo/ Fantástico)		Foolish/Sensible (Pateta/Sensível)	Awful/Nice (Péssimo/ Fantástico)		Foolish/Sensible (Pateta/Sensível)

Abbreviation: RoSAS, Robotic Social Attributes Scale.

associated with both dimensions of PNARS. In our data, this variable was negatively associated with NARHT, but not NATIR. This finding suggests that the items of the warmth dimension seem to capture (or evoke) essentially aspects related to the perception of robots with human traits and less to aspects related to interaction with robots.

Besides PNARS, we also examined the relationships between responses to RoSAS items and Godspeed items. We did this even though the Godspeed scales have not yet been validated for the European-speaking Portuguese population. Nonetheless, we pursued this issue because of the conceptual proximity between these measures and because the Godspeed has also been used many times in studies with Portuguese participants. In this context, we argue that its use and the information collected during this study is potentially informative and valuable for the community.

The Godspeed scales used in this study did not present good psychometric qualities. As proposed by Bartneck et al. (2009), the original model did not fit well with our data, with the items displaying a different organization than the one expected. In our study, the items allowed structure detection explaining approximately 49% of variance. However, the resulting factorial structure (see Table S2) did not correspond conceptually to the dimensions it proposed to measure (i.e., perceived likeability, perceived intelligence, and perceived safety).

The final factorial structure revealed a considerable overlap between items that were supposed to belong to the perceived intelligence dimension and items that were supposed to fit in the perceived safety dimension (Factor 1 of the final model in Table S2). The items belonging to the perceived likeability dimension in the original model were divided into two factors in our analysis (Factor 2 including (Dis)Like; Awful/Nice and (Un)Pleasant and Factor 3 including (Un)Friendly and (Un)Kind; see Tables 3 and S2).

In addition, we cannot compare our findings regarding the factorial structure of the Godspeed scales to other validation and evaluation attempts, given that, to the best of our knowledge, none exist. The original proposal collected and summarized a small number of items that had been used in previous empirical studies to measure the concepts of interest and reported only internal consistency measures (Bartneck et al., 2009). A later qualitative analysis of the use of the Godspeed scales revealed more in terms of its psychometric properties. However, this analysis added two main limitations in its use: one regarding the inconsistency in the use of the initially proposed answer format, and another observation concerning the participants' concerns regarding the repetitiveness of the items and the difficulty in assigning some high-level attributes to the robots.

It is notable in our analysis that, contrary to our expectations, we found no correlation between warmth and perceived likeability. The consistent findings of both models estimated for these scales seem to indicate that (a) warmth is negatively correlated to perceived intelligence, (b) competence is positively correlated with likeability, and (c) discomfort is negatively correlated with all Godspeed dimensions. When considering the original configuration of the Godspeed dimensions, competence presents a positive correlation with perceived intelligence, as expected.

In terms of the temporal reliability of the scale, we found overall poor values of correlation between pre–postmeasurements. However, these poor levels of temporal reliability seem to be partially tied to the sex of the respondent, with women consistently presenting worse values of reliability across all dimensions compared with their male counterparts.

In general, all subscales presented acceptable or good levels of internal consistency. One exception was the set of items that originally composed the subdimension of perceived safety (which can likely be explained, in part, by the weak loadings of the "quiescent" item on this dimension).

Regarding the relation between dimensions of the RoSAS, we found a positive correlation between the warmth and competence dimensions, but no other significant correlations, suggesting that participants who perceive robots as being more competent also perceived them higher in terms of warmth.

This study focused on developing and utilizing a psychometrically validated general scale of robotic social perception in a new language, broadening the tools available to HRI researchers and robot designers. We look forward to the development of RoSAS scales adapted for other cultural settings with a continued focus on the psychometric properties of the scales as it is translated into other languages.

ACKNOWLEDGMENT

CONFLICT OF INTERESTS

The authors declare no conflict of interests.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/hbe2.311.

DATA AVAILABILITY STATEMENT

Data available on request from the authors.

REFERENCES

Anzalone, S. M., Boucenna, S., Ivaldi, S., & Chetouani, M. (2015). Evaluating the engagement with social robots. *International Journal of Social Robotics*, 7(4), 465–478.

Bartneck, C., & Forlizzi, J. (2004). A design-centred framework for social human-robot interaction. *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)*, 591–594.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1), 71–81.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, 6, 149.

Bonani, M., Oliveira, R., Correia, F., Rodrigues, A., Guerreiro, T., & Paiva, A. (2018). What my eyes can't see, a robot can show me. *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility - ASSETS '18*. https://doi.org/10.1145/3234695.3239330

Brislin, R. (1973). Questionnaire wording and translation. In *Cross-cultural research methods* (pp. 32–58). John Wiley & Sons.

Carpinella, C. M., Wyman, A. B., Perez, M. A., & Stroessner, S. J. (2017). The Robotic Social Attributes Scale (RoSAS): Development and validation. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 254–262.

Colina, S., Marrone, N., Ingram, M., & Sánchez, D. (2017). Translation quality assessment in health research: A functionalist alternative to Back-translation. *Evaluation & the Health Professions*, 40(3), 267–293.

Correia, F., Mascarenhas, S. F., Gomes, S., Arriaga, P., Leite, I., Prada, R., Melo, F. S., & Paiva, A. (2019). Exploring prosociality in human-robot teams. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 143–151.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology*, 40, 61–149. https://doi.org/10.1016/s0065-2601(07)00002-0

Dautenhahn, K. (2007). Socially intelligent robots: Dimensions of human-robot interaction. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 362(1480), 679–704.

Edwards, A., Edwards, C., Westerman, D., & Spence, P. R. (2019). Initial expectations, interactions, and beyond with social robots. *Computers in Human Behavior*, 90, 308–314.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41, 1149–1160.

Feinstein, A. R. (1987). The theory and evaluation of sensibility. *Clinimetrics*, 141, 166.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3), 143–166.

Guillemin, F., Bombardier, C., & Beaton, D. (1993). Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines. *Journal of Clinical Epidemiology*, 46(12), 1417–1432.

Gutierrez, F. J., Ochoa, S. F., Cornejo, R., & Vassileva, J. (2019). Designing computer-supported technology to mediate intergenerational social interaction: A cultural perspective. In S. Sayago (Ed.), *Perspectives on Human-Computer Interaction Research with Older People* (pp. 199–214). Springer International Publishing.

Hashemian, M., Paiva, A., Mascarenhas, S., Santos, P. A., & Prada, R. (2019). Social power in human-robot interaction: Towards more persuasive robots. *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 2015–2017.

Ho, C.-C., & MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior*, 26(6), 1508–1518.

Höflich, J. R., & El Bayed, A. (2015). Perception, acceptance, and the social construction of robots—Exploratory studies. In J. Vincent, S. Taipale, B. Sapio, G. Lugano, & L. Fortunati (Eds.), *Social robots from a human perspective* (pp. 39–51). Springer International Publishing.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55.

Katz, J. E. (2017). *Machines that become us: The social context of personal communication technology*. Routledge.

Lang, H., Klepsch, M., Nothdurft, F., Seufert, T., & Minker, W. (2013). Are computers still social actors? In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on—CHI EA '13*. https://doi.org/10.1145/2468356.2468510

Lee, J.-E. R., & Nass, C. I. (2010). Trust in computers. In D. Latusek & A. Gerbasi (Eds.), *Trust and technology in a ubiquitous modern environment* (pp. 1–15). Information Science Reference. https://doi.org/10.4018/978-1-61520-901-9.ch001

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Conference Companion on Human Factors in Computing Systems - CHI '94*. https://doi.org/10.1145/259963.260288

Nikitina, E. A. (2007). Techno-socialization of a human being. *Philosophy & Engineering*, 75.

Nomura, T., Kanda, T., & Suzuki, T. (2006). Experimental investigation into the influence of negative attitudes toward robots on human–robot interaction. *AI & SOCIETY*, 20(2), 138–150.

Noguchi, Y., Kamide, H., & Tanaka, F. (2018). Effects on the self-disclosure of elderly people by using a robot which intermediates remote communication. *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 612–617.

Piçarra, N., Giger, J. C., Pochwatko, G., & Gonçalves, G. (2015). Validation of the Portuguese version of the negative attitudes towards robots scale. *European Review of Applied Psychology*, 65(2), 93–104.

Pan, M. K., Croft, E. A., & Niemeyer, G. (2017). Validation of the robot social attributes scale (RoSAS) for human-robot interaction through a human-to-robot handover use case. *International Journal of Robotics Research*, 36, 5–7.

Pan, M. K. X. J., Matthew K X, Croft, E. A., & Niemeyer, G. (2018). Evaluating social perception of human-to-robot handovers using the robot social attributes scale (RoSAS). In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction - HRI '18*. https://doi.org/10.1145/3171221.3171257

Schumacker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling* (3rd ed.). Routledge Academic.

Spatola, N., Belletier, C., Normand, A., Chausse, P., Monceau, S., Augustinova, M., Barra, V., Huguet, P., & Ferrand, L. (2018). Not as bad as it seems: When the presence of a threatening humanoid robot improves human performance. *Science Robotics*, 3(21), eaat5843.

Stroessner, S. J., & Koya, P. D. (2021). Thinking socially about the nonsocial world. In *Oxford handbook of social cognition*. Oxford University Press.

Strohkorb Sebo, S., Traeger, M., Jung, M., & Scassellati, B. (2018). The ripple effects of vulnerability: The effects of a Robot's vulnerable behavior on Trust in Human-Robot Teams. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 178–186.

Suvei, S.-D., Vroon, J., Somoza Sanchéz, V. V., Bodenhagen, L., Englebienne, G., Krüger, N., & Evers, V. (2018). "I would like to get close to you": Making robot personal space invasion less intrusive with a social gaze Cue. In M. Antona & C. Stephanidis (Eds.), *Universal access in human-computer interaction. Virtual, augmented, and intelligent environments* (pp. 366–385). Springer.

Tsang, S., Royse, C. F., & Terkawi, A. S. (2017). Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine. *Saudi Journal of Anaesthesia*, 11(Suppl 1), S80–S89.

Weiss, A., & Bartneck, C. (2015). Meta analysis of the usage of the Godspeed questionnaire series. *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 381–388.

Ximénez, C. (2006). A Monte Carlo study of recovery of weak factor loadings in confirmatory factor analysis. *Structural Equation Modeling*, 13(4), 587–614.

AUTHOR BIOGRAPHIES

Raquel Oliveira is a PhD candidate in Social and Organizational Psychology at ISCTE-IUL. She has a bachelor's degree in psychology and a masters' degree in Social and Organizational Psychology. Her research interests include humor, human–robot interaction and human–computer interaction in group entertainment settings.

Patrícia Arriaga is an assistant professor at ISCTE-IUL and a researcher at CIS-IUL. She has a PhD in Social and Organizational Psychology by ISCTE (2008), a scholar degree in Psychology (1996) and a master's degree in Clinical Psychology and Psychopathology by ISPA (2000). Her research focus has been on the study of emotions applied to several topics in social and health psychology. More recently she is involved in projects related to the study of human–robot interaction and the development of healthcare multimedia tools to promote well-being.

Steven J. Stroessner is a Professor of Communication at the University of California, Los Angeles. His research examines cognitive and motivational aspects of stereotyping and prejudice, and he has published extensively on social categorization and the utilization of stereotypes in judgment. More recently, his research has examined social categorization processes in judgments of nonsocial entities ranging from shapes to robots.

Ana Paiva is a Full Professor in the Department of Computer Engineering at Instituto Superior Técnico (IST) and a senior researcher of GAIPS—"Group on AI for People and Society" at INESC-ID. She is also the Katherine Hampson Bessell Fellow at the Radcliffe Institute for Advanced Study at Harvard University (2020–2021). Prof. Paiva investigates the creation of complex systems using an agent-based approach, with a special focus on social agents. Her main research focuses on the problems and techniques for creating social agents and robots that can simulate human-like behaviors, be transparent, trustworthy and are able to collaborate with humans in a natural way. She served on the Global Agenda Council in Artificial Intelligence and Robotics of the World Economic Forum and as a member of the Scientific Advisory Board of Science Europe. She is an EuroAI fellow, an ELLIS fellow, and a AAAI Councilor.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.