Computers in Human Behavior 26 (2010) 1508–1518

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices

Chin-Chang Ho, Karl F. MacDorman *

Indiana University School of Informatics, 535 West Michigan Street, Indianapolis, IN 46202, USA

ARTICLE INFO

Article history:
Available online 8 June 2010

Keywords:
Affective appraisal
Embodied agents
Human–robot interaction
Psychometric scales
Social perception

ABSTRACT

Mori (1970) proposed a hypothetical graph describing a nonlinear relation between a character's degree of human likeness and the emotional response of the human perceiver. However, the index construction of these variables could result in their strong correlation, thus preventing rated characters from being plotted accurately. Phase 1 of this study tested the indices of the Godspeed questionnaire as measures of humanlike characters. The results indicate significant and strong correlations among the relevant indices (Bartneck, Kulić, Croft, & Zoghbi, 2009). Phase 2 of this study developed alternative indices with nonsignificant correlations ($p > .05$) between the proposed *y*-axis *eeriness* and *x*-axis *perceived humanness* ($r = .02$). The new *humanness* and *eeriness* indices facilitate plotting relations among rated characters of varying human likeness.

1. Plotting emotional responses to humanlike characters

Mori (1970) proposed a hypothetical graph describing a nonlinear relation between a character's degree of human likeness and the emotional response of the human perceiver (Fig. 1). The graph predicts that more human-looking characters will be perceived as more agreeable up to a point at which they become so human people find their nonhuman imperfections unsettling (MacDorman, Green, Ho, & Koch, 2009; MacDorman & Ishiguro, 2006; Mori, 1970). This dip in appraisal marks the start of the uncanny valley (*bukimi no tani* in Japanese). As characters near complete human likeness, they rise out of the valley, and people once again feel at ease with them. In essence, a character's imperfections expose a mismatch between the human qualities that are expected and the nonhuman qualities that instead follow, or vice versa. As an example of things that lie in the uncanny valley, Mori (1970) cites corpses, zombies, mannequins coming to life, and lifelike prosthetic hands.

Assuming the uncanny valley exists, what dependent variable is appropriate to represent Mori's graph? Mori referred to the *y*-axis as *shinwakan*, a neologism even in Japanese, which has been variously translated as familiarity, rapport, and comfort level. Bartneck, Kanda, Ishiguro, and Hagita (2009) have proposed using *likeability* to represent *shinwakan*, and they applied a *likeability* index to the evaluation of interactions with Ishiguro's android double, the Geminoid HI-1. Likeability is virtually synonymous with interpersonal warmth (Asch, 1946; Fiske, Cuddy, & Glick, 2007; Rosenberg, Nelson, & Vivekananthan, 1968), which is also strongly correlated with other important measures, such as comfortability, communality, sociability, and positive (vs. negative) affect (Abele & Wojciszke, 2007; MacDorman, Ough, & Ho, 2007; Mehrabian & Russell, 1974; Sproull, Subramani, Kiesler, Walker, & Waters, 1996; Wojciszke, Abele, & Baryla, 2009). Warmth is the primary dimension of human social perception, accounting for 53% of the variance in perceptions of everyday social behaviors (Fiske, Cuddy, Glick, & Xu, 2002; Fiske et al., 2007; Wojciszke, Bazinska, & Jaworski, 1998).

Despite the importance of warmth, this concept misses the essence of the uncanny valley. Mori (1970) refers to negative *shinwakan* as *bukimi*, which translates as eeriness. However, eeriness is not the negative anchor of warmth. A person can be cold and disagreeable without being eerie—at least not eerie in the way that an artificial human being is eerie. In addition, the set of negative emotions that predict eeriness (e.g., fear, anxiety, and disgust) are more specific than coldness (Ho, MacDorman, & Pramono, 2008). Thus, *shinwakan* and *bukimi* appear to constitute distinct dimensions.

Although much has been written on potential benchmarks for anthropomorphic robots (for reviews see Kahn et al., 2007; MacDorman & Cowley, 2006; MacDorman & Kahn, 2007), no indices have been developed and empirically validated for measuring *shinwakan* or related concepts across a range of humanlike stimuli, such as computer-animated human characters and humanoid robots. The Godspeed questionnaire, compiled by Bartneck, Kulić, Croft, and Zoghbi (2009), includes at least two concepts, *anthropomorphism* and *likeability*, that could potentially serve as the *x*- and *y*-axes of Mori's graph (Bartneck, Kanda, et al., 2009). Although the

* Corresponding author. Tel.: +1 317 215 7040.
E-mail address: kmacdorm@indiana.edu (K.F. MacDorman).
URL: http://www.macdorman.com (K.F. MacDorman).

0747-5632/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.chb.2010.05.015

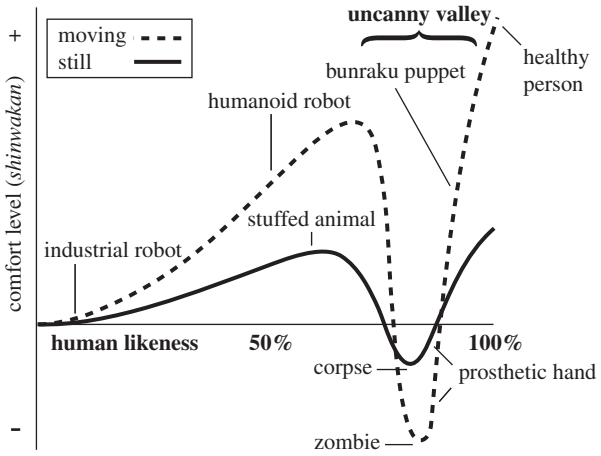

Fig. 1. Mori (1970) proposed a nonlinear relation, which is intensified by movement, between a character's degree of human likeness and the human perceiver's emotional response. The dip in emotional response just before total human likeness is referred to as the uncanny valley.

Godspeed questionnaire lists semantic differential items for each concept, the indices corresponding to these concepts have not been empirically tested as a group for overall reliability and validity. In addition, there is no index corresponding specifically to *eeriness*, a dimension that is arguably distinct from *likeability* but nevertheless important in determining whether a human-looking character has fallen into the uncanny valley.

Phase 1 of the current study evaluates the Godspeed indices based on participant ratings of computer-animated human characters and humanoid robots presented in video clips. The performance of the Godspeed indices in Phase 1 is used in Phase 2 to benchmark progress toward developing a new set of uncanny valley indices. The new set includes *eeriness* as a possible dimension for the *y*-axis in Mori's graph and decorrelates *eeriness* from *humanness* and *warmth*. Indices for *humanness*, *eeriness*, *warmth*, and *attractiveness* were developed in two rounds of testing using five methods of analysis: (1) adjectives that could serve as potential anchors for semantic differential items were selected for each index and rated on their positive (vs. negative) affect, and inversely correlated adjectives that had similar affective ratings were paired in semantic differential items; (2) reliability analysis was used to remove less reliable items from each index; (3) exploratory factor analysis was used to determine the geometric solution of the indices by oblique rotation; (4) correlation analysis was used to decorrelate the indices from interpersonal warmth; and (5) confirmatory factor analysis was used to test their theoretical structure.

2. An empirical analysis of the Godspeed indices

Bartneck, Kulić, et al. (2009) assembled five indices composed of semantic differential items in the Godspeed questionnaire to assist developers in creating embodied social agents. The indices are *anthropomorphism* (Powers & Kiesler, 2006), *animacy* (converted from Likert scales; Lee, Park, & Song, 2005), *likeability* (Monahan, 1998), *perceived intelligence* (Warner & Sugarman, 1996), and *perceived safety* (Kulić & Croft, 2007). The purpose of Phase 1 of this study is twofold: to test for the first time the validity, reliability, and theoretical structure of these indices *as a set* for a range of robots and computer-animated human characters and, specifically, to determine whether *anthropomorphism* and *likeability* are sufficiently decorrelated to serve as *x*- and *y*-axes in plotting people's emotional response to characters that vary in their degree of perceived human likeness. It should be noted that in the past development of these indices, no attempt had been made to decorrelate them from positive (vs. negative) affect or from each other. As an example of this, *anthropomorphism* and *animacy* have a semantic differential item in common, *artificial–lifelike*.

Several of the indices, including *anthropomorphism*, would appear to be correlated with positive (vs. negative) affect, interpersonal warmth, and *likeability*, based on the face validity of the opposing anchors used for their semantic differential items. For example, *fake*, *moving rigidly*, and other anchors used to indicate low anthropomorphism have a negative nuance compared to *natural*, *moving elegantly*, and other anchors used to indicate high anthropomorphism. This trend continues for *animacy* with low animacy anchors like *dead*, *stagnant*, and *apathetic* and high animacy anchors like *alive*, *lively*, and *responsive*; for *perceived intelligence* with low intelligence anchors like *ignorant*, *foolish*, and *irresponsible* and high intelligence anchors like *knowledgeable*, *sensible*, and *responsible;* and for *perceived safety* with low safety anchors like *agitated* and *anxious* and high safety anchors like *calm* and *relaxed*.

Given that interpersonal warmth is the dominant dimension of human social perception and the apparent alignment of the anchors with positive and negative affect, a general concern is that each of the Godspeed indices may not measure the concept after which it was named but instead measures some convolution of that concept and interpersonal warmth. A more specific concern for our study is that, if *anthropomorphism* and *likeability* are strongly correlated, a scatter plot of characters rated along these axes will be highly skewed (Fig. 2). The plot will not accurately depict the characters' scores on the convoluted variable, and topological relations will be distorted.

2.1. Research methods

2.1.1. Participants

Participants were recruited from a list of randomly selected undergraduate students and recent graduates of a nine-campus Midwestern university. Among the 384 participants, 161 (41.9%) were male and 223 (58.1%) were female, 187 (48.7%) were under 20 years old, 162 (42.2%) were 21 to 25 years old, and 35 (9.1%) were over 26 years old. The participants reflected the demographics of the university's undergraduate population (80.1% non-Hispanic white, 6.9% African-American, 3.4% Asian, 3.0% Hispanic, and 6.6% foreign or unclassified). With respect to the sample's representativeness of the undergraduate population as a whole, the measurement error range was ±5.0% at a 95% confidence level.

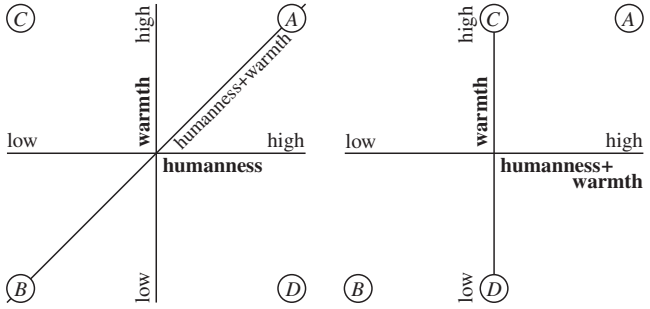

Fig. 2. Plotting an index that is a composite of two or more dimensions on a single axis distorts topological relations among observations. To illustrate this, four characters, labeled A, B, C, and D, are plotted against the *humanness* and *warmth* axes for the graph on the left and the *humanness + warmth* and *warmth* axes for the graph on the right. For the graph on the right, the degree of *humanness* of the low humanness character C and the high humanness character D cannot be distinguished. In addition, C is closer to A than B, and D is closer to B than A, although the distances should be equal.

There were no significant differences among the studies reported in this paper by gender or age.

2.1.2. Materials and procedures

Each participant viewed 10 video clips presented one at a time in random order (see Fig. 3). There were five video clips of three-dimensional computer-animated characters and five of robots. The video clips were displayed using a width of 480 pixels and a height of 360 pixels, which is a 4:3 aspect ratio. The clips were 15–30 s in length. Clips were played in a continuous loop while participants answered a survey on the figure featured in each video clip.

The survey consisted of the Godspeed questionnaire, which is composed of five indices and 24 semantic differential items. The *anthropomorphism* index has five items, the *animacy* index has six items, the *likeability* index has five items, the *perceived intelligence* index has five items, and the *perceived safety* index has three items (Table 1).

2.1.3. Statistical analysis

Cronbach's α was used to measure the reliability of each index. Confirmatory factor analysis was used to verify whether the 24 semantic differential items divide into five factors corresponding to the five Godspeed indices. If the results of confirmatory factor analysis were inconsistent with the construct dimensions, the items could not represent the concepts of the indices. In addition, correlation analysis was used to evaluate the relation among the indices and to test their discriminant validity. Multidimensional scaling (MDS) was used to create a (Euclidean) distance matrix for all pairs of the 24 semantic differential items to approximate their distance from each other in a space that has been reduced

Fig. 3. The five video clips on the top row contain computer-animated human characters from the films (1) *Final Fantasy*: *The Spirits Within*, (2) *The Incredibles*, and (3) *The Polar Express*, (4) an Orville Redenbacher popcorn advertisement, and (5) a technology demonstration of the *Heavy Rain* video game. The remaining five video clips contain (6) iRobot's Roomba 570, (7) JSK Laboratory's Kotaro, (8) Hanson Robotics's Elvis and (9) Eva, and (10) Le Trung's Aiko.

Table 1
Structural coefficients for the Godspeed indices.

Items[a]	Anthropomorphism	Animacy	Likeability	Perceived intelligence	Perceived safety
Machinelike–Humanlike	.89	–	–	–	–
Artificial–Lifelike	.87	–	–	–	–
Fake–Natural	.85	–	–	–	–
Unconscious–Conscious	.76	–	–	–	–
Moving rigidly–Moving elegantly	.76	–	–	–	–
Mechanical–Organic	–	.88	–	–	–
Artificial–Lifelike	–	.87	–	–	–
Dead–Alive	–	.79	–	–	–
Stagnant–Lively	–	.64	–	–	–
Apathetic–Responsive	–	.59	–	–	–
Inert–Interactive	–	.57	–	–	–
Awful–Nice	–	–	.86	–	–
Unpleasant–Pleasant	–	–	.85	–	–
Dislike–Like	–	–	.83	–	–
Unfriendly–Friendly	–	–	.81	–	–
Unkind–Kind	–	–	.81	–	–
Ignorant–Knowledgeable	–	–	–	.81	–
Unintelligent–Intelligent	–	–	–	.79	–
Incompetent–Competent	–	–	–	.78	–
Foolish–Sensible	–	–	–	.74	–
Irresponsible–Responsible	–	–	–	.70	–
Agitated–Calm	–	–	–	–	.84
Anxious–Relaxed	–	–	–	–	.70
Surprised–Quiescent	–	–	–	–	.19
Cronbach's α	.91	.88	.92	.87	.60
Model	χ^2	df	GFI	AGFI	
	3927.25	242	.86	.82	
	NFI	CFI	RMR	RMSEA	
	.98	.98	.086	.088	

[a] Items are sorted by the factor loading of each index.

from 24 to 2 dimensions. The distance matrix was used to visualize similarities and dissimilarities among the items. Internal reliability and correlation analysis were performed using SPSS, confirmatory factor analysis was performed using LISREL, and multidimensional scaling was performed using MATLAB.

2.2. Results

To confirm the reliability and the validity of the Godspeed indices, an internal reliability test was conducted. The results showed that the *likeability* and *anthropomorphism* indices had the highest reliability with a Cronbach's α of .92 and .91, respectively. The Cronbach's α of *animacy* and *perceived intelligence* was .88 and .87, respectively. However, *perceived safety* had low reliability with a Cronbach's α of .60, which is below the standard .70 cutoff (Nunnally, 1978).

Confirmatory factor analysis was used to test the theoretical structure of the Godspeed indices. Table 1 shows the factor loadings of the 24 semantic differential items. In the model, two goodness-of-fit indices (RMR = .086; RMSEA = .088) exceeded the standard .05 cutoff, indicating that the 24 semantic differential items did not fit well in the structure of these five indices (χ^2 = 3927.25, CFI = .98, NFI = .98, GFI = .86, AGFI = 0.82; Bentler, 1990; Chin & Todd, 1995; Gefen, Straub, & Boudreau, 2000). A serious problem was that several factor loadings could not reach a high level, such as *stagnant–lively*, *inert–interactive*, and *apathetic–responsive* for *animacy* and *surprised–quiescent* for *perceived safety*. The result is that the latent constructs could not capture more than half their variances.

Another serious problem was the significant and extremely high correlation between *anthropomorphism*, *likeability*, *animacy*, and *perceived intelligence* (Table 2). The correlations ranged from .67 for *anthropomorphism* and *perceived intelligence* to .89 for *anthropomorphism* and *animacy*. This suggests that those concepts had no discriminant validity. In other words, they were all measuring the same concept instead of measuring distinct concepts.

Multidimensional scaling was performed on the 24 semantic differential items. Fig. 4 shows that semantic differential items belonging to the *anthropomorphism* and *animacy* indices are distributed across a large overlapping region. Although the *likeability* items are packed closely together, they are wholly contained within the region circumscribed by the *anthropomorphism* and *animacy* items. The MDS results indicate that the *anthropomorphism*, *animacy*, and *likeability* indices are unable to measure distinctly their corresponding concepts.

The conclusion that the Godspeed indices lack discriminant validity is further supported by the fact that the spread of data points in a scatter plot followed a diagonal line of humanness: all the robots were located in the lower-left area, and the computer-animated human characters were located in the upper-right area (Figs. 5–7). *Likeability* was significantly (p = .000) and highly correlated with *anthropomorphism* (r = .73), *animacy* (r = .74), and *perceived intelligence* (r = .71). These findings indicate that the Godspeed indices could not measure the intended concepts independently of positive (vs. negative) affect. In addition, the *anthropomorphism* index could not separate the robots by their degree of humanness despite a nonanthropomorphic robot, Roomba 570, being included in the group.

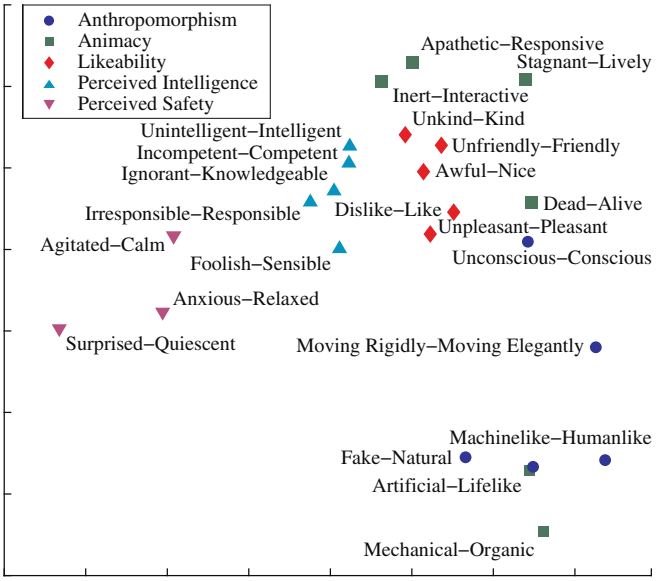

Fig. 4. Multidimensional scaling of the 24 semantic differential items was performed based on participant ratings of the figures in the 10 video clips. Items from the *anthropomorphism* and *animacy* indices are spread out across a large overlapping region, which includes the *likeability* items.

3. The development of humanness, warmth, eeriness, and attractiveness indices

The results of Phase 1 of this study found that the Godspeed indices did not represent their concepts independently of positive (vs. negative) affect. Hence, in Phase 2 an alternative set of indices is developed to measure participants' attitudes toward anthropomorphic characters: *perceived humanness*, *warmth*, *eeriness*, and *attractiveness*.

The first three indices are motivated by the original graph of the uncanny valley proposed by Mori (1970). Studies on the uncanny valley typically manipulate as an independent variable a character's "objective" humanness—the human photorealism of the character's morphology, skin texture, motion quality, or other formal property (MacDorman, Coram, Ho, & Patel, 2010; MacDorman et al., 2009; Seyama & Nagayama, 2007). However, it is also useful to have a corresponding measure of its subjective or perceived humanness to check whether the objective manipulation is having the intended effect. Interpersonal warmth is useful to include, because it is the dominant dimension of human social perception and strongly correlated with concepts identified with *shinwakan*, the *y*-axis of Mori's graph, such as comfort level, likeability, and rapport.

Table 2
Correlation between anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety.

	Anthropomorphism	Animacy	Likeability	Perceived intelligence	Perceived safety
Anthropomorphism	–				
Animacy	.89***	–			
Likeability	.73***	.74***	–		
Perceived Intelligence	.67***	.72***	.71***	–	
Perceived Safety	.06**	−.01	.20***	.17***	–

** p < .01 (2-tailed).
*** p < .001 (2-tailed).

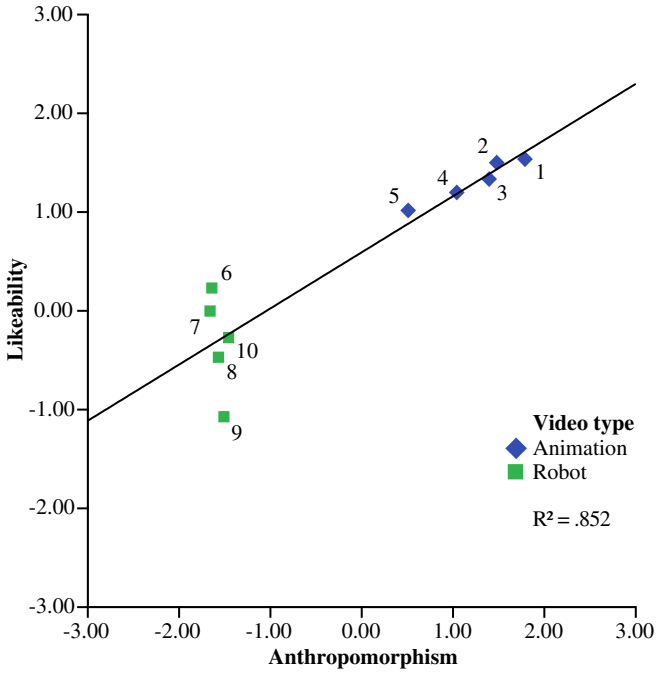

Fig. 5. The *anthropomorphism* and *likeability* indices of the Godspeed questionnaire are significantly and strongly correlated ($p = .000$, $r = .73$). The ratings of the computer-animated human characters are nearly collinear, as are the ratings of the robots. The *anthropomorphism* index is unable to discriminate the robots by their degree of humanness. The humanoid robot, Kotaro, was rated as having slightly lower *anthropomorphism* than the nonanthropomorphic robot, Roomba 570.

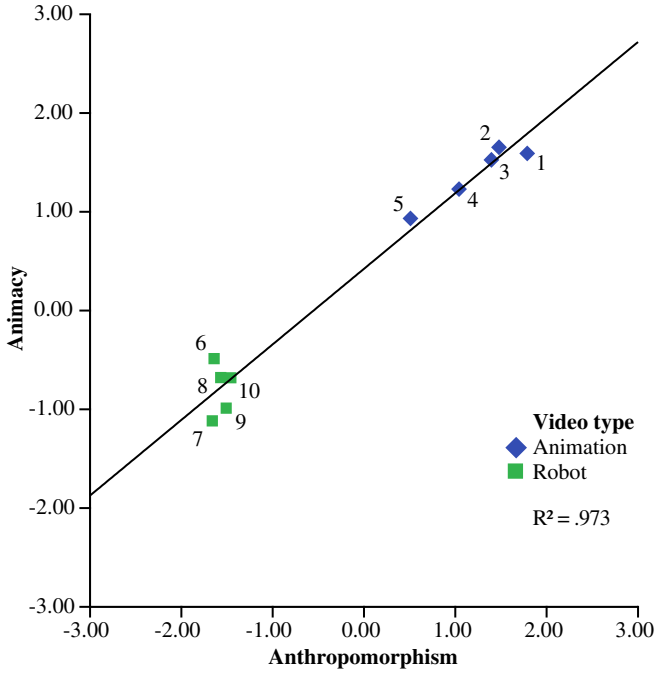

Fig. 6. The *anthropomorphism* and *animacy* indices of the Godspeed questionnaire are significantly and strongly correlated ($p = .000$, $r = .89$). This indicates they may be measuring the same concept. The ratings of the computer-animated human characters are nearly collinear, and there is little separation among the ratings of the robots.

Eeriness, which is conceptually distinct from negative warmth (i.e., interpersonal coldness), would need to be included in any set of indices on the uncanny valley, as it corresponds to the phenomenon to be explained.

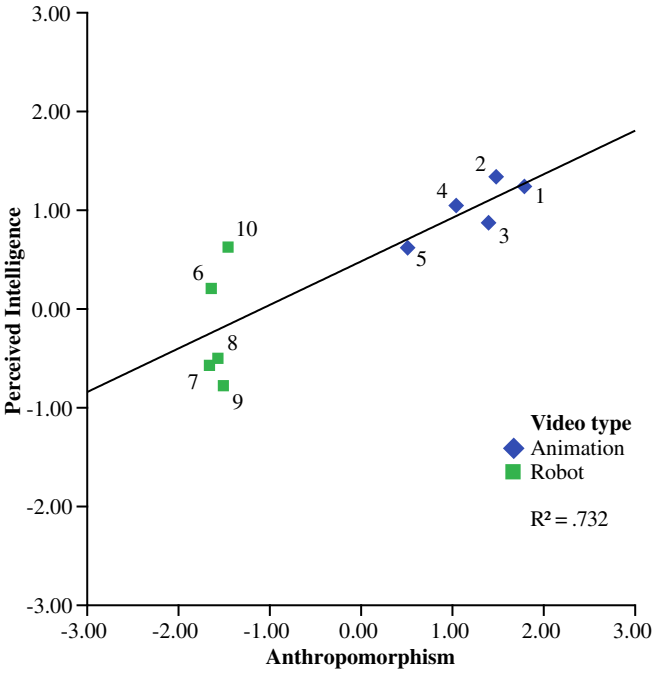

Fig. 7. The *anthropomorphism* and *perceived intelligence* indices of the Godspeed questionnaire are significantly and strongly correlated ($p = .000$, $r = .67$).

An *attractiveness* index is included, because physical attractiveness is an important dimension in explanations of the uncanny valley based on evolved perceptual and cognitive mechanisms for mate selection and pathogen avoidance (MacDorman & Ishiguro, 2006; MacDorman et al., 2009). Bilateral symmetry, clear skin, certain proportions of the face and body, and other observable markers of attractiveness are correlated with reproductive fitness as measured by a range of physiological variables, including sperm count, strength of female orgasm, hormonal and immune system levels, and the ability to conceive (Jasienska, Ziomkiewicz, Ellison, Lipson, & Thune, 2004; Jones, Little, & Perrett, 2004; Manning, Scutt, & Lewis-Jones, 1998; Thornhill & Gangestad, 1993; Thornhill, Gangestad, & Comer, 1995). There is an extensive literature exploring the evolutionary and cultural basis for perceptions of attractiveness and their pervasive impact on human behavior (Cunningham, Roberts, Barbee, Druen, & Wu, 1995; Jones, 1995; Langlois et al., 1987; Langlois et al., 2000). Attractiveness is known to influence many kinds of decisions, even without principled reasons, including decisions of moral consequence (Cunningham, 1986). Therefore, it is important to control for the effects of attractiveness in studies on the uncanny valley.

3.1. Research goal

The goal of Phase 2 of this study is to develop valid and reliable indices for *perceived humanness*, *warmth*, *eeriness*, and *attractiveness* based on corresponding semantic differential items, such that *perceived humanness* and *eeriness* are not significantly correlated with each other or with *warmth* or *attractiveness*. The naïve development of *perceived humanness* and *eeriness* indices could confound these dimensions with interpersonal warmth. If *eeriness*, for example, were strongly correlated with interpersonal warmth, wicked but artfully rendered villains might be rated eerier than amiable but uncanny-looking heroes (e.g., the queen in Walt Disney's 1937 hand-animated film *Snow White* versus the conductor in Robert Zemeckis's 2004 computer-animated film *The Polar Express*). Such an index would not be able to detect characters that

had fallen into the uncanny valley as described by Mori (1970). In this study, decorrelation between indices was achieved for *eeriness* but only partly achieved for *perceived humanness*.

Semantic differential items were used in Phase 2, because they can reduce acquiescence bias (i.e., the tendency of participants to agree with statements) without lowering psychometric quality (Friborg, Martinussen, & Rosenvinge, 2006; Lorr & Wunderlich, 1988). To decorrelate the *humanness*, *eeriness*, and *attractiveness* indices from interpersonal warmth, the opponent adjective pairs of their semantic differential items went through a process of selection to find adjectives that have about the same level of positive (vs. negative) affect. These adjectives are paired in semantic differential scales so the indices that accumulate their values are not correlated with positive (vs. negative) affect. In addition, this study attempts to adhere to the following guidelines in constructing *humanness*, *eeriness*, and *attractiveness* indices: (1) the opponent adjective pairs should be moderately or strongly inversely correlated; (2) items corresponding to a single, unidimensional concept should load on the same factor when applying exploratory factor analysis as a heuristic tool for index development (Comrey, 1978); (3) the positive and negative anchors of eeriness and humanness adjective pairs should be nearly uncorrelated with the *warmth* or *pleasure* indices, and the attractiveness item pairs should have at most a medium correlation; (4) there should be at least three semantic differential scales per index to enable the estimation of reliability; and (5) the reliability of the indices should be acceptable (Cronbach's $\alpha \geqslant .70$).

3.2. Methods

3.2.1. Participants

In the initial round of testing, there were 19 participants, 13 (68.4%) male and 6 (31.6%) female, of whom 7 (36.8%) were 21–25 years old, 4 (21.1%) were 26–30, 5 (26.3%) were 31–35, and 3 (15.8%) were over 36. Most participants were human–computer interaction (HCI) graduate students, young professionals, and HCI-related professionals.

In the second round of testing, participants were recruited from a random selection of undergraduate students and recent graduates of a nine-campus Midwestern university. Among the 253 participants, 112 (44.3%) were male and 141 (55.7%) were female, 216 (85.4%) were under 25 years old, 20 (7.9%) were 26–30, and 17 (6.7%) were over 31. The participants reflected the demographics of the university's undergraduate population. The measurement error range was ±6.16% at a 95% confidence level.

3.2.2. Materials and procedures

The video clips and method of presentation were the same as in the previous study. Each participant viewed 10 video clips presented one at a time in random order (see Fig. 3). There were five video clips of three-dimensional computer-animated characters and five of robots. The video clips were displayed using a width of 480 pixels and a height of 360 pixels, which is a 4:3 aspect ratio. Most clips were 15–30 s in length. Clips were played in a continuous loop while participants answered a survey on the figure featured in each video clip. The initial round of the survey consisted of 22 semantic differential items: seven from the *perceived humanness* index, eight from the *eeriness* index, and seven from the *attractiveness* index. The second round of the survey consisted of 29 semantic differential items: 10 from the *humanness* index, 8 from the *eeriness* index, and 11 from the *attractiveness* index.

3.2.3. Statistical analysis

Internal reliability was used to measure how reliable items were for their indices in each round of testing. Exploratory factor analysis, which applied the principal components analysis method and the Promax rotation, was used to verify that the semantic differential items loaded on factors corresponding to their named concepts. In addition, *artificial–natural* in the *humanness* index, *reassuring–eerie* in the *eeriness* index, and *unattractive–attractive* in the *attractiveness* index were chosen as "sanity check" items to verify the correctness of indices. A sanity check item has high face validity but does not necessarily meet the other criteria for an item, such as being correlated with interpersonal warmth. If the results of factor analysis varied from the sanity check's dimension and showed low factor loadings, new items should be developed and added to the index in the next round. Correlation analysis showed the relation between indices and verified the discriminant validity of indices during testing. Confirmatory factor analysis was used to verify the theoretical structure of the new set of uncanny valley indices. Finally, multidimensional scaling was used to visualize similarities and dissimilarities among the semantic differential items by reducing the dimensionality of the space from 19 to 2 dimensions. Internal reliability, exploratory factor analysis, and correlation analysis were performed using SPSS, confirmatory factor analysis was performed using LISREL, and multidimensional scaling was performed using MATLAB.

3.3. Results

3.3.1. Humanness index

A pool of seven items was initially selected for the *humanness* index (see Table 3). *Artificial–natural* was the sanity check for the *humanness* index. The overall internal reliability of the initial test was relatively high (Cronbach $\alpha = .85$). The initial exploratory factor analysis with no iterations showed all items loaded on a single factor that explained 57.33% of the variance. The reliability was improved by removing *genderless–male or female*, *uncommunicative–bigmouthed*, and *automatic–deliberate*.

These items were replaced with *inanimate–living*, *mechanical movement–biological movement*, and *synthetic–real* in the second round of testing. The internal reliability in the second round of testing remained the same. As with the initial round of testing, exploratory factor analysis extracted (with no iterations) one major factor that explained 60.79% of the variance. However, the newly added items contributed higher factor loadings than those of *genderless–male or female*, *uncommunicative–bigmouthed*, and *automatic–deliberate*.

In the final version of the index, *artificial–natural*, *human-made–humanlike*, *without definite lifespan–mortal*, *inanimate–living*,

Table 3
Reliability and factor loadings of the humanness index.

Items[a]	Round 1	Round 2	Final
Artificial–Natural[b]	.83	.87	.90
Human-made–Humanlike	.82	.85	.88
Innocent of Morals–Aware of Right and Wrong[d]	.82	.77	–
Without Definite Lifespan–Mortal	.81	.84	.85
Genderless–Male or Female[d]	.71	.63	–
Uncommunicative–Bigmouthed[d]	.66	.62	–
Automatic–Deliberate[d]	.62	.52	–
Inanimate–Living[c]	–	.86	.88
Mechanical Movement–Biological Movement[c]	–	.86	.86
Synthetic–Real[c]	–	.86	.90
Total variance explained	57.33%	60.79%	68.96%
Cronbach's α	.85	.85	.92

[a] Items are sorted by the factor loading of the initial round of testing.
[b] The sanity check.
[c] Items added in the second round of testing.
[d] Items excluded from the final version.

mechanical movement–biological movement, and *synthetic–real* were the measurement items. Therefore, the final version of the humanness index would retain six items. Its internal reliability was high (Cronbach's α = .92), and it explained 68.96% of the variance.

3.3.2. Eeriness index

A pool of eight items was initially selected for the *eeriness* index (see Table 4). *Reassuring–eerie* was the sanity check for the *eeriness* index. The overall internal reliability in the initial round of testing was .80. The initial exploratory factor analysis with three iterations showed that two major factors were extracted. *Reassuring–eerie*, *numbering–freaky*, *bland–uncanny*, and *ordinary–supernatural* loaded on the first factor, which explained 43.42% of the variance. The internal reliability of the first factor was .76. *Unemotional–hair-raising*, *uninspiring–spine-tingling*, *boring–shocking*, and *predicable–thrilling* loaded on the second factor, which explained 19.80% of the variance. The internal reliability of the second factor was .79.

Because the initial results met the reliability criterion, the second round of testing was followed by exploratory factor analysis to check whether the items represented the eeriness index appropriately. Although the internal reliability of the second round of data was .74, the exploratory factor analysis result with three iterations was similar to the initial testing. *Unemotional–hair-raising*, *uninspiring–spine-tingling*, *boring–shocking*, *predicable–thrilling*, and *bland–uncanny* loaded on the first dimension, which explained 38.40% of the variance. *Reassuring–eerie*, *numbering–freaky*, and *ordinary–supernatural* loaded on the second dimension, which explained 22.93% of the variance.

Because the two dimensions explained sufficient variance and were both relevant to the concept of eeriness, all items in the *eeriness* index were retained in the final version. For follow-up confirmatory factor analysis, the factor corresponding to the *reassuring–eerie*, *numbering–freaky*, and *ordinary–supernatural* items was referred to as *eerie*, and its internal reliability was .71; the factor corresponding to the *unemotional–hair-raising*, *uninspiring–spine-tingling*, *boring–shocking*, *predicable–thrilling*, and *bland–uncanny* items was referred to as *spine-tingling*, and its internal reliability was .81. Therefore, the final version of the *eeriness* index would retain eight items that explained 62.04% of the variance and held an overall internal reliability of .74.

3.3.3. Attractiveness index

A pool of seven items was initially selected for the *attractiveness* index (see Table 5). Opponent adjectives that were rated as having similar levels of positive (vs. negative) affect were paired in semantic differential items. *Unattractive–attractive* was the sanity check for the *attractiveness* index. The initial internal reliability was .78. The initial exploratory factor analysis with three iterations extracted two major factors. *Unpretentious–alluring*, *prim–eye-catching*, *modest–sensual*, *unadorned–showy*, and *plain-featured–racy* loaded on the first factor, which explained 44.09% of the variance. Only *homely–slick* was grouped with *unattractive–attractive* in the second factor, which explained 14.83% of the variance.

The initial result's first factor did not contain *unattractive–attractive* and thus did not appear to be measuring attractiveness. Therefore, four items were added in the second round of testing: *ugly–beautiful*, *repulsive–agreeable*, *crude–stylish*, and *messy–sleek*. The internal reliability of the data in the second round of testing was .84. Although exploratory factor analysis extracted two factors in three iterations, the four newly added items loaded on the same factor as *unattractive–attractive*, and this factor explained 39.75% of the variance. The cronbach's α of these five items was .90. The final version of the *attractiveness* index would retain these five items, which explained 70.93% of the variance. Although these items had high reliability and face validity, the opponent adjectives did not have the same level of positive (vs. negative) affect. Thus, the items would be unlikely to meet the goal of decorrelating *attractiveness* from *warmth*.

3.3.4. Pleasure and warmth indices

Sad–happy, *bad–good*, *terrible–wonderful*, and *annoyed–pleased* comprised the *pleasure* index in the initial round of testing. The internal reliability of the *pleasure* index was acceptable (Cronbach's α = .79). The *pleasure* index was used to assess the correlations among indices. If the *attractiveness*, *humanness*, and *eeriness* indices correlated highly with the *pleasure* index, it means that the positive (vs. negative) affect in these indices might dilute their discriminant validity. *Cold-hearted–warm-hearted*, *hostile–friendly*, *spiteful–well-intentioned*, *ill-tempered–good-natured*, and *grumpy–cheerful* comprised the *warmth* index in the second round of testing. The internal reliability of the *warmth* index was high (Cronbach's

Table 4
Reliability and factor loadings of the eeriness index.

Items[a]	Round 1		Round 2		Final	
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2
Reassuring–Eerie[b]	.91	−.34	−.22	.87	−.22	.87
Numbing–Freaky	.80	.06	.05	.82	.05	.82
Ordinary–Supernatural	.68	.13	.20	.67	.20	.67
Bland–Uncanny	.68	.16	.70	.09	.70	.09
Unemotional–Hair-raising	−.14	.85	.75	−.23	.75	−.23
Uninspiring–Spine-tingling	.05	.82	.78	.08	.78	.08
Predictable–Thrilling	−.08	.75	.76	−.09	.76	−.09
Boring–Shocking	.32	.66	.77	.17	.77	.17
Total variance explained	43.42%	19.80%	38.40%	22.93%	38.40%	22.93%
Cronbach's α	.76	.79	.81	.71	.81	.71
Overall Cronbach's α	.80		.74		.74	

[a] Items are sorted by the factor loading of the initial round of testing.
[b] The sanity check.

Table 5
Reliability and factor loadings of the attractiveness index.

Items[a]	Round 1		Round 2		Final
	Factor 1	Factor 2	Factor 1	Factor 2	
Unpretentious–Alluring[d]	.75	.07	.22	.57	–
Modest–Sensual[d]	.75	.02	−.10	.70	–
Plain-featured–Racy[d]	.74	−.05	−.09	.77	–
Unadorned–Showy[d]	.73	−.05	−.03	.71	–
Prim–Eye-catching[d]	.73	−.01	.07	.62	–
Homely–Slick[d]	−.15	.92	.35	.26	–
Unattractive–Attractive[b]	.21	.69	.84	.05	.87
Repulsive–Agreeable[c]	–	–	.88	−.18	.82
Ugly–Beautiful[c]	–	–	.86	.04	.88
Messy–Sleek[c]	–	–	.81	−.04	.79
Crude–Stylish[c]	–	–	.80	.06	.82
Total variance explained	44.09%	14.83%	39.75%	16.32%	70.93%
Cronbach's α	.79	.49	.87	.72	.90
Overall Cronbach's α	.78		.84		.90

[a] Items are sorted by the factor loading of the initial round of testing.
[b] The sanity check.
[c] Items added in the second round of testing.
[d] Items excluded from the final version.

α = .88). Like the *pleasure* index, the *warmth* index in the second round of testing was designed to assess its correlation with other indices. If any index showed a high correlation with the *warmth* index, its items should be modified to eliminate this correlation.

3.3.5. Validation of the final version of the indices

Based on two rounds of testing, five items were constructed for the final version of the *attractiveness* index, eight items were constructed for the *eeriness* index, and six items were constructed for the *humanness* index (Tables 3–5). Confirmatory factor analysis was used to test the theoretical structure of the final set. Table 6 shows the factor loadings for the 19 semantic differential items of the final set. Although one goodness-of-fit index (RMSEA = .075) slightly exceeded the cutoff of .05, the other goodness-of-fit indices indicated that the 19 semantic differential items fit moderately well within the structure of these indices (χ^2 = 1229.29, CFI = .97, NFI = .97, GFI = .91, AGFI = 0.88; Bentler, 1990; Chin & Todd, 1995; Gefen et al., 2000).

The correlation analysis indicated that the indices retained their construct validity (Table 7). In the final version, the *attractiveness* index had no significant correlation with *eeriness* (r = −.03, p = .316). The correlation of the *attractiveness* and *eeriness* indices with positive (vs. negative) affect was effectively eliminated. In addition, the *eeriness* index had no significant correlation with the *humanness* index (r = .02, p = .514).

Multidimensional scaling was performed on the 19 semantic differential items. Fig. 8 shows that semantic differential items belonging to the *humanness*, *eeriness*, and *attractiveness* indices are in three distinct, nonoverlapping regions. The three items belonging to the *eerie* subfactor and the five items belonging to the *spine-tingling* subfactor of the *eeriness* index (listed in Table 6) are also widely separated. These MDS results indicate that the *perceived humanness*, *eeriness*, and *attractiveness* indices can measure distinctly their corresponding concepts.

Table 6
Structural coefficients for the semantic differential items.

Items[a]	Perceived Humanness	Eeriness		Attractiveness
		Eerie	Spine-tingling	
Artificial–Natural	.89	–	–	–
Synthetic–Real	.87	–	–	–
Inanimate–Living	.86	–	–	–
Human-made–Humanlike	.84	–	–	–
Mechanical Movement–Biological Movement	.83	–	–	–
Without Definite Lifespan–Mortal	.80	–	–	–
Reassuring–Eerie	–	.79	–	–
Numbing–Freaky	–	.69	–	–
Ordinary–Supernatural	–	.55	–	–
Uninspiring–Spine-tingling	–	–	.75	–
Boring–Shocking	–	–	.75	–
Predictable–Thrilling	–	–	.66	–
Bland–Uncanny	–	–	.63	–
Unemotional–Hair-raising	–	–	.63	–
Unattractive–Attractive	–	–	–	.87
Ugly–Beautiful	–	–	–	.87
Repulsive–Agreeable	–	–	–	.78
Crude–Stylish	–	–	–	.75
Messy–Sleek	–	–	–	.69
Cronbach's α	.92	.71	.81	.90
Model	χ^2	df	GFI	AGFI
	1229.29	146	.91	.88
	NFI	CFI	RMR	RMSEA
	.97	.97	.23	.075

[a] Items sorted by the factor loading of each index.

Table 7
Correlation between the attractiveness, eeriness, humanness, and warmth indices in the final version.

	Attractiveness	Eeriness	Humanness	Warmth
Attractiveness	–			
Eeriness	−.03	–		
Humanness	.61***	.02	–	
Warmth	.62***	−.05	.66***	–

*** p < .001 (2-tailed).

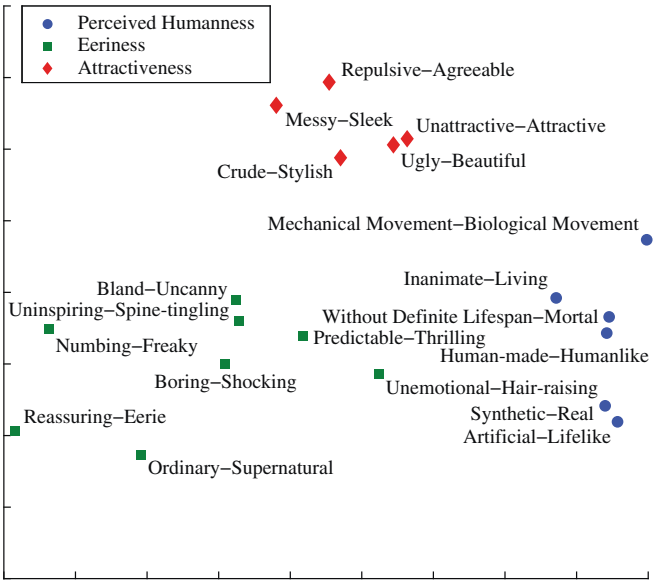

Fig. 8. Multidimensional scaling of the 19 semantic differential items was performed based on participant ratings of the figures in the 10 video clips. Items from the *perceived humanness*, *eeriness*, and *attractiveness* indices are widely separated.

The scatter plot shows that *humanness* and *eeriness* were decorrelated (Fig. 9), and *warmth* and *eeriness* were also decorrelated (Fig. 10). The data points did not follow a diagonal line as they had in the Godspeed indices. The insignificant correlation of the *eeriness* and *humanness* indices revealed that the final version of these indices could have good discriminant validity and high reliability. The *eeriness* index also had an insignificant correlation with the *warmth* index (r = −.05, p = .083). Although the attractiveness index yielded a high correlation with the *humanness* index (r = .61, p = .000), the data points vertically aligned into two main groups. Specifically this analysis showed that the *attractiveness* and *humanness* indices were somewhat less affected by positive (vs. negative) affect than *anthropomorphism* in the Godspeed indices. Although the *humanness* index was not correlated with the *eeriness* index after two rounds of testing, the *humanness* index maintained a high correlation with the *warmth* index (r = .66, p = .000). This analysis indicated that the notion of *warmth* might strongly overlap with the concept of *humanness* in practical circumstances. It is difficult to obtain discriminant validity; however, this may be improved in future studies.

4. Discussion

In Phase 1 of this study, the results of the validity analysis identified several problems with the Godspeed indices. The reliability

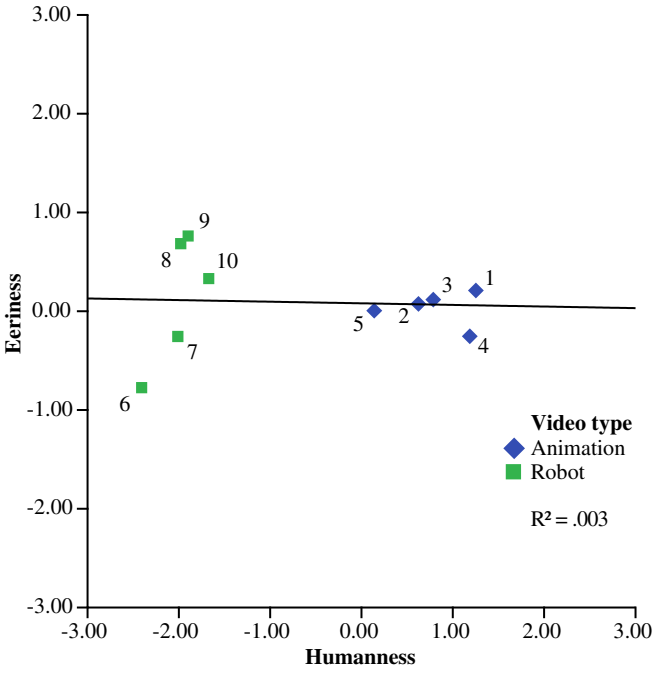

Fig. 9. The developed *humanness* and *eeriness* indices are not significantly correlated ($p = .514$, $r = .02$).

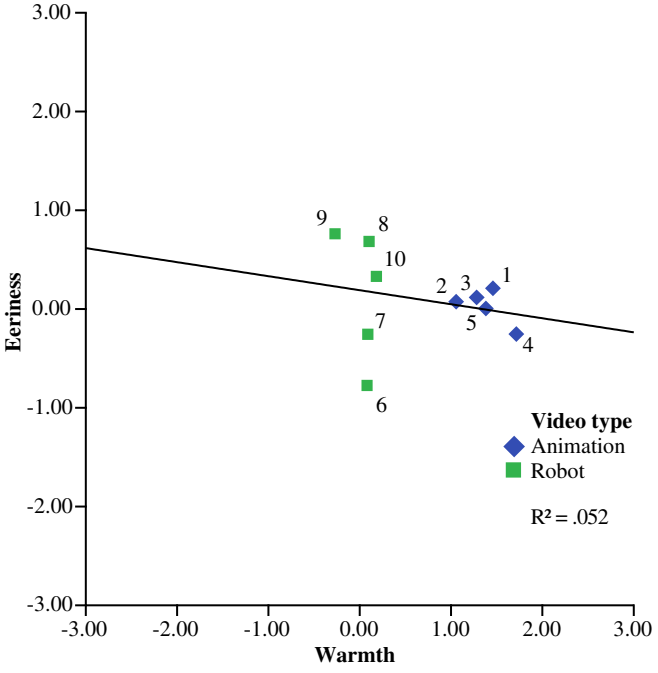

Fig. 10. The developed *warmth* and *eeriness* indices are not significantly correlated ($p = .083$, $r = -.05$).

of *perceived safety* was below the standard .70 cutoff. Confirmatory factor analysis also found inconsistencies in these indices and indicated that several items should be removed. However, the most serious problem was that *anthropomorphism*, *animacy*, *likeability*, and *perceived intelligence* were highly correlated with each other. This correlation indicates that they may be measuring the same concept, not separate concepts. These findings indicate the Godspeed indices are not appropriate as distinct concepts for evaluating anthropomorphic agents.

Therefore, Phase 2 included a new set of uncanny valley indices. After two rounds of testing, the developed indices for anthropomorphic characters' *attractiveness*, *eeriness*, and *humanness* were shown to have high internal reliability. With respect to computer-animated human characters and robots, these indices demonstrate the bipolarity of the semantic space for assessing people's emotional responses and judgments of personality traits (Bentler, 1969; Gärling, 1976; Lorr & Wunderlich, 1988; Rosenberg et al., 1968; Van Schuur & Kiers, 1994). Exploratory factor analysis was used to determine which items were retained for each index, and confirmatory factor analysis was used to verify the theoretical structure of the indices. Exploratory factor analysis demonstrated a comprehensive strategy for model selection prior to the validation by confirmatory factor analysis (Gerbing & Hamilton, 1996). In general, these indices appear to be valid for measuring their putative concepts.

4.1. Limitations and future work

The new indices were developed and validated with a particular set of stimuli, but it is important to retest them with other sets of stimuli. A limitation of the current set is that there were more nonhuman characteristics in the humanoid robots than in the animated human characters. To increase the variation within each group, less polished animations should be included, such as those rendered by video game software engines, and more polished human-looking robots should also be included, such as the Geminoid F developed by Hiroshi Ishiguro's laboratory at Osaka University and Kokoro Co. Ltd.

There is also considerable individual variation in emotional responses to humanoid robots and animated human characters. For example, although some participants were disturbed by the digital resurrection of the businessman Orville Redenbacher, other participants accepted the character as the real person. It is important to explore demographic factors that may influence the intensity of emotional responses. Although our study did not find age and gender to be significant factors in our population of undergraduates, these participant variables may be significant in a more heterogeneous sample that includes a broader range of ages. Past research has indicated that differences of culture and levels of exposure to robots can have a significant influence on attitudes (MacDorman, Vasudevan, & Ho, 2009). It is important to test the indices with different populations.

It is also important to apply external criteria to assess the validity of the developed indices. For example, the microdynamics of interaction between an embodied agent and a human being can indicate the extent to which the human being is responding to the agent as if it were human (Cassell & Tartaro, 2007). The same information can also indicate an aversive response when the interaction breaks down. Nonverbal behavior, such as gaze frequency and duration, have been used to determine preference between still and computer-animated monkeys in experiments on the uncanny valley that used macaque monkeys as subjects (Steckenfinger & Ghazanfar, 2009), and similar methods have also been applied to human infants and adults in the study of attractiveness. Facial expressions, which convey emotional state, can be measured by optical motion tracking or electromyography. These kinds of behavioral metrics can be used to test the predictive validity of the developed indices, as can physiological variables, such as heart rate, respiration, and galvanic skin response, which can increase in response to fear, an emotion associated with uncanny stimuli (Ho et al., 2008). Functional magnetic resonance imaging (fMRI) can be used to correlate response strength on the indices with brain areas that have been identified with emotions associated with the uncanny valley (e.g., fear and anxiety in the central and lateral amygdala

and medial hypothalamus, Panksepp, 2006; disgust in the anterior insular cortex and frontal operculum; Jabbi, Bastiaansen, & Keysers, 2008).

5. Conclusion

The set of uncanny valley indices developed in the current study are new measures for human perceptions of anthropomorphic characters that reliably assess four relatively independent individual attitudes. Bartneck, Kulić, et al. (2009) note that developing indices for robots can benefit robot developers. Comparing different robots and robot settings by means of the same index will help developers in making design decisions. The indices developed in this study have four advantages. First, they have excellent psychometric properties. The factor structure remains constant for both male and female participants and across two rounds of testing. Second, the internal reliability of the four indices is high. Third, the *eeriness* index, which could serve as the *y*-axis in Mori's graph, not only measures its named concept well but also is decorrelated from the *humanness*, *warmth*, and *attractiveness* indices. The apparent independence of the *humanness* and *eeriness* indices enables anthropomorphic characters to be plotted along nearly orthogonal axes, as implied by Mori's (1970) original graph of the uncanny valley. Confirmatory factor analysis was used to verify the theoretical structure of the indices. The results indicate the development of robust instruments for the dimensions of *attractiveness*, *eeriness*, *humanness*, and *warmth*. Fourth, the stimuli presented in this study were not limited to humanlike robots; they included computer-generated human characters. This widens the range of stimuli to which the indices may be applied.

Acknowledgments

The authors would like to express their gratitude to Himalaya Patel, Wade Mitchell, and the anonymous reviewers for their thoughtful suggestions for improving this paper. The IUPUI/Clarian Research Compliance Administration has approved this study (EX0903-35B). This study was supported by an IUPUI Signature Center grant.

References

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus other. *Journal of Personality and Social Psychology, 93*(5), 751–763.

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*(3), 259–290.

Bartneck, C., Kanda, T., Ishiguro, H., & Hagita, N. (2009). My robotic doppelganger: A critical look at the uncanny valley theory. In *Proceedings of the 18th IEEE international symposium on robot and human interactive communication* (pp. 269–276). Toyama, Japan.

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.

Bentler, P. M. (1969). Semantic space is (approximately) bipolar. *Journal of Psychology, 71*(1), 33–40.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 170*(2), 238–246.

Cassell, J., & Tartaro, A. (2007). Intersubjectivity in human–agent interaction. *Interaction Studies, 8*(3), 391–410.

Chin, W. W., & Todd, P. A. (1995). On the use, usefulness, and ease of use of structural equation modeling in MIS research: A note of caution. *MIS Quarterly, 19*(2), 237–246.

Comrey, A. L. (1978). Common methodological problems in factor analytic studies. *Journal of Consulting and Clinical Psychology, 46*(4), 648–659.

Cunningham, M. R. (1986). Measuring the physical in physical attractiveness: Quasi- experiments on the sociobiology of female facial beauty. *Journal of Personality and Social Psychology, 50*(5), 925–935.

Cunningham, M. R., Roberts, A. R., Barbee, A. P., Druen, P. B., & Wu, C.-H. (1995). "Their ideas of beauty are on the whole the same as ours?" Consistency and variability in cross-cultural perception of female physical attractiveness. *Journal of Personality and Social Psychology, 68*(2), 261–279.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from status and competition. *Journal of Personality and Social Psychology, 82*(6), 878–902.

Friborg, O., Martinussen, M., & Rosenvinge, J. H. (2006). Likert-based vs. semantic differential-based scorings of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience. *Personality and Individual Differences, 40*(5), 873–884.

Gärling, T. (1976). A multidimensional scaling and semantic differential technique study of the perception of environmental settings. *Scandinavian Journal of Psychology, 17*(1), 323–332.

Gefen, D., Straub, D., & Boudreau, M. (2000). Structural equation modeling and regression: Guidelines for research practice. *Communications of the Association for Information Systems, 4*(7), 1–79.

Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling, 3*(1), 62–72.

Ho, C.-C., MacDorman, K., & Pramono, Z. A. D. (2008). Human emotion and the uncanny valley: A GLM, MDS, and ISOMAP analysis of robot video ratings. In *Proceedings of the third ACM/IEEE international conference on human–robot interaction* (pp. 169–176). March 11–14, Amsterdam, The Netherlands.

Jabbi, M., Bastiaansen, J., & Keysers, C. (2008). A common anterior insula representation of disgust observation, experience and imagination shows divergent functional connectivity pathways. *PLoS ONE, 3*(8), e2939.

Jasienska, G., Ziomkiewicz, A., Ellison, P., Lipson, S., & Thune, I. (2004). Large breasts and narrow waists indicate high reproductive potential in women. *Proceedings of the Royal Society of London: Biological Sciences, 271*(1545), 1213–1217.

Jones, D. (1995). Sexual selection, physical attractiveness, and facial neoteny: Cross-cultural evidence and implications. *Current Anthropology, 36*(5), 723–748.

Jones, B. C., Little, A. C., & Perrett, D. I. (2004). When facial attractiveness is only skin deep. *Perception, 33*(5), 569–576.

Kahn, P. H., Jr., Ishiguro, H., Friedman, B., Kanda, T., Freier, N. G., Severson, R. L., et al. (2007). What is a human? Toward psychological benchmarks in the field of human–robot interaction. *Interaction Studies, 8*(3), 363–390.

Kulić, D., & Croft, E. (2007). Physiological and subjective responses to articulated robot motion. *Robotica, 25*, 13–27.

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin, 126*(3), 390–423.

Langlois, J. H., Roggman, L. A., Casey, R. J., Ritter, J. M., Rieser-Danner, L. A., & Jenkins, V. Y. (1987). Infant preferences for attractive faces: Rudiments of a stereotype. *Developmental Psychology, 23*(3), 363–369.

Lee, K. M., Park, N., & Song, H. (2005). Can a robot be perceived as a developing creature? *Human Communication Research, 31*(4), 538–563.

Lorr, M., & Wunderlich, R. A. (1988). A semantic differential mood scale. *Journal of Clinical Psychology, 44*(1), 33–36.

MacDorman, K. F. & Cowley, S. J. (2006). Long-term relationships as a benchmark for robot personhood. In *Proceedings of the 15th IEEE international symposium on robot and human interactive communication* (pp. 378–383). September 6–9, Hatfield, UK.

MacDorman, K. F., Coram, J. A., Ho, C.-C., & Patel, H. (2010). Gender differences in the impact of presentational factors in human character animation on decisions of ethical consequence. *Presence: Teleoperators and Virtual Environments, 19*(3).

MacDorman, K. F., Green, R. D., Ho, C.-C., & Koch, C. (2009). Too real for comfort: Uncanny responses to computer generated faces. *Computers in Human Behavior, 25*(3), 695–710.

MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in social and cognitive science research. *Interaction Studies, 7*(3), 297–337.

MacDorman, K. F., & Kahn, P. H. Jr., (2007). Introduction to the special issue on psychological benchmarks of human–robot interaction. *Interaction Studies, 8*(3), 359–362.

MacDorman, K. F., Ough, S., & Ho, C.-C. (2007). Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison. *Journal of New Music Research, 36*(4), 283–301.

MacDorman, K. F., Vasudevan, S. K., & Ho, C.-C. (2009). Does Japan really have robot mania? Comparing attitudes by implicit and explicit measures. *AI & Society, 23*(4), 485–510.

Manning, J. T., Scutt, D., & Lewis-Jones, D. I. (1998). Developmental stability, ejaculate size and sperm quality in men. *Evolution and Human Behavior, 19*(5), 273–282.

Mehrabian, A., & Russell, J. (1974). *An approach to environmental psychology*. Cambridge, MA: MIT Press.

Monahan, J. L. (1998). I don't know you but I like you: The effects of nonconscious affect on person perception. *Human Communication Research, 24*, 480–500.

Mori, M. (1970). *Bukimi no tani* (the uncanny valley). *Energy, 7*(4), 33–35.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Panksepp, J. (2006). Emotional endophenotypes in evolutionary psychiatry. *Progress in Neuro-Psychopharmacology and Biological Psychiatry, 30*(5), 774–784.

Powers, A. & Kiesler, S. (2006). The advisor robot: Tracing people's mental model from a robot's physical attributes. In *Proceedings of the first ACM SIGCHI/SIGART conference on human–robot interaction* (pp. 218–225). March 2–3, Salt Lake City, Utah, USA.

Rosenberg, S., Nelson, C., & Vivekananthan, P. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology, 9*(4), 283–294.

Seyama, J., & Nagayama, R. S. (2007). The uncanny valley: The effect of realism on the impression of artificial human faces. *Presence: Teleoperators and Virtual Environments, 16*(4), 337–351.

Sproull, L., Subramani, M., Kiesler, S., Walker, J. H., & Waters, K. (1996). When the interface is a face. *Human–Computer Interaction, 11*(2), 97–124.

Steckenfinger, S. A., & Ghazanfar, A. A. (2009). Monkey visual behavior falls into the uncanny valley. *Proceedings of the National Academy of Sciences, 106*(43), 18362–18366.

Thornhill, R., & Gangestad, S. W. (1993). Human facial beauty: Averageness, symmetry, and parasite resistance. *Human Nature, 4*(3), 237–269.

Thornhill, R., Gangestad, S. W., & Comer, R. (1995). Human female orgasm and mate fluctuating asymmetry. *Animal Behaviour, 50*(6), 1601–1615.

Van Schuur, W. H., & Kiers, H. A. L. (1994). Why factor analysis often is the incorrect model for analyzing bipolar concepts and what model to use instead. *Applied Psychological Measurement, 18*(2), 97–110.

Warner, R. M., & Sugarman, D. B. (1996). Attributions of personality based on physical appearance, speech, and handwriting. *Journal of Personality and Social Psychology, 50*, 792–799.

Wojciszke, B., Abele, A. E., & Baryla, W. (2009). Two dimensions of interpersonal attitudes: Liking depends on communion, respect depends on agency. *European Journal of Social Psychology, 39*(6), 973–990.

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin, 24*(12), 1245–1257.