Proceedings of the 52nd Hawaii International Conference on System Sciences | 2019

# Perceived Intelligence and Perceived Anthropomorphism of Personal Intelligent Agents: Scale Development and Validation

Sara Moussawi
Carnegie Mellon University
smoussaw@andrew.cmu.edu

Marios Koufaris
Baruch College, CUNY
Marios.koufaris@baruch.cuny.edu

## Abstract

*Personal intelligent agents are systems that are autonomous, aware of their environment, continuously learning and adapting to change, able to interact using natural language and capable of completing tasks within a favorable timeframe in a proactive manner. Examples include Siri and Alexa. Several unique characteristics distinguish these agents from other traditional information systems. Of particular interest in this work are characteristics of intelligence and anthropomorphism. This paper describes the process of developing two new measures with satisfactory psychometric properties that can be adapted by researchers to assess the users' perceptions of intelligence and anthropomorphism of PIAs. The measures are validated using data collected from 232 experienced PIA users.*

## 1. Introduction

Intelligent and human-like software is becoming more widespread especially with the development and refinement of personal intelligent agents (PIAs) like Siri and Alexa that are available through various technological devices. Anecdotal evidence suggests that users are developing an emotional connection with their PIAs while relying on them to complete their everyday tasks more efficiently. At the same time, while user expectations of PIAs are rising, so are incidents of user disappointment. For example, users can get frustrated if the agent doesn't recognize their voice or isn't providing correct or reliable answers to queries [31].

While evidence is pointing to users perceiving these agents to be human-like and intelligent, there are no established scales that measure these perceptions *specifically* for PIAs that are embedded in various technological devices (e.g., smartphones, Amazon's Echo, etc.). Existing measures for perceived intelligence focus on user perceptions of robots, where intelligence is partly determined by perceptions of physical appearance and movement. Measures of perceived anthropomorphism also focus on users' perceptions of robots' human-likeness based on facial features and mental capacities (Appendix, Table 1). In the absence of reliable measures, research on PIAs will be restricted to theorizing about perceptions of intelligence and anthropomorphism. Hence, the development of these measures is both timely and necessary. In this paper, we describe the development of reliable and valid scales for assessing perceptions of intelligence and anthropomorphism in PIAs. Given that research on user interaction with personal intelligent agents is nascent, developing these scales will provide researchers with standardized scales for future work exploring PIAs and similar systems. PIAs' unique characteristics distinguish them from other traditional information systems and have design implications. As such, an investigation of these characteristics should be of practical and theoretical relevance.

### 1.1. PIAs

PIAs are personalized intelligent software systems that perform actions in the place of and at the request of humans. A PIA operates within a specific user's context and is capable of formulating precise queries when interacting with the user. Characteristics of personalization, autonomy, awareness of the environment, learning and adaptation to change, communication, and task completion and pro-activeness have been commonly associated with intelligent agents and/or PIAs in the literature (Table 1). Personalization relates to the PIA's ability to respond to the user's requests and provide information based on user-specific preferences and history [27]. With time, a PIA acquires information and becomes able to better predict the user's behavior as it learns from patterns, information, and errors. Autonomy refers to the PIA's ability to operate upon command and without

URI: https://hdl.handle.net/10125/59452
ISBN: 978-0-9981331-2-6

the user's continuous intervention in every step [38; 45; 40; 27]. The agent can perform tasks on behalf of the user in an independent manner without the user's constant interference. For instance, when asked about the nearest movie theaters, the PIA conducts a search for theaters in the city, gets the current location of the user, calculates the distance to each theater, compares the results, and reports the finalized list to the user. The PIA is typically able to detect conditions in its physical and virtual environments (tactile input from user, ambient sound, other applications, data, etc.). The PIA is also usually aware of its own limitations in relation to both the physical and virtual environments [38; 45]. Learning and adapting to change relates to the PIA's ability to adapt its behavior based on prior events and new circumstances, and thus exhibit the ability to learn from change and newly acquired information [38; 27]. This is possible due to the underlying technology infrastructure, i.e., the complex choice and process models and learning algorithms, that give PIAs the ability to leverage every piece of information they acquire about the user and her environment and to learn from it. Additionally, the PIA is able to interact via natural language processing and language production abilities, allowing it to communicate with the user [38; 45; 27]. It is able to understand the users' verbal, text-based, and in some cases gestural requests and can ask follow-up questions. Furthermore, the agent is able to produce language to communicate with the user in the form of voice and/ or text. Task completion and pro-activeness relates to the PIA's ability to complete tasks within a favorable and expected timeframe for the user and be able to find and process the necessary information for completing its tasks [38; 45]. PIAs need to be able to set and pursue tasks on their own in anticipation of future user needs in a pro-active manner and should be able to provide the user with a useful answer.

**Table 1. PIA's characteristics**

| Characteristic | Definition |
|---|---|
| Personalization – discussed in [27] | Ability to respond to the user's specific requests and provide information based on user-specific preferences and history. |
| Autonomy – discussed in [38; 45; 40; 27] | Ability to operate upon command and without the user's continuous intervention in every step. |
| Awareness of/ Reactivity to the Environment – discussed in [38; 45] | Ability to detect conditions in its physical environment (tactile input from user, ambient sound, etc.) as well as its virtual environment (other applications, data, etc.). |
| Learning and Adaptation to Change – discussed in [38; 27] | Ability to adapt its behavior based on prior events and new circumstances, and thus exhibit the ability to learn from change and newly acquired information. |
| Communication – discussed in [38; 45; 27] | Ability to interact via natural language processing and language production abilities, allow the PIA to communicate with the user. |
| Task Completion and Pro-activeness – discussed in [38; 45] | Ability to complete tasks within a favorable and expected timeframe for the user, and be able to find and process the necessary information for completing its tasks. |

## 1.2. Intelligence

Objective definitions of intelligence in the Artificial Intelligence (AI) literature share qualities like goal-achievement, problem-solving, speed, flexibility, learning, improvement, and environmental awareness [24]. The Turing test provides an operational definition of intelligence in computers where a computer is considered intelligent if it is able to deceive the human interrogator into thinking that it is not a computer but a human too [38]. Definitions of intelligence in the psychology literature focus on components such as knowledge, mental abilities, learning, understanding, and reasoning [24]. Based on the PIAs' characteristics (Table 1) and the relevant literature, we define *perceived intelligence* as the formed perceptions about the extent to which a PIA's behavior is efficient, useful, goal-directed, and autonomous with an effectual output and an ability to produce and process natural language. The most important dimensions of PIA intelligence are: autonomy, physical world awareness, virtual world awareness, pro-activeness, completion time, communication ability, logical reasoning, learning ability, and output quality.

A thorough exploration of prior measures for perceived intelligence revealed that in the human-robot interaction (HRI) literature, intelligence of the system (robot) is measured by asking users to rate their perceptions of the robot's competence (incompetent/ competent), knowledge (ignorant/ knowledgeable), responsibility (irresponsible/ responsible), intelligence (unintelligent/intelligent), and sensibleness (foolish/ sensible) [3, 4; 34; 42]. We believe that the use of a single scale item of "intelligence" in this scale does not provide enough depth in measuring this construct. A more comprehensive scale should capture a more multidimensional measure of perceived intelligence by evaluating capacities that enable the PIA to seem intelligent such as effectiveness, autonomy, communication, goal-directed and useful behavior.

### 1.3. Anthropomorphism

The use of human-like features in technological artifacts like robots and PIAs aims to improve the agent's ability to engage in meaningful social interactions. Such interactions require the employment of human-like qualities in form or behavior [14]. Anthropomorphism relates to the user's attribution of human capacities to a non-human agent. Objects are generally perceived to be human-like when they possess features or characteristics that reflect emotions, cognition, or intention. Any object might be perceived to be human-like including cars and Coca-Cola bottles [1]. More than one view exists for conceptualizing and measuring users' perception of the system's human-likeness. Kiesler's [20; 21; 22; 35] work explored robots' sociability, human-likeness and machine-likeness, while Bartneck et al.'s [2; 3; 4] investigated the robots' movement, artificialness, fakeness, and consciousness. Waytz et al.'s [43; 44] work focused on the centrality of mental capacities and argued that the presence of mental capacities is both a necessary and sufficient condition for humanness. Finally, Haslam's [18]

work highlighted that humanness can be understood as a set of features or attributes that are either uniquely or typically human. Traits rated high in human uniqueness include agreeableness, openness and civility (polite, broad-minded, analytical, talkative). Traits rated high on human nature include extraversion, emotionality, warmth, openness, and agency (curious, imaginative, friendly, emotional, passionate). In this paper, we adopt Haslam's [18] view. That is, a PIA can be perceived as fluent, respectful, or funny (uniquely human attributes), or as friendly, happy or caring (human nature features). For instance, some of the currently available PIAs communicate with the user by listening and talking back with varying intonations and pitches, to appear more human-like. A thorough exploration of prior measures for perceived anthropomorphism are presented in the Appendix, Table 1.

## 2. Instruments development process

This paper aimed to develop measures to assess users' perceptions of intelligence and anthropomorphism. While it is possible to objectively assess intelligence and possibly anthropomorphism, it is not feasible to explore users' perceptions through objective measures (strength of the algorithm, presence of a voice, etc.).

Guided by established scale development guidelines [12; 19; 41; 26], we used a systematic approach, involving different methods to develop, refine, and validate scales for perceived intelligence and perceived anthropomorphism. The process was a multi-step one, including an extensive search of the literature, discussions with academic experts and experienced PIA users, card sorting exercises, pre-testing with users, a pilot test, and a confirmatory test through a study [5; 9; 32; 19].

### 2.1. Conceptual development and items generation

The first step of the scale development process consists of precisely defining the construct's conceptual domain and theme (i.e., set of attributes and characteristics) [26]. The previous three sections of this paper explained the conceptual domain research and described and defined the fundamental attributes of the two constructs.

Subsequently, a thorough analysis of the characteristics and dimensions revealed that both constructs are reflective in nature. That is, the sub-dimensions can exist independently and are viewed as manifestations of the focal constructs.

**Table 2. Scale development process**

| | |
|---|---|
| **Step1** | Develop a conceptual definition for the constructs |
| **Step 2** | Generate items to represent the constructs |
| **Step 3** | Assess the semantic content of the items and scales refinement – Pre-test with 262 PIA users and in-depth interviews with 2 experienced PIA users and 4 domain experts |
| **Step 4** | Assess the content validity of the items and scales refinement- sorting with a total of 3 users |
| **Step 5** | Collect data for pilot test |
| **Step 6** | Scales purification and refinement |
| **Step 7** | Gather data from a new sample and reexamine scales properties |
| **Step 8** | Assess scales validity |

The next step in the process is selecting the set of items for each construct [26]. Psychometricians emphasize a careful selection of the initial set of items used for measurement from the domain-sampling model. This model assumes the presence of a domain of content corresponding to the latent variable that the researcher is interested in measuring. With reflective indicators, items are chosen randomly from the universe of items relating to the construct of interest [12]. In this case, proper selection of the items ensures content validity [11; 33; 7]. After conducting a comprehensive review of the literature

on PIAs in AI, IS, and HRI and exploring practitioner-oriented publications, we generated 15 initial candidate items for each of the two constructs. For perceived intelligence, we developed new items based on the definition of PI without adapting any items from prior scales. For perceived anthropomorphism, and in line with our definition and prior literature, we compiled a set of items from prior studies, namely Waytz et al. [44], Kiesler et al. [22] and Powers and Kiesler [35]. We did not adapt the complete scales from these studies since they were intended to measure a wide variety of constructs, such as machine- and human-likeness, sociability, and mental capacities, and they were developed in disparate contexts, such as robots or cars, that are different than ours. We aimed to generate a set of items that covered the entire scope for each construct, as we defined them.

## 2.2. Conceptual refinement and items modification

The next step consisted of assessing the semantic content of the initial 30 items. We conducted around 262 pre-test, pre-structured interviews with PIA users of different use experiences. 52% of interviewees were male and 47% were female. The interviewees used Apple's Siri (72%), Google Now (23%), Microsoft's Cortana (2%), and other PIAs (2.7%). Around 57% of the participants have been using their agent for more than a year, 26% between 5 and 12 months, 12% between one and four months and 5% used it for less than a month.

This group of users seemed appropriate to capture all the relevant aspects of perceptions of intelligence and anthropomorphism. As a result of the interviews, we reworded and removed a few items.

We then conducted in-depth interviews with two experienced PIA users and contacted four experts in the domain to assess and comment on the relevance of the items as well as their semantic content. The goal of these interviews was to let users and experts comment on the items' relevance to the definitions of perceived intelligence and perceived anthropomorphism. As a result of these discussions, many items were reworded and some items were removed for low relevance and lack of clarity.

Examples of items that were reworded or removed in step 3 include: "*the personal intelligent agent can store what it knows*", "*the personal intelligent agent is able to think, i.e. it can answer questions and draw conclusions*", "*the personal intelligent agent learns*", "*the personal intelligent agent is able to detect patterns*", "*the personal intelligent agent is able to produce the best rational outcome*", for perceived intelligence; and *"the personal intelligent agent's voice tone is humanlike", "the personal intelligent agent wants to assist me with my daily tasks", "the personal intelligent agent is compassionate", "the personal intelligent agent is capable of showing love",* for perceived anthropomorphism.

The items that best fit the definition of each construct were retained, yielding 9 items for perceived intelligence, and 14 items for perceived anthropomorphism.

For content validity, i.e., to assess how well the items tapped into the construct, we conducted two rounds of sorting with two experienced users (1 male and 1 female, average age = 27, average of 1.5 years of experience with PIAs) and one IS researcher. The raters were representative of the population of interest and were deemed to be a good choice to reliably distinguish between the aspects of the content domain without being overwhelmed by the items [26]. Judges were first presented with the definition of the construct, and then a randomly sorted list of items. The instructions explained the exercise and asked each judge to rank the items from most to least relevant to the definition, with an ambiguous / does not match definition option. After sorting the items, judges explained their rationale. All three judges considered that for the purpose of humanness a PIA does not need to show intelligence. Hence, we removed the items adapted from Waytz et al. [44]: "*The personal intelligent agent is smart*"; "*The personal intelligent agent could feel what is happening around it*"; "*The personal intelligent agent could anticipate future user needs*"; "*The personal intelligent agent could set and pursue tasks by itself*". Additionally, one judge found the 6$^{th}$ item in the perceived intelligence scale to be unclear. Hence, we changed the original item "*the personal intelligent agent can communicate successfully with the user*" into two items: one that captured the PIA's reception and comprehension of the user's commands ("*the personal intelligent can understand my commands*"), and one that captured the comprehensible communication with the user ("*the personal intelligent can communicate with me in an understandable manner*"). The result of this step was 10 items for perceived intelligence and 9 items for perceived anthropomorphism (Tables 3 and 4). The card-sorting exercise resulted in an overall hit ratio of 88%. A measurement of the overall frequency with which judges placed items within the intended theoretical construct is indicative of the reliability of the classification scheme [30].

| Table 3. Refined items for perceived anthropomorphism |
|---|
| PA1. The personal intelligent agent is able to speak like a human. |
| PA2. The personal intelligent agent can be happy. |
| PA3. The personal intelligent agent can feel love. |
| PA4. The personal intelligent agent can get upset at times. |
| PA5. The personal intelligent agent can get frustrated at times. |
| PA6. The personal intelligent agent can be friendly. |
| PA7. The personal intelligent agent can be respectful. |
| PA8. The personal intelligent agent can be funny. |
| PA9. The personal intelligent agent can be caring. |

An examination of the hits per construct showed that for perceived intelligence all placements were within the target construct, while for perceived anthropomorphism 34 out of 42 (81%) were within the target. These results indicate that the items represent reasonable measures for perceived intelligence and perceived anthropomorphism. Given the thoroughness of the previous two pre-test steps, and the expertise of and level of agreement between the judges, we considered a total of three judges to be acceptable for this step.

| Table 4. Refined items for perceived intelligence |
|---|
| PI1.The personal intelligent agent is able to operate without my intervention. |
| PI2. The personal intelligent agent is aware of the physical world (e.g., its user, its location, etc.). |
| PI3. The personal intelligent agent is aware of the virtual world (e.g., other applications, the Internet, data, etc.). |
| PI4. The personal intelligent agent is able to set and pursue tasks by itself in anticipation of future user needs. |
| PI5. The personal intelligent agent can complete tasks quickly. |
| PI6. The personal intelligent agent can understand my commands. |
| PI7. The personal intelligent agent can communicate with me in an understandable manner. |
| PI8. The personal intelligent agent can find and process the necessary information for completing the tasks. |
| PI9. The personal intelligent agent can adapt its behavior based on prior events. |
| PI10. The personal intelligent agent is able to provide me with a useful answer. |

We retained 10 items for perceived intelligence and 9 for perceived anthropomorphism anticipating that around half of these items will be retained for use in the final scales [19]. A scale of four to five items is an effective means to reduce response bias caused by boredom and fatigue while maintaining internal consistency reliability [19].

## 2.3. Pilot study

In this stage of scale development, the items were presented to a sample that was representative of the actual population of users. The goal was to test the nomological network and to examine the psychometric properties of the new measures [19]. We used the items that survived content validity assessments to measure the constructs under examination.

This pilot study consisted of a cross-sectional survey distributed to experienced PIA users. We filtered participants by asking them about their prior use of PIAs. Participants who reported using a PIA at least twice in the last month were redirected to our questionnaire.

Participants were undergraduate college students, recruited from a subject pool, at a Northeastern university in the U.S. 29 responses were dropped due to missing data leaving 249 complete and valid data records for analysis. This sample size is sufficient for exploratory factor analysis purposes with a 1:13, item-to-response ratio [39; 19; 26].

Around 76% of the subjects were Apple's Siri users, 21% were Google Now users, 1.2% were Microsoft's Cortana users, and 2% were Amazon's Echo users. 63% of users were using their agent for more than a year. About 62% of the subjects were 18 to 20 years old and 22% were 21 to 23 years old. Around 46% of participants were female, 52% male, and 2 preferred not to specify their gender.

A preliminary data check for data quality showed that all scores for skewness and kurtosis were within the -2 and +2 range (skewness between -1.199 and 0.519 and kurtosis between -1.168 and 1.773), suggesting no serious deviations from the normality assumption. We also screened for outliers using Cook's D values.

Using SPSS, we then conducted an exploratory factor analysis using principal components analysis with oblique rotation. Our goal with the PCA was to remove items that did not load on the appropriate construct [10; 28]. We conducted a PCA using an oblique rotation because we assumed that perceived intelligence and perceived anthropomorphism items are correlated based on prior literature [16; 13; 28]. We went through a series of iterations to eliminate items with low loadings on all factors or high loadings on more than one factor. We retained items that clearly loaded on a single appropriate factor [19]. This process resulted in 5 items for perceived intelligence and 6 items for perceived anthropomorphism (Tables 5 and 6).

We kept indicators PI5 (loading = 0.79), PI6 (loading = 0.77), PI7 (loading = 0.72), PI8 (loading = 0.77), and PI10 (loading = 0.79) for perceived intelligence that captured the effectiveness, communication ability, environment awareness, autonomy, pro-activeness, and output speed and

correctness. We kept indicators PA1 (loading = 0.54), PA2 (loading = 0.79), PA6 (loading = 0.81), PA7 (loading = 0.66), PA8 (loading = 0.67), and PA9 (loading = 0.79) for perceived anthropomorphism that captured typically and uniquely human features. PA1 loaded high on perceived intelligence as well (0.50). Since it is theoretically essential to keep PA1, we keep it in this round.

**Table 5. Final instrument for perceived anthropomorphism**

| |
|---|
| **Pant1***. The personal intelligent agent is able to speak like a human (PA1) |
| **Pant2**. The personal intelligent agent can be happy (PA2) |
| **Pant3**. The personal intelligent agent can be friendly (PA6) |
| **Pant4**.The personal intelligent agent can be respectful (PA7) |
| **Pant5**.The personal intelligent agent can be funny (PA8) |
| **Pant6**.The personal intelligent agent can be caring (PA9) |
| * We provide the old codes for each item (starting with the acronym PA) as well as the new ones (in bold) starting with acronym Pant. |

Next, we assessed the reliability of the measuring instrument. Cronbach's alpha is a commonly used and accepted measure for reliability. Cronbach's alpha values for perceived intelligence and perceived anthropomorphism were 0.83 and 0.81 respectively. Both coefficients are above 0.70 indicating strong item covariance and that the sampling domain has been adequately captured [10; 33].

Following that, we assessed the convergent validity of the items. AVE values for perceived intelligence and perceived anthropomorphism were respectively 0.60 and 0.51, both above 0.50 [26], indicating that the majority of the variance in the items is shared with the latent construct. Data collected during the pilot stage was not used in subsequent stages.

**Table 6. Final instrument for perceived intelligence**

| |
|---|
| **Pint1***. The personal intelligent agent can complete tasks quickly (PI5) |
| **Pint2**. The personal intelligent agent can understand my commands (PI6) |
| **Pint3**. The personal intelligent agent can communicate with me in an understandable manner (PI7) |
| **Pint4**. The personal intelligent agent can find and process the necessary information for completing the tasks (PI8) |
| **Pint5**. The personal intelligent agent is able to provide me with a useful answer (PI10) |
| * We provide the old codes for each item (starting with the acronym PI) as well as the new ones (in bold) starting with the acronym Pint. |

# 3. Test of the nomological validity

## 3.1. Data collection and measures

In order to test the nomological validity (predictive ability) of the two scales, we ran the same study as in the pilot but with a new set of subjects. We tested the nomological validity of PI and PA in the context of post-adoption continuance of use by existing PIA users. Our nomological network was based on the unified model of IT continuance [6]. This model proposes that post-adoption users assess their pre-adoption expectations against perceived performance. Expectations are either confirmed (resulting in confirmation) or not (resulting in disconfirmation).

We expect that with PIAs, perceptions of intelligence and anthropomorphism shape disconfirmation of expectations, i.e. perceptions and expectations of performance. Additionally, we anticipate that the cognitive nature of perceptions of intelligence will impact perceptions of usefulness as it relates to how effectively the agent is.

The disconfirmation of expectations, satisfaction with use, continuance intention, subjective norms, and perceived usefulness scales were adapted from Bhattacherjee and Lin [6].

## 3.2. Sample

Participants were undergraduate college students at a Northeastern university in the U.S. A total of 252 subjects were recruited from a subject pool to participate in this study over a period of one month. 20 data records were marked and excluded from the data analysis resulting in a total of 232 complete and valid data records for data analysis. Participation was voluntary and students received course credit upon completion of the questionnaire. The course credit awarded was constant among all subjects and was not subject to performance or other factors. About 51% of the subjects were 18 to 20 years old, 27% were 21 to 23 years old, around 43% were male and 57% were female. Around 78% of the participants were Apple's Siri users, 16% were Google Assistant users, 4% were Microsoft's Cortana users and 2% were Amazon's Echo users.

## 3.3. Measurement model

Confirmatory factor analysis of the measurement model confirmed the good psychometric properties of our scales [32]. The outer loadings for the constructs were all statistically significant and above 0.6.

We then assessed the internal consistency and convergent and discriminant validity of the constructs within the model with SmartPLS [36]. The results are presented in Table 7. The composite reliability values

(above 0.84) demonstrated high reliability for all constructs (Table 7).

Convergent validity indicates that the items reflect the same latent variable. We evaluated convergent validity using composite reliability scores (all above 0.70) [15]. Convergent validity can be further assessed through the square root of the AVEs for all constructs, which were above the minimum level of 0.70 (Table 7).

To ensure discriminant validity, we used the Fornell-Larcker criterion [15; 17]. The square roots of the AVEs for the constructs were all larger than their correlations with other variables, hence ensuring satisfactory discriminant validity [17].

**Table 7. Cronbach's alpha, composite reliability, latent variable correlations, and square root of the AVE**

| | CA | CR | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|---|---|
| (1) | 0.84 | 0.90 | 0.87 | | | | | | |
| (2) | 0.85 | 0.90 | 0.63 | 0.83 | | | | | |
| (3) | 0.78 | 0.84 | 0.27 | 0.37 | 0.68 | | | | |
| (4) | 0.80 | 0.86 | 0.48 | 0.53 | 0.36 | 0.75 | | | |
| (5) | 0.86 | 0.90 | 0.67 | 0.78 | 0.28 | 0.50 | 0.80 | | |
| (6) | 0.88 | 0.92 | 0.69 | 0.73 | 0.38 | 0.57 | 0.70 | 0.86 | |
| (7) | 0.84 | 0.90 | 0.26 | 0.41 | 0.34 | 0.14 | 0.44 | 0.33 | 0.87 |

(1) Continuance intention; (2) Disconfirmation of expectations; (3) Perceived anthropomorphism; (4) Perceived intelligence; (5) Perceived usefulness; (6) Satisfaction; (7) Subjective norms.
CA: Cronbach's Alpha, CR: Composite Reliability, AVE: Average Variance Extracted
The values in bold in the diagonal cells in the correlations part of the table are the square root of the AVE for the corresponding constructs.

### 3.4. Structural model and nomological validity

We used a bootstrapping procedure to assess the significance of the path coefficients and predictive power of the model. The results of the component analysis indicated that 62% of the variance in perceived usefulness is explained by the perceptions of intelligence and disconfirmation of expectations, and 31% of the variance in disconfirmation of expectations is explained by the perceptions of intelligence and anthropomorphism.

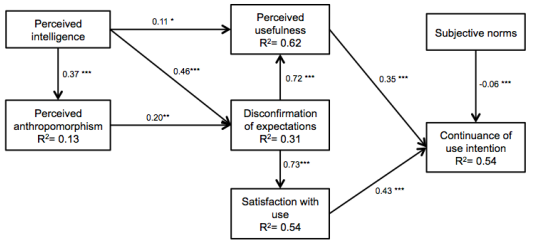


**Figure 1. Results of research model test (*** p<0.001, ** p<0.01, * p<0.05)**
Controlled for are habit, frequency of use, and tenure (all results were non-significant)

We present the results in Figure 1. Based on these results, perceptions of intelligence and anthropomorphism both related to performance of the PIA changed after continuing interaction with the PIA, which resulted in a disconfirmation of expectations. This is explained in the model through the significant relationships between perceived intelligence and perceived anthropomorphism with disconfirmation of expectations. Additional tests also showed that the effect of perceived intelligence on perceived usefulness is partially mediated by disconfirmation of expectations.

## 4. Discussion

This study described the process of creating reliable and valid scales for perceived intelligence and perceived anthropomorphism. We developed the measurement instruments following a systematic approach and conducted thorough validity and reliability tests of both measures aiming to provide dependable measures to be used in future research.

Following a thorough review of the relevant literature, we explained that the one existing measure for perceived intelligence and the measures for perceived anthropomorphism are not effective in a PIA context mainly because they are either too broad or better suited to a different context (e.g., interaction with robots). Hence, prior measures do not capture the core characteristics that make PIAs and similar systems unique.

Guided by the definition of perceived intelligence, we developed and refined a scale to measure the construct. The five final scale items reflect capacities of efficiency, autonomy, production and processing of natural language, useful and goal-directed behavior.

For perceived anthropomorphism, our definition and scale were guided by Haslam's view on humanness. We distinguished between PIA's features that are typically vs. uniquely human. The six final scale items reflect attributes of human uniqueness, such as civility, openness, and agreeableness (e.g. fluent, respectful, funny) and attributes of human nature, including friendliness, emotion, and passion (e.g. friendly, happy, caring). The first item in the perceived anthropomorphism scale ("*the PIA is able to speak like a human"*) had high loadings on both perceived intelligence and perceived anthropomorphism in the pilot study. We kept the item in the scale based on theoretical support.

## 5. Contributions

We believe that the theoretical and practical contributions of this research are significant. On the theoretical level, we have developed and validated two new scales that reliably measure perceptions of intelligence and anthropomorphism of perceived intelligence agents by their users. These measures will be critical for researchers who examine the increasingly complex relationship between users and PIAs of all types. It is especially important that the measures we developed are for the perceptions of intelligence and anthropomorphism, because those two characteristics are key to differentiating PIAs (and other systems that employ some form of artificial intelligence) from many of the systems that IS researchers have examined to date, such as decision support systems, enterprise resource systems, and social media platforms. In order for any research in PIA use and development to be complete, both intelligence and anthropomorphism must be taken into account, in our opinion.

An important aspect of both measures that we developed in this paper is that they are system-agnostic. Both measure perceptions regarding the PIA's behavior, intention, and ability but are not concerned with issues of appearance, interface design, or physical manifestation. This means that both measures can continue to be used even as PIAs evolve into more complex and capable systems or as the way they are available to users might change, such as through technology implants or intelligent rooms or homes. This can be an important factor in achieving continuity in research in this area.

On the practical level, we believe that this research and the two measures we have developed can be used in the development of more useful and better accepted PIAs. If developers are able to reliably measure how intelligent and human-like users perceive PIAs to be, they can ensure that they develop systems that enjoy maximum user acceptance.

## 6. Conclusion

Given the pace at which users are exploring and adopting PIAs, it is vital to explore the unique characteristics of these systems. In this work, we focused on intelligence and anthropomorphism. We describe the process of systematically developing measurement instruments for user perceptions of intelligence and anthropomorphism. Future research can adopt these measures to assess users' perceptions when interacting with this new technology.

## 7. References


[1] Aggarwal, P., & McGill, A. L. (2007). Is that car smiling at me? Schema congruity as a basis for evaluating anthropomorphized products. *Journal of Consumer Research, 34*(4), 468-479.

[2] Bartneck, C., Kanda, T., Ishiguro, H., & Hagita, N. (2009). *My robotic doppelgänger-A critical look at the uncanny valley.* Paper presented at the RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication.

[3] 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.

[4] Bartneck, C., Verbunt, M., Mubin, O., & Al Mahmud, A. (2007). *To kill a mockingbird robot.* Paper presented at the Human-Robot Interaction (HRI), 2007 2nd ACM/IEEE International Conference on.

[5] Benlian, A., Koufaris, M., & Hess, T. (2011). Service quality in software-as-a-service: Developing the SaaS-Qual measure and examining its role in usage continuance. *Journal of Management Information Systems, 28*(3), 85-126.

[6] Bhattacherjee, A., & Lin, C.-P. (2015). A unified model of IT continuance: three complementary perspectives and crossover effects. *European Journal of Information Systems*.

[7] Bohrnstedt, G. W. (1970). Reliability and validity assessment in attitude measurement. *Attitude measurement*, 80-99.

[8] Chin, M. G., Yordon, R. E., Clark, B. R., Ballion, T., Dolezal, M. J., Shumaker, R., & Finkelstein, N. (2005). *Developing and anthropomorphic tendencies scale.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

[9] Christophersen, T., & Konradt, U. (2006). The development of a formative and a reflective scale for the assessment of on-line store usability. *Studies, 55*, 20.

[10] Churchill Jr, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of marketing research*, 64-73.

[11] Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319-340.

[12] DeVellis, R. F. (1991). Guidelines in scale development. *Scale Development: Theory and Applications. Newbury Park, Calif: Sage, 5191*.

[13] Dillon, W. R., & Goldstein, M. (1984). *Multivariate Analysis Methods and Applications*. New York: John Wiley & Sons.

[14] Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and autonomous systems, 42*(3), 177-190.

[15] Fornell, C., & Larcker, D. F. (1981). Structural equation models with unobservable variables and measurement error: Algebra and statistics. *Journal of marketing research*, 382-388.

[16] Hair, J. F. (2009). *Multivariate data analysis* (7 ed.): Pearson.

[17] Hair Jr, J. F., Hult, G. T. M., Ringle, C., & Sarstedt, M. (2016). *A primer on partial least squares structural equation modeling (PLS-SEM)* (Second ed.): Sage Publications.

[18] Haslam, N., Loughnan, S., Kashima, Y., & Bain, P. (2008). Attributing and denying humanness to others. *European review of social psychology, 19*(1), 55-85.

[19] Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational research methods, 1*(1), 104-121.

[20] Kiesler, S., & Goetz, J. (2002a). Machine trait scales for evaluating mechanistic mental models of robots and computer-based machines. *Unpublished manuscript, Carnegie Mellon University. Downloadable at* https://www.cs.cmu.edu/~kiesler/anthropomorphism-org/pdf/Machine_scale.pdf.

[21] Kiesler, S., & Goetz, J. (2002b). *Mental models of robotic assistants.* Paper presented at the CHI'02 extended abstracts on Human Factors in Computing Systems.

[22] Kiesler, S., Powers, A., Fussell, S. R., & Torrey, C. (2008). Anthropomorphic interactions with a robot and robot-like agent. *Social Cognition, 26*(2), 169-181.

[23] Komiak, S. Y., & Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS quarterly*, 941-960.

[24] Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and applications, 157*, 17.

[25] MacDorman, K. F. (2006). *Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the uncanny valley.* Paper presented at the ICCS/CogSci-2006 long symposium: Toward social mechanisms of android science.

[26] MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS quarterly, 35*(2), 293-334.

[27] March, S., Hevner, A., & Ram, S. (2000). Research commentary: an agenda for information technology research in heterogeneous and distributed environments. *Information Systems Research, 11*(4), 327-341.

[28] McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research, 13*(3), 334-359.

[29] Minato, T., Shimada, M., Itakura, S., Lee, K., & Ishiguro, H. (2005). *Does gaze reveal the human likeness of an android?* Paper presented at the Development and Learning, 2005. Proceedings. The 4th International Conference on.

[30] Moore, G. C., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research, 2*(3), 192-222.

[31] Moussawi, S. (2018). *User Experiences with Personal Intelligent Agents: A Sensory, Physical, Functional and Cognitive Affordances View*. Paper presented at the Proceedings of the 2018 ACM SIGMIS Conference on Computers and People Research, Buffalo-Niagara Falls, NY, USA.

[32] Nevo, S., & Wade, M. R. (2008). *Development and validation of scales to measure the strategic potential of IT-enabled resources: A resource-based approach.* Paper presented at the Hawaii International Conference on System Sciences, Proceedings of the 41st Annual.

[33] Nunnally, J. (1978). Psychometric methods: New York: McGraw-Hill.

[34] Parise, S., Kiesler, S., Sproull, L., & Waters, K. (1996). *My partner is a real dog: cooperation with social agents.* Paper presented at the Proceedings of the 1996 ACM conference on Computer supported cooperative work.

[35] Powers, A., & Kiesler, S. (2006). *The advisor robot: tracing people's mental model from a robot's physical attributes.* Paper presented at the Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction.

[36] Ringle, C. M., Wende, S., & Becker, J. M. (2015). SmartPLS 3.

[37] Ruijten, P. A., Bouten, D. H., Rouschop, D. C., Ham, J., & Midden, C. J. (2014). *Introducing a rasch-type anthropomorphism scale.* Paper presented at the Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction.

[38] Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A modern approach* (Third ed. Vol. 25). New Jersey: Pearson Education.

[39] Schwab, D. P. (1980). Construct validity in organizational behavior. *Research in organizational behavior, 2*(1), 3-43.

[40] Shoham, Y. (1993). Agent-oriented programming. *Artificial intelligence, 60*(1), 51-92.

[41] Straub, D. W. (1989). Validating instruments in MIS research. *MIS quarterly*, 147-169.

[42] Warner, R. M., & Sugarman, D. B. (1986). Attributions of personality based on physical appearance, speech, and handwriting. *Journal of personality and social psychology, 50*(4), 792.

[43] Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5*(3), 219-232.

[44] Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52*, 113-117.

[45] Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The knowledge engineering review, 10*(02), 115-152.

# 8. Appendix

| Table 1. Anthropomorphism Measures | | |
|---|---|---|
| **Source** | **Measure** | **Fit within PIA context** |
| MacDorman [25] | A 9-point mechanical versus humanlike scale for assessing the human likeness of what is being viewed. | One-item scale / Does not capture the concept well |
| Chin et al. [8] | A 208-item scale developed to measure self-reported anthropomorphic tendencies. | 208 items/ Not an efficient measure for this study |
| Minato et al. [29] | Behavioral measurement that analyzed differences in participants' observed responses when looking at a robot. | Behavioral measurement/ outside the methodological scope for this study |
| Aggarwal and McGill [1] | Two questions asking participants whether the car resembles a person. | Two questions designed to assess whether participants view the car as a person/ Not an efficient measure for this study |
| Kiesler and Goetz [21] | For experiment 1: adapted existing scales of sociability, intellect, and personality; and created a measure of mechanistic models using the ratings: complex, obsolete, intuitive, works quickly, usable, durable, powerful, reliable, accurate.<br>For experiment 2: created a measure adapting the Big Five and and also used the measure of mechanistic mental models. | Not efficient measures for this study. |
| Kiesler and Goetz [20] | Five 5-point scales (efficiency, maintenance, durability, safety, and information technology) used to measure users' mechanistic mental models of robots and computers. | Not an efficient measure for this study/ Scales are more appropriate with robots and PCs. |
| Powers and Kiesler [35] | Items adapted from previous research. Used six different scales to measure sociability (cheerfulness, friendliness, warmth, happiness, likable, sympathy, compassionate, gentle, tender, emotion, attractiveness), knowledge (competence, knowledge, intelligence, expert, reliability, usefulness, trustworthiness, likable), dominance (strong personality, assertive, dominant, dominance, power), humanlikeness (natural, humanlike, like a human, lifelike, moves like a human, has a mind), masculinity, and machinelikeness (machinelike). | Not an efficient measure for this study / Measured humanelikeness, machinelikeness, and sociability as three different constructs |
| Bartneck et al. [2] | Transformed Powers and Kiesler's [35] scales into one 7-point semantic differential scale: fake/natural, machinelike/humanlike, unconscious/ conscious, artificial/ lifelike and moving rigidly/ moving elegantly. | Scale focuses on the robot's appearance and movement and is more appropriate with robots. |
| Kielser et al. [22] | Two 7-point scales to measure lifelikeness with 4 items (humanlike, lifelike, machinelike (rev.), natural) and traits with 6 sub-scales (dominant -4 items, trustworthy -5 items, sociable -10 items, responsive -6 items, competent -14 items, respectful -3 items). | Not an efficient measure for this study / Scale is more appropriate with robots. |
| Waytz et al. [43] | Individual differences in anthropomorphism questionnaire (IDAQ) – a 15-item scale that measures individual differences in anthropomorphism (attributions of intentions, consciousness, emotions) in adults. | Not an efficient measure for this study / Scale does not tap on the concept that we are trying to measure. |
| Waytz et al. [44] | A 10-point 4-item anthropomorphism scale that asked participants to rank how smart the car was, how well it could feel what was happening around it, how well it could anticipate what was about to happen, and how well it could plan a route. | Not an efficient measure for this study / Scale focuses solely on the mental functions of the car. |
| Bartneck et al. [3] | A 5-point 5-item anthropomorphism scale for users to report their impression of the robot: fake/natural, machinelike/humanlike, unconscious/conscious, artificial/lifelike, moving rigidly/ moving elegantly | Not an efficient measure for this study / Scale is more appropriate with robots. |
| Ruijten et al. [37] | A 25-item Rasch scale tested with robots based on Haslam's [18] notion of typically vs. uniquely human. | Not an efficient measure for this study / Psychometric scale. |
| This paper | A 7-point 6-item anthropomorphism scale based on Haslam's [18] notion of typically vs. uniquely human that asked the user to rate how well can the agent: speak like a human, be happy, friendly, respectful, funny, caring. | Relevant in a PIA context. |