The Artificial-Social-Agent Questionnaire: Establishing the long and short questionnaire versions

Siska Fitrianie
Delft University of Technology
Delft, The Netherlands
s.fitrianie@tudelft.nl

Merijn Bruijnes
Utrecht University
Utrecht, the Netherlands
m.bruijnes@uu.nl

Fengxiang Li
School of Business Administration
Northeastern University
Shenyang, China
1810438@stu.neu.edu.cn

Amal Abdulrahman
Delft University of Technology
Delft, the Netherlands
a.abdulrahman@tudelft.nl

Willem-Paul Brinkman
Delft University of Technology
Delft, the Netherlands
w.p.brinkman@tudelft.nl

ABSTRACT

We present the ASA Questionnaire, an instrument for evaluating human interaction with an artificial social agent (ASA), resulting from multi-year efforts involving more than 100 Intelligent Virtual Agent (IVA) researchers worldwide. It has 19 measurement constructs constituted by 90 items, which capture more than 80% of the constructs identified in empirical studies published in the IVA conference 2013-2018. This paper reports on construct validity analysis, specifically convergent and discriminant validity of initial 131 instrument items that involved 532 crowd-workers who were asked to rate human interaction with 14 different ASAs. The analysis included several factor analysis models and resulted in the selection of 90 items for inclusion in the long version of the ASA questionnaire. In addition, a representative item of each construct or dimension was selected to create a 24-item short version of the ASA questionnaire. Whereas the long version is suitable for a comprehensive evaluation of human-ASA interaction, the short version allows quick analysis and description of the interaction with the ASA. To support reporting ASA questionnaire results, we also put forward an ASA chart. The chart provides a quick overview of the agent profile.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Intelligent agents**;

KEYWORDS

Artificial social agent; user study; evaluation instrument; questionnaire; construct validity

IVA '22, September 6–9, 2022, Faro, Portugal

ACM ISBN 978-1-4503-9248-8/22/09.
https://doi.org/10.1145/3514197.3549612

ACM Reference Format:
Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. 2022. The Artificial-Social-Agent Questionnaire: Establishing the long and short questionnaire versions. In *ACM International Conference on Intelligent Virtual Agents (IVA '22), September 6–9, 2022, Faro, Portugal.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3514197.3549612

INTRODUCTION

In this paper, we present the *Artificial Social Agent Questionnaire*, an instrument for evaluating human interaction with an artificial social agent (ASA), resulting from multi-year efforts involving more than 100 Intelligent Virtual Agent (IVA) researchers worldwide. This ongoing collaboration[1] was motivated by discussions at the IVA community and, specifically, at IVA conference workshops on methodology in 2018 and 2019. It aims to develop a validated standardized questionnaire instrument for evaluating human interaction with ASAs. Different types of these ASAs exist, ranging from text-based chatbots to computer-controlled virtual humanoid agents to virtual and physical robots. As such, this effort is relevant for all sub-fields of the ASA community including IVA, Human-Robot Interaction (HRI), and more. With the variety of ASAs that exist, the questionnaire: (i) can make a standardized statement about the quality of the ASA; (ii) can make a statement about the various aspects and dimensions expected to be relevant to capturing an ASA's quality; and (iii) is grounded in examples of current and popular ASAs. Researchers can use this questionnaire as a common base to measure and describe the interaction experience with an ASA instead of following "the old-fashioned trend" of creating a new measurement instrument in every new study [7]. Using the ASA questionnaire, researchers can establish a 'broad-spectrum impression' of their ASA's quality that is comparable with other researchers' agents. This allows us to, together as a community, address the methodological challenge in our field, specifically, issues related to: comparing different agents, replicating scientific findings, validating claims, and establishing the impact of our ASAs.

Previous and Current Work

The previous work in this community project can be found in the Open Science Foundation's Work-group of Artificial Social

[1] Join our effort at OSF work-group of Artificial Social Agent Evaluation Instrument, https://osf.io/6duf7/.

Agent Evaluation Instrument[1]. We first established the focus of our efforts: the interaction between the user and ASA. As depicted in Figure 1, both parties (i.e., human user and ASA) communicate, take into account each other's contribution, and involve in a certain process to achieve a certain outcome. Thus, we exclude the pre-existing notions of the user and agent that can be established prior to interaction and also exclude the context-dependent (often task-related) processes and outcomes.

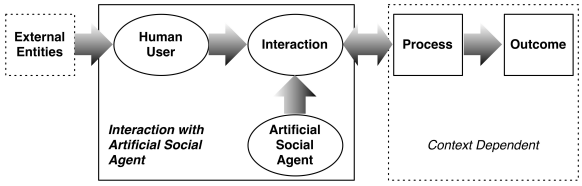

Figure 1: A world model of human-ASA interactions: the instrument measures only the 'Interaction with ASA'.

Within the interaction focus, previous efforts identified a unified set of 19 constructs and their dimensions that capture more than 80% of constructs used in empirical studies published at the IVA conference between 2013 to 2018 [8]. A construct is a concept or subject matter that one wishes to measure. Some constructs are multidimensional that can consist of two or more underlying dimensions, e.g., the Human-Like Appearance and Natural Behavior are dimensions of the construct Agent's Believability. For these constructs and their dimensions, the community composed 131 questionnaire items [6]. These items were established in three phases. Firstly, a group of eight experts (i.e., members of the work-group) collected the initial 431 potential construct items. Secondly, twenty experts rated whether items measure (only) their intended construct. After removing items that were found not appropriate for their intended construct or on the other hand, appropriate for multiple constructs, this step resulted in 207 content-validated questionnaire items. Thirdly, a reliability analysis was conducted, involving 192 crowd-workers who were asked to rate a video of a human interacting with an ASA using the content-validated items. The selected 131 out of 207 items (in 19 constructs) showed a respectable average level of reliability with Cronbach's α range [.60 .. .87].

In this paper, we present the results of the construct validity determination of the item set. Here we examine whether related items operate in a consistent manner (construct validity), specifically, if an item converges with the items of the same construct or dimension (convergent validity) and diverges from items of other constructs or dimensions (discriminate validity) [10]. For this examination, we performed factor analysis to explore the theoretical factor structure of the constructs and their dimensions by analysing their internal consistency and the underlying relationships between these constructs (and dimensions) and between items. The analysis was based on data obtained from crowd-workers (n = 532); each rated one of fourteen different ASAs. The result of the analysis allowed us to select items that both converge and discriminate for the long version of the questionnaire. Based on this final set, we selected representative items of constructs or dimensions for a short version allowing for a quick and simple evaluation of an ASA. The dimensions can be depicted on a web chart for the purpose of visualising the evaluation. This study was approved by the data management officer and the Human Research Ethics Committee TUDelft (no. 1402/18-12-2020).

METHOD

Participants

Determining sample size, by considering the rule-of-thumb for conducting factor analysis, is about 4 to 10 participants per observed variable [2], resulting in a minimum of 524 participants (i.e., 131 items * 4 participants). Furthermore, we conducted a simulation-run based on the theoretical model (of constructs) and found that, with 131 items in 19 constructs, the model could achieve convergence with a minimum of 406 participants. Finally, to ensure equal distribution of participants over each of the agents, we needed 532 participants (i.e., 14 agents * 38 participants). Fortunately, this fell within our available budget with some room to mitigate participant attrition (i.e., recruit replacements for excluded participants).

We recruited 567 crowd-workers from an online crowd-sourcing platform and 532 (95.8%) were included in the analysis based on two criteria: (1) having a compatible internet browser with the video format used in the study; and (2) correctly answering at least 12 out of 15 attention check questions. 33 participants failed the video check, and only 2 participants failed the attention check (i.e., the number of correctly answered-attention-questions: M = 14.83, SD = .91, range [1 .. 15]). Participants were paid for their time according to the crowd-platform's regulations.

Material

We decided to use videos of human-ASA interactions as stimuli material, which allowed us to collect data on interaction across a series of ASA. Essential for establishing the set of ASAs was that it would create an opportunity for the ratings across ASAs to be, to a degree, non-related on constructs and dimensions. We wanted to avoid a situation where, for example, very likable ASAs always had a human-like appearance, thereby providing no opportunity to examine discriminant validity between items of the constructs agent's likeability and human-like appearance. Therefore, prior to the current study, nine experts were involved in collecting 56 (30-second) video clips corresponding to 56 different agents. The agents vary in types (e.g., robots, chatbots, voice assistants, virtual agents, and real animals), domains (e.g., education, healthcare, personal assistant, and entertainment), environments (i.e., reality, mixed reality, virtual reality, and augmented reality), and development stages (i.e., high or low fidelity agents, partial or full functionality of the system). Some videos show conversations between agents and users, while other videos show how users use agents to help them with some tasks. The videos and related documents (including discussion notes) are online available[2].

As we mentioned before, we set out to select a set of agents that more-or-less cover the range of the constructs we intend to measure, as we need variability among observed items in the constructs, while at the same time allowing some degree of independence between the rating of items from different constructs and dimensions.

[2]https://osf.io/q2xur/wiki/home/.

Table 1: The stimuli used in the study: 13 ASAs and one animal (a dog).

Agent	Description	Modality	Comm. Language	Embodiment	Mobility	ASA-Score
iCAT	Cat-like robot developed by Philips	V, A, T	Spoken & body language	Physical	Stationary	-2
DeepBlue	Chess playing computer developed by IBM	V	Symbolic language	Disembodied	NA	7
Amy	Virtual healthcare agent [11]	V, A	Spoken & body language	Virtual	Virtual	9
Furby	Toy resembling a hamster or owl-like creature developed by Tiger Electronics	V, A, T	Spoken, body & non-language	Physical	Limited	13
Siri	Virtual assistant developed by Apple	A	Spoken language	Disembodied	NA	13
HAL 9000	Fictional character in A Space Odyssey	A	Spoken language	Disembodied	NA	14
Poppy	Virtual human from SEMAINE [12]	V, A	Spoken & body language	Virtual	Stationary	14
Sim Sensei	Virtual healthcare agent [3]	V, A	Spoken & body language	Virtual	Stationary	17
CHAPPiE	Robot character in CHAPPiE	V, A, T	Spoken & body language	Physical	Physical	18
Aibo	Robotic dog developed by Sony	V, A, T	Spoken, body, symbolic & non-language	Physical	Physical	20
Sarah	Customer service from Digital Humans	V, A	Spoken & body language	Virtual	Stationary	22
Nao	Humanoid robot from Aldebaran Robotics	V, A, T	Spoken & body language	Physical	Physical	23
Marcus	Cyborg character in Terminator	V, A, T	Spoken & body language	Physical	Physical	25
Dog	Domesticated carnivore, 'man's best friend'	V, A, T	Spoken, body & non-lang.	Physical	Physical	29

Note: Modality = communication modalities, i.e. V= Visual, A = Auditory, T = Tactile; Comm. Language = language used (by human and/or the agent) to communicate, i.e. spoken, body language (i.e. facial expression, head-, legs-, arms-, hands- or body motion), symbolic (e.g. buzzers, lights, cards), and non-language vocalization (e.g. vocal sounds without words, bark); NA = not applicable; ASA-Score = the (rounded up) total score on the ASA chart.

Therefore, we included agents that ranged relatively from the highest to the lowest score on the different constructs. To determine this, three experts predicted the ratings of each agent on each construct (high, medium, or low) and used the result to calculate the correlation between agents. We selected a set of agents (n = 14: 13 ASAs and one dog) that had the least correlation with each other and across the constructs to ensure diverse agent rates across the constructs. Table 1 shows the list of agents used in this study.

To gather the data, we used Qualtrics and the online crowdsourcing platform Prolific Academic. Further, data processing and analysis were conducted using R (v4.0.4) with factor analysis libraries from the package psych (v2.1.3) and lavaan (v0.6-8). Analyses scripts and data are online available [5].

The ASA questionnaire items have a 7-point scale on an interval between 'disagree' (value of -3) to 'agree' (value of 3) with the middle point (value of 0) for 'neither agree nor disagree'. They are formulated as singular statements in such a way they can easily be changed so that they can be answered by a person who interacted with an agent (i.e., first-person point of view) and by someone who observed interaction with an agent (i.e., third-person point of view). A previous study showed that scores differences between points of view were limited [6]. Therefore, large-scale testing through crowd-sourcing, using videos (third-person perspective), is feasible.

Design and Procedure

We asked participants to rate (a video of) interaction between a human user and an ASA using the questionnaire. The human-ASA interaction was displayed in a 30-second video clip, which was randomly selected from the fourteen selected agents. All participants received the same 131 questionnaire items plus 15 attention check questions in random order. Particularly, the items were set up to the third-person point of view, e.g., "The user likes <the agent>" instead of "I like <the agent>", where <the agent> was replaced with the name of the agent that the participant viewed.

All participants had to determine the compatibility of their internet browser with the video format. Only participants who met this criterion could continue with the study. Participants read the instructions [5] and gave their online informed consent. Then, after watching the video, they rated the questionnaire items based on what they have seen in the video. The first viewing of the video was enforced by preventing participants to advance to the next page for the duration of the video, and they could re-watch the video at any time during the remainder of the experiment. Participants had to answer all of the 131 items and fifteen attention check questions before they could submit their answers.

Data Preparation and Analysis Plan

A goal of the analysis was to: (i) lessen the number of items that correlate with other constructs/dimensions; and (ii) balance the coverage within each construct/dimension and the discriminatory power between constructs/dimensions. Prior to the analysis, the observed data was standardized to allow us to compare the ratings between different ASAs. For this, we calculated the mean and the standard deviation per item per ASA. Then, for each observed value of an item, we subtracted the mean and divided it by the standard deviation based on its corresponding ASA.

In the following sections, we describe how we determined the final item set and the short version of the ASA questionnaire, and we suggest how to present the insights in an ASA chart.

The Final Questionnaire Items. Running an admissible (i.e., no negative variances are found and no a non-positive definite matrix returned) second-order analysis based on one theoretically grounded conceptual model proved impossible due to the model complexity [1]. To solve this problem, we broke up the theoretical model into

smaller models. We first created models containing overlapping (with strong co-linearity) constructs/dimensions to create 'worst-case scenarios' (i.e., if a smaller model is able to resolve these closely related constructs, the full model is likely also able to resolve them). To do this, we took two sequential analysis (ovals in Figure 2):

(1) Convergent validity analysis: Analysing individual constructs' convergence. A Confirmatory Factor Analysis (CFA) was carried out for each construct in isolation to verify the internal consistency of the construct and, for those that have more than one dimension, to reduce the co-linearity between dimensions.
(2) Discriminant validity analysis: Analysing constructs in groups. The constructs (and their dimensions) were grouped into smaller models based on grouping established with an Exploratory Factor Analysis (EFA) on the construct and dimensions scores (i.e. the predicted latent scores derived from the CFA analysis of individual constructs in the convergent validity analysis). Then, these models were analysed with a CFA. Here, we removed items that had a low discriminatory power between constructs. The remaining items are the final item set of the questionnaire. However, note that discriminant validity is relevant from a statistical point of view, yet due to the expert perspective in our previous work, content validity took precedent. In other words, we optimized for discriminant validity but accepted correlations between constructs [14]: experts are interested in constructs that 'apparently' overlap.

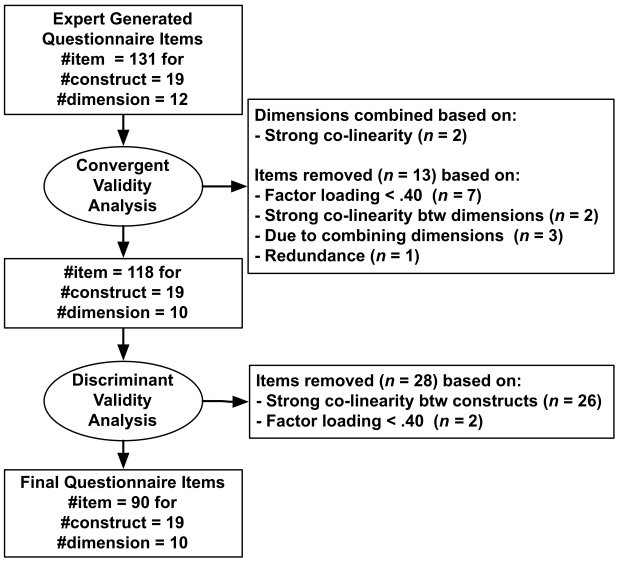

Figure 2: Analysis flow

We carried out all factor analysis via the maximum likelihood method and the promax rotation method, which allows factors to be correlated. Note that here, all dimensions are treated as separate constructs, as dimensions often strongly correlate with other dimensions of their construct. We observed the model's Comparative Fit Index (CFI) value as an indicator of good model fit. The CFI has a range of 0 to 1 and values closer to 1 are a sign of a better fit [1].

In each analysis, we aimed to reduce the number of items in constructs (and dimensions), with a minimum of at least three remaining items. An item could be removed when it had:

- a low standardised factor load score. The items are expected to load highly on their intended construct (or dimension) and low on others (i.e., no cross-loading). A priori, a factor loading smaller than .40 was established [9] as unacceptable for this study; or
- a high modification index with other constructs/dimensions or with another item from another construct/dimension. Modification indices suggest that additional links in the model structure would improve the Chi-square of model fit [13]. In our case, a modification index higher than 3.841 (i.e., 0.05 critical value for Chi-square difference test with one degree of freedom [1]) of an item associated with another construct/dimension indicates that the item has low discriminatory power for its intended construct (or dimension).

The model (where the item was removed) still had to achieve convergence and be admissible; and the removal of the item should result in a higher CFI score. Additionally, links between two items in the same construct/dimension based on modification indices were considered and added to the model for improving the model fit. These processes were repeated until: (i) all models achieved convergence, were admissible, and had a good fit (with $CFI \geq .95$); and (ii) no overlapping constructs/dimensions in the models. An item was removed according to these rules and according to the decision of four judges who examined whether or not it made theoretical sense to remove or retain the item. The judges' discussion continued until a unanimous agreement was reached.

The Short Version of the Questionnaire. We aimed to select one representative item from each construct (and dimension) to create a short version of the ASA questionnaire. A representative item should: (a) have a high standardized factor loading (preferably the highest), and (b) be able to theoretically represent its intended construct/dimension. To assess to what extent the short version could be a substitute for the long version, we analysed the correlation and absolute mean difference between the long and short versions of the questionnaire based on the raw observed data ($n = 532$). Here, for the long version, we used the mean of item scores of each construct/dimension. Finally, we compared whether EFAs of the short and long questionnaires would result in a similar grouping of constructs/dimensions.

ASA Chart. To present the results of the ASA questionnaire in a standardized and easy-to-view manner, we propose the ASA-(web) chart. In the web chart, each of the constructs/dimensions is displayed on a semi-circle radiating outward. We organised them on the chart in groups following the factorial groupings used in the discriminant validity analysis. Within each group, we used the correlation scores between constructs/dimensions as a reference to place them next to each other. The higher correlation score the closer they were set with each other.

RESULT

Participants ($n = 532$) took on average 19.5 minutes ($SD = 8$) to complete the experiment (including reading the instruction, filling

in their consent, and (re)watching the video if needed). The samples of 131 questionnaire items were then standardized, with on average 38 samples (SD = 1.24, range [36 .. 39]) for each of the 14 ASAs. This section presents the results of the analysis. First, we describe the final questionnaire items, then the short version, and finally how to create the ASA chart.

The ASA Questionnaire Items

Preliminary convergent validity analysis of the individual constructs revealed that the dimensions, in three out of four constructs that have dimensions, might not necessarily form one construct (i.e., statistically non-admissible models). To address this, the judges discussed whether to combine such dimensions or to treat them as individual constructs during the analysis. Firstly, the judges decided to combine two dimensions in the construct Performance (i.e., Agent's Performance and User's Performance), because the performance of 'the team' is dependent on the actions of both parties. Secondly, in the construct User-Agent Alliance, the judges agreed to combine the dimensions Task Alliance and Social Alliance, as performing a task in a social setting means both are relevant for the user-agent alliance. Thirdly, the judges decided to analyse two dimensions in the construct Emotional Experience (i.e User's Emotion Presence and Agent's Emotional Intelligent Presence) as separate constructs because the agent's and user's emotion presence are independent. Finally, these (theoretically based) decisions allowed us to run the statistical models.

Convergent Validity Analysis. Analysing individual constructs' convergence [5], aiming for three items per construct, resulted in the removal of 13 (9.9%) out of 131 items. As illustrated in Figure 2, seven items were removed due to their low factor loading ($\beta < .40$). Two items were dropped due to a high correlation between dimensions. Further, the combination of two dimensions in the construct Performance led to the removal of three items that measured *only* the user's performance as ASA researchers might be interested predominantly in the ASA's performance. Finally, one item was taken out from the Construct Personality because of similarity. The judges agreed to remove the redundant item with the lowest factor loading. This resulted in 118 (90.1%) convergent validated items in 19 constructs (CFI M = .99, SD = .02, range [.96 .. 1]).

Discriminant Validity Analysis. The discriminate validity analysis was carried out based on 24 constructs/dimensions resulting from the convergent validity analysis, i.e., 5 dimensions from the construct Agent's Believability, 2 dimensions from the construct Emotional Experience, and 17 individual constructs. The EFA resulted in five factors of constructs/dimensions (with standardized factor load ranging [.41 .. .97], see Table 2). Additionally, three constructs/dimensions were loaded on more than one factor (i.e., Natural Behavior, Human-Like Behavior, and Agent's Enjoyability). As mentioned before, however, these factors are **not** intended to have meaning. The factorial grouping allows us to analyse the constructs/dimensions in five separated models, compared to one large (non-admissible) model.

The discriminant validity analysis [5], aiming for three items per construct, resulted in the removal of 28 items, i.e. due to a very low factor load ($\beta < .40$; 2 items), or due to a poor discriminatory power

Table 2: Initial grouping of 24 constructs/dimensions

ID	Construct/ Dimension	Factors 1	2	3	4	5
NA	1.3 Natural Appearance	.97				
HLA	1.1 Human-Like Appear..	.90				
NB	1.4 Natural Behavior	.64				.41
HLB	1.2 Human-Like Behavior	.59				.57
PF	3. Performance		.89			
AC	13. Agent's Coherence		.77			
UAA	7. User Acceptance ..		.74			
AT	15. Attitude		.71			
AA	12. Agent's Attentiveness		.71			
UE	9. User's Engagement		.65			
UT	10. User's Trust		.64			
AU	2. Agent's Usability		.63			
AE	8. Agent's Enjoyability		.54		.50	
AAS	1.5 Appearance Suitability		.53			
IIS	17. ..Impact on Self-image		.50			
AI	14 Agent's Intentionality		.47			
AEI	18.1 Agent's Emotional ..			.97		
UEP	18.3 User's Emotion ..			.82		
UAI	19. User-Agent Interplay			.68		
SP	16. Social Presence			.64		
SC	5. Agent's Sociability			.53		
SPP	6.1 Agent's Personality ..			.52		
UAL	11. User-Agent Alliance			.46		
AL	4. Agent's Likability				.54	

Note: The construct/dimension numbering following [4]: <construct no>.<dimension no>

Table 3: Classification of items

Factor threshold [15]	Abs(Load score)	#Item n = 90
Excellent	0.71 - 1.00	32 (35.6%)
Very good	0.63 - 0.70	21 (23.3%)
Good	0.55 - 0.62	20 (22.2%)
Fair	0.45 - 0.54	14 (15.6%)
Poor	0.32 - 0.44	2 (.02%)
	< .32	1 (.01%)

for their intended construct ($MI > 3.84$; 26 items). This resulted in 90 validated items (i.e. 76.3% out of 118 items) in 19 constructs (on average 4 items per construct/dimension, see Figure 3). The factor loadings of the items range from .31 to .86 (M = .64, SD = .11, Table 3). One item (in the dimension Natural Behavior in the construct Agent's Believability) has a factor loading of $\beta < .40$. However, the judges decided to not remove the item as the dimension already has three items only. Additionally, the internal consistency test of the constructs/dimensions shows an average reliability of Cronbach's α = .72 (SD = .07, ranging from .60 to .86).

We investigated whether the removal of the items impacted the factorial models. The EFA was re-run and showed that there were now four factorial models, without constructs overlapping into

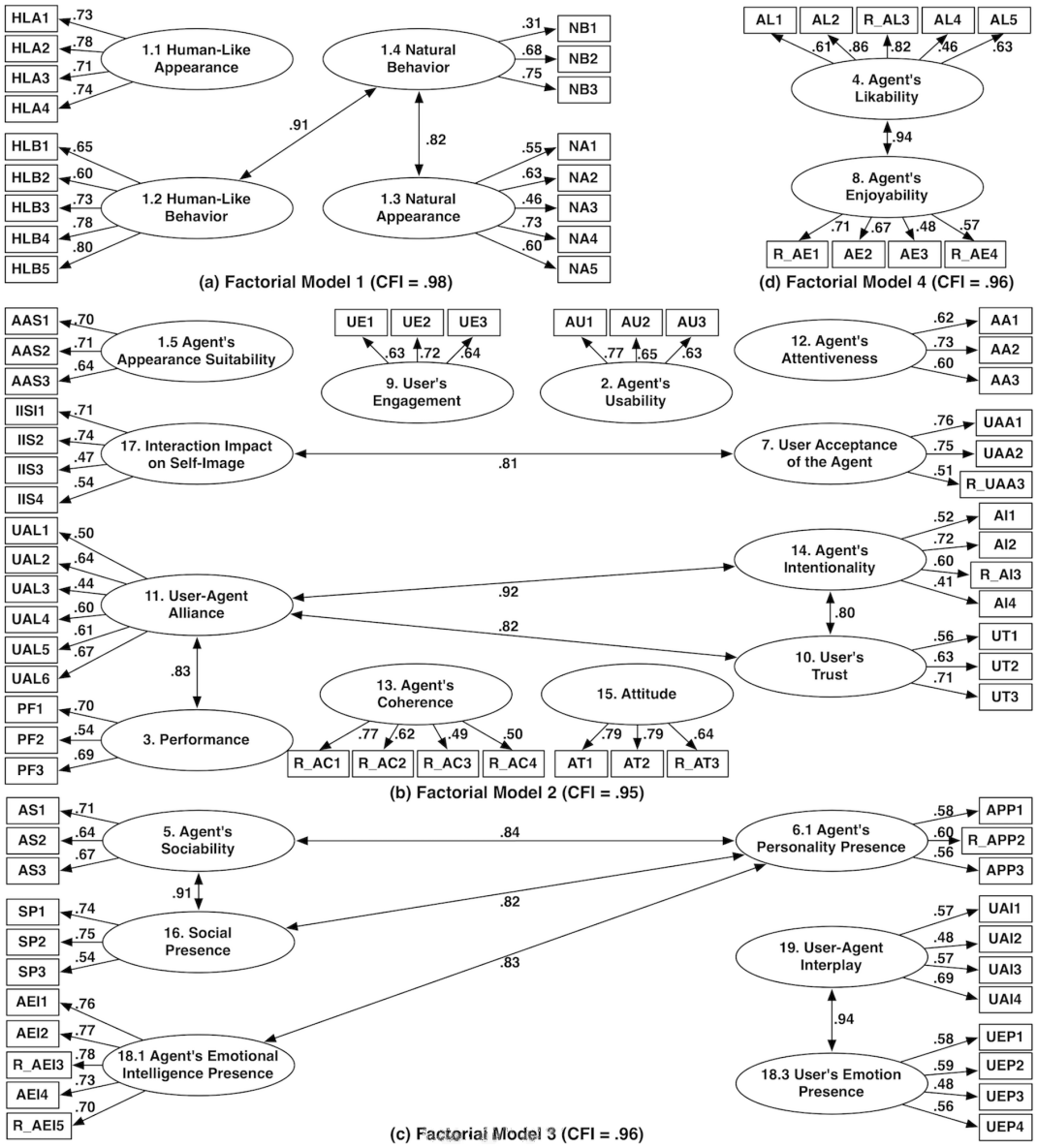

Figure 3: Confirmatory factor analysis diagrams. Links between constructs $\rho \geq .8$ are shown. Note: The construct/dimension numbering following [4]: <construct no>.<dimension no>

multiple factors (Table 4, left side). The CFI scores of the models (CFI range = [.95 .. .98], $M = .96$, $SD = .01$, see Figure 3) show a very good fit. This implies that our a priori expectation, the items are grouped as manifestations of their underlying construct/dimensions, matches with the observed data. In Figure 3, we only show correlations between constructs that are of marginal and moderate concern ($\rho \geq .8$) [14]. The 90 items constitute the full/long ASA questionnaire [4].

The Short Version of the ASA Questionnaire

The short version includes one representative item for each construct/dimension. Each representation item was determined based on the factor loading and the theoretical representation, which was decided by the four judges. These 24 representative items serve as the short version of the ASA questionnaire, see Table 5.

The absolute mean difference between any construct's or dimension's mean on the one hand, and its representative item on the other, was small (range [.00 .. .61], $M = .20$, $SD = .18$). The short and long version of the ASA questionnaire were highly correlated (range ρ =[.71 .. .93], $M = .82$, $SD = .05$). Running an EFA on the representative items showed a similar grouping of the constructs/dimensions as in the Discriminant Validity Analysis, see Table 4 - right side. Only three constructs (in bold) were grouped

Table 4: The grouping of 24 constructs/dimensions and 24 representative items into 4 factors

Factor	24 Constructs/Dimensions	24 Representative Items
1	1.1 Human-Like Appearance, 1.2 Human-Like Behavior, 1.3 Natural Appearance, 1.4 Natural Behavior;	1.1 Human-Like Appearance, 1.2 Human-Like Behavior, 1.3 Natural Appearance, 1.4 Natural Behavior, **5. Agent's Sociability**, **11. User-Agent Alliance**, **16. Social Presence**;
2	1.5 Agent's Appearance Suitability, 2. Agent's Usability, 3. Performance, 7. User's Acceptance of the Agent, 9. User's Engagement, 10. User's Trust, **11. User-Agent Alliance**, 12. Agent's Attentiveness, 13. Agent's Coherence, 14. Agent's Intentionality, 15. Attitude, 17. Interaction Impact on Self-Image;	1.5 Agent's Appearance Suitability, 2. Agent's Usability, 3. Performance, 7. User's Acceptance of the Agent, 9. User's Engagement, 10. User's Trust, 12. Agent's Attentiveness, 13. Agent's Coherence, 14. Agent's Intentionality, 15. Attitude, 17. Interaction Impact on Self-Image;
3	**5. Agent's Sociability**, 6.1 Agent's Personality Presence, **16. Social Presence**, 18.1 Agent's Emotional Intelligence Presence, 18.3 User's Emotion Presence, 19. User-Agent Interplay;	6.1 Agent's Personality Presence, 18.1 Agent's Emotional Intelligence Presence, 18.3 User's Emotion Presence, 19. User-Agent Interplay;
4	4. Agent's Enjoyability, 8. Agent's Likeability;	4. Agent's Enjoyability, 8. Agent's Likeability;

Note: The construct/dimension numbering following [4]: <construct no>.<dimension no>

Table 5: The short version of the ASA questionnaire

ID	Item
HLA	[The agent] has the appearance of a human
HLB	[The agent] has a human-like manner
NA	[The agent] seems natural from its outward appearance
NB	[The agent] reacts like a living organism
AAS	[The agent]'s appearance is appropriate
AU	[The agent] is easy to use
PF	[The agent] does its task well
AL	I like [the agent]
AS	[The agent] can easily mix socially
APP	[The agent] has a distinctive character
UAA	[I / The user] will use [the agent] again in the future
AE	[R] [The agent] is boring
UE	The interaction captured [my / the user's] attention
UT	[I / The user] can rely on [the agent]
UAL	[The agent] and [I / the user] have a strategic alliance
AA	[The agent] is attentive
AC	[R] [The agent]'s behavior does not make sense
AI	[R] [The agent] has no clue of what it is doing
AT	[I see / The user sees] the interaction with [the agent] as something positive
SP	[The agent] is a social entity
IIS	Others would encourage [me / the user] to use [the agent]
AEI	[R] [The agent] is emotionless
UEP	The emotions [I feel / the user feels] during the interaction are caused by [the agent]
UAI	[The agent]'s and [my / the user's] emotions change to what [we / they] do to each other

Note: Codes in the items: [R] refers to a reverse-scoring questionnaire item; [The agent] can be replaced with the ASA's name; and [.. / ..], e.g. [I am / The user is], means to use either one.

differently. This indicates that the short ASA questionnaire is a good representation of the full ASA questionnaire.

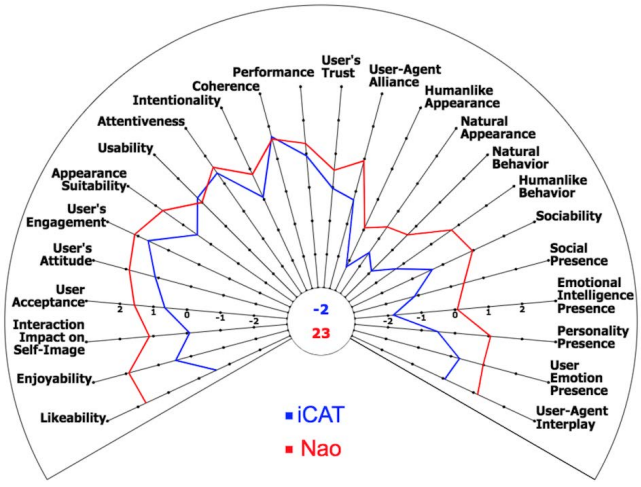

Figure 4: An example of ASA chart: iCAT (Philips) and Nao (Aldebaran Robotics)

ASA Chart

An ASA chart is an informative visual tool to display the scores of an ASA on the 24 constructs/dimensions on a two-dimensional plane. The scores are normalised to a 7-point scale with an interval between -3 to 3 and 0 as the middle point. Each score of a construct/dimension is depicted on an axis that emerges from a common central point. We arranged the constructs/dimensions on the ASA chart based on their distance in the factor analysis and theoretical similarities (see Table 2 and 4). The total score, rounded up, of all the constructs/dimensions is displayed in the middle of the chart which is called the *ASA-score*. The scripts for generating ASA charts are online available [4].

Comparing the ASA chart of two illustrative agents, iCat and Nao (see Figure 4), demonstrates the insights that these charts bring. Here, the chart represents values based on the mean of the 90 questionnaire items calculated from the corresponding agents' raw

observed data. The ASA-scores of all the ASAs used in our study can be seen in Table 1.

DISCUSSION AND CONCLUSION

This paper presents the final 90 items of the ASA questionnaire that can measure the 19 constructs in which IVA researchers are most interested. The items have acceptable reliability and good convergent and discriminant validity. Correlations between some constructs and dimensions exist, reflecting the overlap between the theoretical constructs of interest to the community. The expert input in the development of this questionnaire means that we place more emphasis on content validity (i.e., we measure what the community studies) and less on discriminant validity (i.e., we allow overlap between constructs). Additionally, there are potential causal links between constructs that can explain the observed correlations (e.g., on the one side Agent's Personality Presence, and the other Agent's Sociability, Social Presence, and Emotional Intelligence Presence).

To allow the community to compare agents, we need a widely used standardised measure. This means that such a measure needs to be brief and easy to administer. To this end, we identified 24 representative items, one for each construct and dimension. We suggest that all IVA researchers use this short ASA questionnaire whenever they are evaluating their agents. Researchers can use all the items of the constructs that they are most interested in to increase the power of the measurement for those constructs.

Finally, we developed the ASA chart to display the scores on each construct in one glance. This allows for easy comparison between different artificial social agents: the ASA chart shows the profile of an ASA concerning to users' views about the ASA and their interaction. The arrangement of constructs in the ASA chart suggests subsets that conform with the expert-construct categorisation in [8], i.e. from the left to right: agent's social traits, agent's basic properties, human-agent interaction quality, agent's role performance, and human impressions left after the interaction.

An ASA-score is the summation of all ASA constructs' scores, if you score high on all constructs you will have a high ASA-score, and if your agent scores low on all constructs the ASA-score will be low. For example (see Table 1), although both are a disembodied ASA, HAL 9000 (ASA-score = 14) is more sociable than Siri (ASA-score = 13). On the other hand, Marcus (ASA-score = 25), a cyborg in the Terminator movie, has a higher fidelity than a Nao robot (ASA-score = 23). Therefore, open questions still remain: what does the ASA-score mean? But also, how would a human score on this scale? And what would that score then mean, perhaps a gold standard?

Next, in spite of our previous study comparing (imagined) first and third-person perspectives [6], an evaluation with participants who interacted with an ASA (real first-person perspective) remains to be done. Despite that the rating of an ASA might be different from real interaction and a video experience, we assume that the correlations between items and constructs is unlikely to differ. The latter is important because this study was essentially a correlation analysis. These questions, however, are left for future exploration. For now, the next step of the project are: (1) Determine the generalisation performance of the long and short questionnaire versions (i.e. cross-validation: fit model on data set from a new set of ASAs); (2) Determine criteria validity (i.e., predictive validity: agreement with predicted future observations) and concurrent validity (e.g., agreement with other 'valid' measures); (3) Translate the questionnaire (i.e., forward/backward translation); and (4) Develop a normative data set.

ACKNOWLEDGMENTS

This publication is sponsored by the Dutch 4TU - Humans and Technology, Pride and Prejudice project.

REFERENCES

[1] Niels J. Blunch. 2013. *Introduction to Structural Equation Modeling using IBM SPSS Statistics and AMOS* (2nd eds. ed.). SAGE, City Road, London. https://doi.org/10.4135/9781526402257

[2] Henrica C. W. de Vet, Herman J. Adér, Caroline B. Terwee, and François Pouwer. 2005. Are factor analytical techniques used appropriately in the validation of health status questionnaires? A systematic review on the quality of factor analysis of the SF-36. *Quality of Life Research* 14 (2005), 1203–1218. https://doi.org/10.1007/s11136-004-5742-3

[3] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo, and Louis-Philippe Morency. 2014. SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support. In *Proc. of AAMAS '14* (Paris, France). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1061–1068.

[4] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. 2022. Artificial Social Agent Questionnaire Instrument. (2022). https://doi.org/10.4121/19650846

[5] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. 2022. Data and analysis underlying the research into the Artificial-Social-Agent Questionnaire: Establishing the long and short questionnaire versions. (2022). https://doi.org/10.4121/19758436

[6] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, and Willem-Paul Brinkman. 2021. Questionnaire Items for Evaluating Artificial Social Agents - Expert Generated, Content Validated and Reliability Analysed. In *Proc. of IVA'21*. ACM, NY, USA, 84–86. https://doi.org/10.1145/3472306.3478341

[7] Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Amal Abdulrahman, and Willem-Paul Brinkman. 2019. What Are We Measuring Anyway? - A Literature Survey of Questionnaires Used in Studies Reported in the Intelligent Virtual Agent Conferences. In *Proc. of IVA'19*. ACM, NY, USA, 159–161. https://doi.org/10.1145/3308532.3329421

[8] Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Andrea Bönsch, and Willem-Paul Brinkman. 2020. The 19 Unifying Questionnaire Constructs of Artificial Social Agents: An IVA Community Analysis. In *Proc. of IVA'20*. ACM, NY, USA, 1–8. https://doi.org/10.1145/3383652.3423873

[9] Eva Knekta, Christopher Runyon, and Sarah Eddy. 2019. One Size Doesn't Fit All: Using Factor Analysis to Gather Validity Evidence When Using Surveys in Your Research. *CBE life sciences education* 18, 1 (2019), rm1. https://doi.org/10.1187/cbe.18-04-0064

[10] Neuman W. Lawrence. 2014. *Social Research Methods: Qualitative and Quantitative Approaches* (7 ed.). Pearson Education Limited, Harlow.

[11] Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. 2013. I Can Help You Change! An Empathic Virtual Agent Delivers Behavior Change Health Interventions. *ACM Trans. Manage. Inf. Syst.* 4, 4 (2013), 1–28. https://doi.org/10.1145/2544103

[12] Gary Mckeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2013. The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *Affective Computing, IEEE Transactions on* 3 (08 2013), 5–17. https://doi.org/10.1109/T-AFFC.2011.20

[13] Yves Rosseel. 2012. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software* 48, 2 (2012), 1–36. http://www.jstatsoft.org/v48/i02/

[14] Mikko Rönkkö and Eunseong Cho. 2022. An Updated Guideline for Assessing Discriminant Validity. *Organizational Research Methods* 25, 1 (2022), 6–14. https://doi.org/10.1177/1094428120968614

[15] James E. Sallis, Geir Gripsrud, Ulf Henning Olsson, and Ragnhild Silkoset. 2021. Factor Analysis. In *Research Methods and Data Analysis for Business Decisions: A Primer Using SPSS*. Springer International Publishing, Cham, 223–243. https://doi.org/10.1007/978-3-030-84421-9_12