The 19 Unifying Questionnaire Constructs of Artificial Social Agents
An IVA Community Analysis

Fitrianie, Siska; Bruijnes, Merijn; Richards, Deborah; Bönsch, Andrea; Brinkman, Willem Paul

DOI
10.1145/3383652.3423873

Publication date
2020

Document Version
Final published version

Published in
Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents, IVA 2020

The 19 Unifying Questionnaire Constructs of Artificial Social Agents: An IVA Community Analysis

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

Merijn Bruijnes
Delft University of Technology
Enschede, the Netherlands
m.bruijnes@tudelft.nl

Deborah Richards
Macquarie University
Sydney, NSW, Australia
deborah.richards@mq.edu.au

Andrea Bönsch
Visual Computing Institute
RWTH Aachen University
Aachen, Germany
boensch@vr.rwth-aachen.de

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

ABSTRACT

In this paper, we report on the multi-year Intelligent Virtual Agents (IVA) community effort, involving more than 80 researchers worldwide, researching the IVA community interests and practises in evaluating human interaction with an artificial social agent (ASA). The effort is driven by previous IVA workshops and plenary IVA discussions related to the methodological crisis on the evaluation of ASAs. A previous literature review showed a continuous practise of creating new questionnaires instead of reusing validated questionnaires. We address this issue by examining questionnaire measurement constructs used in empirical studies between 2013 to 2018 published in the IVA conference. We identified 189 constructs used in 89 questionnaires that are reported across 81 studies. Although these constructs have different names, they often measure the same thing. In this paper, we, therefore, present a unifying set of 19 constructs that captures more than 80% of the 189 constructs initially identified. We established this set in two steps. First, 49 researchers classified the constructs in broad theoretically based categories. Next, 23 researchers grouped the constructs in each category on their similarity. The resulting 19 groups form a unifying set of constructs, which will be the basis for the future questionnaire instrument of human-ASA interaction.

CCS CONCEPTS

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

KEYWORDS

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

ACM Reference Format:
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 *IVA '20: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20), October 19–23, 2020, Virtual Event, Scotland Uk*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3383652.3423873

IVA '20, October 19–23, 2020, Virtual Event, Scotland Uk

ACM ISBN 978-1-4503-7586-3/20/09…$15.00
https://doi.org/10.1145/3383652.3423873

1 INTRODUCTION

We investigated which measurement constructs in user-evaluations are used by the Intelligent Virtual Agent (IVA) community. A variety of Artificial Social Agents (ASAs) exist, ranging from chat bots and computer-controlled virtual humanoid agents to virtual and physical robots. User-evaluations of these ASAs are non-trivial and reoccurring tasks in science. The work presented in this paper is part of a larger effort that includes all sub-fields of the ASA community. It aims at developing a validated standardised questionnaire instrument for evaluating human interaction with ASAs that: (i) has an ability to make a standardised statement about quality of the ASA; (ii) is grounded in examples of current and popular ASAs; and (iii) has an ability to make a statement about the various aspects and dimensions expected to be relevant to capture an ASA's quality. Note that the aim is not to replace questionnaires for specific constructs, but to provide a broad-spectrum questionnaire to give a 'general impression' of the ASA's quality. To achieve these goals, a call went out to several mailing lists to create an open work-group to share ideas and efforts. Currently, over 80 people (self-selected to) participate in the work-group's open science framework platform[1]. The work group has put forward a plan consisting of multiple steps, including: (1) Determine the conceptual model (i.e., examine existing questionnaires and foster discussions among experts); (2) Determine the constructs and dimensions (i.e., check face validity among experts and grouping of existing constructs); (3) Determine an initial set of construct items (i.e., content validity analysis: reformulate items into easy to understand and 'ASA-appropriate' questionnaire items); (4) Confirmatory factor analysis to examine construct validity; (5) Establish the final item set with the provision to create a long and short questionnaire version; (6) Determine criteria validity (i.e., predictive validity: agreement with predicted future observations) and concurrent validity (e.g., agreement with other 'valid' measures); (7) Translate the questionnaire (i.e., forward/backward translation); and (8) Develop a normative data set.

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

The effort is motivated by discussions across the IVA community and at IVA conference workshops on methodology in 2018 and 2019. It was noted that a methodological crisis is occurring in research in the field of IVA and specifically in the evaluation of ASAs. As a common practice, after constructing an agent and modelling its deliberative (physical and/or social) controls [5], researchers sought to establish an empirically grounded understanding of the agent, its interaction with humans, and how it ultimately causes certain outcomes in domains, such as health, entertainment, and education [2]. However, it is difficult to compare agents that are used in studies, and even more difficult to replicate the scientific findings or validate claims of the impact of these studies. This is due to the trend to create new measurement instruments instead of reusing an existing one. In fact, analysing questionnaires that were reported across 81 papers in IVA conference proceedings between 2013 to 2018, Fitrianie et al. [3] found that the vast majority of questionnaires (more than 76%) were only used by a single study. We therefore argue that we should move towards creating a unified measurement instrument, which researchers could use as a common base to measure a set of shared constructs to describe the interaction experience with an ASA.

In this paper, we show our progress towards the unified measurement instrument for ASAs. In essence, the approach taken was to group the set of 189 constructs on their similarity and regard each group as a unifying construct. However, directly grouping 189 constructs was mentally too hard. Using a divide and conquer approach, we therefore broke up the task into two manageable steps. First, the work group members assigned 189 constructs to seven theory-based categories, referred to as concepts. Second, in seven card-sorting tasks, the work group members grouped constructs, which were associated with a concept, into groups based on their similarity. The new groups were used to define the set of 19 unifying constructs, which is the main contribution of this paper. In addition, the paper also provides an insight into the reliability of this process by quantifying the amount of agreement of the work group members and also the total number of constructs that can be captured by the unifying set.

2 STEP 1 - THEORY-BASED CATEGORISATION

The first step towards the unifying set was to group the 189 constructs identified in the literature into theoretical categories. For this, the work group first decided on the scope of the measurement instrument, and secondly identified categories within this scope. Once this was established, members of the work group categorised each construct. Below we explain these steps and the results.

2.1 Scoping the Measurement Instrument

To scope our measurement instrument we focus only on the interactions between humans and ASAs. However, interactions exist in a wider model of pre-existing entities on the one hand and a context dependent process leading to an outcome beyond the interaction on the other hand (see Figure 1).

People's beliefs and expectations towards an ASA are influenced by pre-existing factors external to the current interaction, such as demographics and personality of the user, prior experiences with ASAs, and formal education, but also the context of interaction, such as environmental settings (e.g., public or private interactions), the domain type (e.g., health, entertainment, education), the agent's embodiment, and the interaction method (e.g., free natural speech or selection of presented buttons). Although many studies capture these factors, and we recognise the potential importance of these factors in addressing the goal of a particular study, they are not in the scope of our measurement instrument and as such are considered as *external entities*. This is because these entities are predefined before the initial interaction and often static during subsequent interactions.

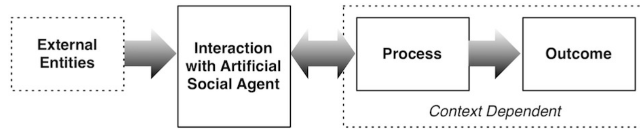


Figure 1: A world model of human-ASA interactions.

An *interaction* between a human and an ASA can be considered as a series of communication acts/events. The purpose of such communication is often not only developing a relationship, in fact, a developed relationship facilitates jointly performing some task [4]. A joint task is a *process* that consists of actions embedded in some context geared towards some outcome. Some of the actions may be jointly accomplished and some may be accomplished more or less individually resulting in an *outcome*. We assert that people (deliberately) engage in a process to produce an outcome. For example, individuals engage in an interaction with a weight-loss therapist ASA to reduce their weight as an outcome. They make conscious attempts (i.e., engage in the process) to affect behaviour that is under their control, such as controlling eating behaviour and doing physical exercises. A therapist agent can facilitate a user in this process by, for example, recommending food and guiding physical exercises. The beliefs that people have toward an ASA and the interaction shape their (continued) involvement with the ASA [1] and thus with the facilitation that the agent offers. The interaction with the agent itself shapes the beliefs that the user has towards the agent. In other words, the process of interaction has as a (potential) outcome an evolved belief of the user. For example, a pleasant interaction with an agent might increase the user's belief that future interactions with the agent will also be pleasurable. Summarising, we are interested only in the measurement of the interactions with ASAs.

We zoom in on the concepts that are necessary to describe human-ASA interaction in Figure 2. There must be two (sets of) agents: 'real' social agents (i.e., human users) and artificial social agents. Additionally, there has to be a reciprocal information exchange between these parties: they have to communicate and take into account what the other party has contributed. Each party possesses certain properties that lead to, and influence, the interaction. On the other hand, the interaction itself has properties that influence both parties.

2.2 Theoretical Concepts

We present an ad-hoc model that serves as a theoretical starting point to develop a community shared consensus that will underpin

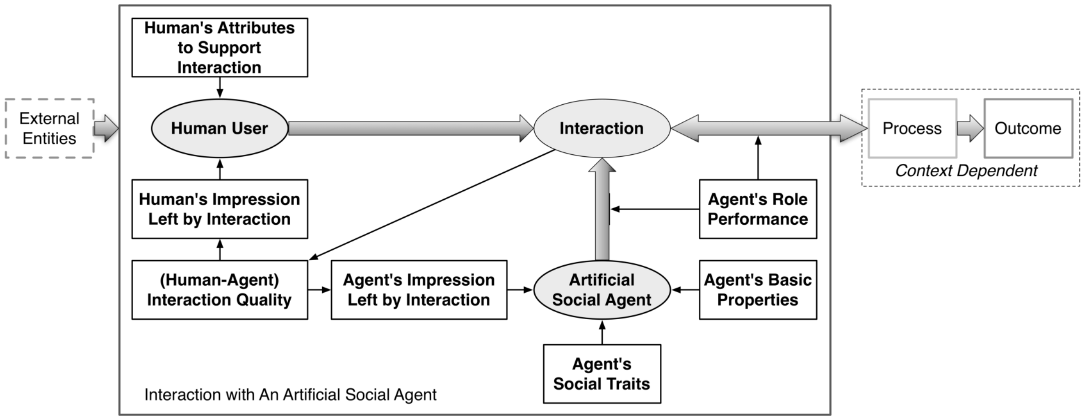


Figure 2: The relationship between successive clusters of factors regarding human-ASA interaction, with the box "interaction with an artificial social agent" defined in more detail.

our measurement instrument (i.e., "we need something to start the conversation"). The success of the initial grouping step depended on the constructs the work group members could agree upon assigning to a specific concept. Therefore, these concepts were defined at a relatively generic level, covering three broad clusters of concepts, namely the ASA, the human user, or the interaction between the two. Below we describe these clusters in more detail, with illustrative examples of constructs that were associated with a specific concept by the work group (see Section 2.3). In the following we give definitions of the concepts of the model of human-ASA interactions (Figure 2) and accompany each concept with examples of measurement constructs identified from IVA literature [3].

2.2.1 Concepts related to Artificial Social Agent. An ASA, as an autonomous and intelligent entity, has basic properties, social traits, and abilities to perform its role as a social agent in an interaction. The *Agent's Basic Properties* concept is defined as the ability to act independently based on its acquired knowledge and skill in response to varying situations and past experience. Examples of constructs that belong to this concept are agent's competence and perceived intelligence (see Table A1). The *Agent's Social Traits* is related to the abilities of an ASA to interact with the human in a social, warm, sensitive and personal manner. To be able to do this an ASA could include skills to handle its own emotions, to understand the other's emotion or to influence the other's emotion. Constructs classified into this concept relate to the emotional skills of ASAs (see Table A1), such as agent's personality and whether the agent is perceived as sociable. The *Agent's Role Performance* refers to how well an ASA performs its role (during the interaction) as a supporter of the user's intended outcomes. Depending on the nature of the process involving the human-ASA interaction, an ASA has a specific set of roles to perform. Its role can be as a supportive entity that helps a user to conduct certain tasks. Agent's perceived utility is an example of a construct that comprises how well the ASA supports the user (see Table A1).

2.2.2 Concepts related to Human User. A human user has certain attributes that allow him/her to interact with an ASA. Hence, the concept *Human's Attribute to Support Interaction* is related to the user's capacity to engage in an interaction with ASAs. Examples of identified constructs are the user's performance and self-efficacy (see Table A1).

2.2.3 Concepts related to Interaction. The interaction leaves an impression on the actors involved in the interaction. This impression affects how these actors conduct themselves in the interaction. Based on this view, the *Human's Impression Left by Interaction* concept is defined as the effect the interaction had on the user and how he/she responds to the interaction. Constructs attached to this concept relate to an idea, feeling, or opinion of the user about the ASAs and the interaction, such as User's Perceived Enjoyment and Trust (see Table A1). On the other hand, the *Agent's Impression Left by Interaction* refers to the effect the interaction has on an ASA and how it subsequently responds to the interaction. Constructs in this concept measure the ASA's feelings about others (including the human user) and the interaction, for example agent's botheredness and perceived positiveness (see Table A1). Finally, the interaction itself forms a central concept to describe the interactive interplay between the actors. The interaction itself too has properties that deal mostly with the quality of communication. This idea motivates the concept *(Human-Agent) Interaction Quality*. It refers to how a user characterises his/her interaction with an agent (e.g., 'was I understood'). Often this quality explains the mechanism with which an interaction results in a particular user's impression of the agent. Perceived behavioural interdependence and (user-agent) bonding are a few examples of identified constructs (see Table A1).

2.3 Method for Construct Categorisation

Fitrianie et al. [3] claimed that few questionnaires were re-used in multiple IVA studies and most questionnaires were used once. However, we argue that there are multiple attempts across the studies to measure similar underlying concepts. In the previous section we proposed a model of concepts that are relevant to describe human-ASA interactions. The analysis in this section aimed at assigning the constructs found in the literature to the concepts of our model based on a majority view of the participants. The appropriateness of proposed concepts, was examined by analysing how many of all the constructs could be successfully assigned with the majority support. Also, insight into the distribution of constructs shows the focus of current human-ASA interaction studies in the IVA community.

Members of the work group were invited to classify, as independent coders, the 189 constructs into the model. Specifically, constructs could be assigned to the seven concepts defined in the box 'Interaction with Artificial Social Agent', or to one of the three exogenous categories (i.e. external entities, process, and outcome). Additionally, an 'other' rating was available for constructs that did not fit in any category. For these constructs coders could suggest a new category. The constructs were presented to the coders in a randomised order.

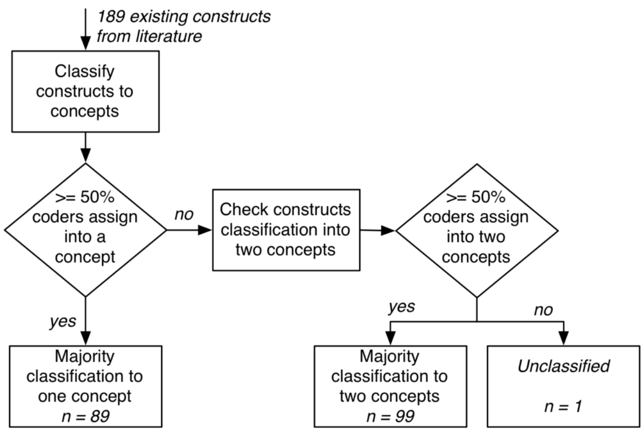


Figure 3: Analysis process of construct classifications in concepts.

A total of 49 coders were involved in the study, with 12 coders (24.5%) classifying all 189 constructs. The average number of coders per construct was 13 and the minimum number of coders was 12. The classification of a construct in a concept was accepted if at least 50% of coders assigned the construct to that concept (see the flowchart in Figure 3). For constructs that could not be assigned to a single concept, a classification of a construct in two concepts was accepted if the majority and at least 50% of coders assigned the construct to those two concepts. If a construct cannot be classified into at most two concepts, we regarded the construct as unclassified. Finally, new concepts proposed by the coders were collected. Related data and files to this study are available online at our Open Science Foundation-repository.[2]

2.4 Result of Construct Categorisation

Of the coders, 65.8% agreed to classify 89 constructs (47.1%) to one concept and 66.5% of coders agreed to classify 99 constructs (52.4%) to two concepts (see Table 1). Note that 11 constructs were classified in 'other' (about 6%), and as such fall outside the scope of our measurement instrument. Examples of these 'other' constructs are: Computer Literacy, User's Personality Trait, and User's Preference Role in Decision Making. One construct could not be classified: no agreement was reached for Perceived Similarity. This leaves

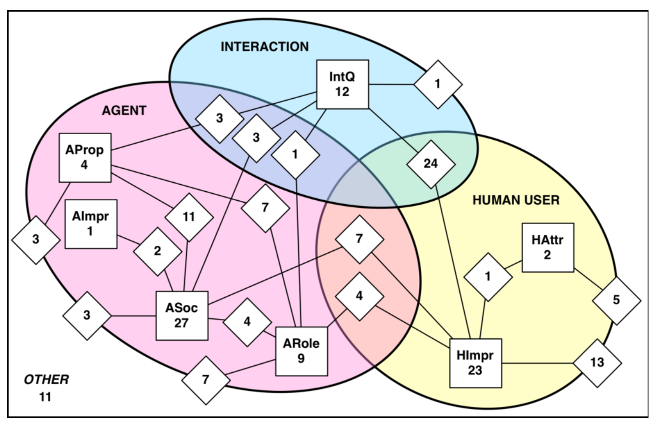
[2]Results of Study 1: Defining Categories, https://osf.io/82vtr/wiki/home/

Table 1: The number of constructs classified into concepts defined in Figure 2

Concept	*One*	%	*AA*	*Two*	%	*AA*	*Tot*
AProp	4	2.1	58.7	24	12.7	32.9	28
ASoc	27	14.3	72.8	30	15.9	32.3	57
ARole	9	4.8	72.4	23	12.2	30.0	32
AImpr	1	.5	61.5	2	1.1	26.1	3
IntQ	12	6.3	60.2	32	16.9	33.3	44
HImpr	23	12.2	61.2	49	25.9	35.0	72
HAttr	2	1.1	59.2	6	3.2	27.3	8
OTHER	11	5.8	63.4	32	16.9	35.0	43
Unclassified							1
Sum	89	47.1		99	52.4		189

Note: AProp = Agent's Basic Properties; ASoc = Agent's Social Traits; ARole = Agent's Role Performance; AImpr = Agent's Impression ...; IntQ = Interaction Quality; HImpr = Human's Impression ...; HAttr = Human's Attributes ... ; *OTHER* = categories outside the step "Interaction with ASA"; *One* = #classifications to one concept; *Two* =#classifications to two concepts; *Tot* = total #classifications; % = #classification / 189 constructs; *AA* = average (%) of coders that agree on classification.

177 (93.7%) constructs that were attributed to the seven theoretical concepts.



Note: Squares = #classification to one concept; Diamonds = #classification to two concepts; AProp = Agent's Basic Properties; ASoc = Agent's Social Traits; ARole = Agent's Role Performance; AImpr = Agent's Impression ...; IntQ = Interaction Quality; HImpr = Human's Impression ...; HAttr = Human's Attributes ... ; and OTHER = categories outside the step "Interaction with ASA"

Figure 4: The number of constructs classified into concepts defined in Figure 2.

The composition of the classifications is shown in Figure 4. It shows the number of assignments to singular concepts in squares and assignments to two concepts in diamonds. The colored zones indicate whether the concepts are related to properties of the agent, human, or interaction. The results give insight into the research focus of the community. For example, the concept Human's Impression Left by Interaction (HImpr) has the most constructs attributed to it (38.1%), which was followed by the Agent's Social Traits (ASoc) (30.2%). Of the constructs in the cluster Interaction (blue area in Figure 4) most constructs (72.7%) are attributed to two concepts (i.e., most are in the diamonds). Less constructs were attributed to the agent's basic properties, its role performance, and the quality of the interaction itself. The least number of constructs were assigned to concepts measuring people's perception of the impact that the interaction has on the agent, and that aim to capture the human's capacity to engage in an interaction with an ASA.

2.5 Discussion on Categorisation

The results show that work group members were able to agree on the assignment to a single or combination of two theoretical concepts for 93.7% of the constructs. The finding indicates remarkable agreement on how to split up the constructs into more manageable groups for the second card-sorting task step. It shows that our model provides good coverage for describing the measurement constructs that were extracted from IVA papers, but more importantly, we found that all seven concepts had a number of constructs assigned to them and that the constructs are somewhat distributed. Overall, both the set of constructs collected and the model cover the currently utilised measurement constructs, which arguably represent the research interests of ASA community.

The composition of the classification to two concepts shows that the focus of studies into human-ASA interaction emphasises evaluating the social aspect of the agent *and* its impact on the user. This can be seen from the fact that on one hand, about 30% of constructs that were classified in concepts related to the ASA (pink area in Figure 4) are part of the concept Agent's Social Traits (ASoc). On the other hand, more than half of the constructs assigned to the concept Interaction Quality (54.4%) are shared with the concept Human's Impression Left by Interaction.

Additional concepts were proposed by (four) coders for constructs that they could not classify: usability, social presence, agent's embodiment, agent's appearance, agent's attributes to support interaction, human-agent relationship, spatial presence, virtual environment of agent, and human's physiological state during interaction. However, the proposed concepts are already covered in the existing constructs or outside the scope. For example, suggested concepts that cannot be generalised across different ASAs (e.g., measures specific to embodiment, modality, or display technology) or that cannot be measured in the scope of a questionnaire (e.g., physiological measures) were excluded.

3 STEP 2 - GROUPING CONSTRUCTS AND DEFINING THE UNIFYING CONSTRUCT SET

The next step was to explore whether agreement existed between work group members on how groups could be formed from constructs that were assigned to a concept. If so, these groups would form the basis for a unifying construct. In other words, this step aimed to group constructs that, while formulated and defined differently, measure the same (or very similar) underlying thing in the eyes of work group members.

3.1 Method

The work in this step is split in two parts: the grouping of constructs and using these groupings to define the unifying construct set. Related data and files to this study are available online at our Open Science Foundation-repository.[3]

Table 2: The number of groups in each of the seven card-sorting tasks

Task	*#Constructs in Task*	*#Groups*	*Average Agreement*	*#Excluded Constructs*
AProp	28	5	56.2	6
ASoc	57	11	58.5	13
ARole	32	9	60.3	2
AImpr	3	1	69.6	1
IntQ	44	11	63.6	6
HImpr	72	12	57.3	11
HAttr	8	3	72.7	2
Total	177	52		25*

Note: AProp = Agent's Basic Properties; ASoc = Agent's Social Traits; ARole = Agent's Role Performance; AImpr = Agent's Impression ...; IntQ = Interaction Quality; HImpr = Human's Impression ...; HAttr = Human's Attributes ... ; *#Construct intask* = #constructs initially assigned; *Average agreement* = average (%) of coders that agree on classification; *#Excluded constructs* = #constructs (from that task) excluded in analysis; * = excluded from all tasks;

3.1.1 Establishing Construct Groups. Members of the work group were invited to classify, as independent coders, to organise measurement constructs into groups using seven card-sorting tasks. Each task corresponded to a concept in Figure 2. In a task, each card represented a measurement construct that was classified to that concept. The number of constructs in each task differs (see *#Construct* in Table 2). Constructs that were attributed to two concepts (i.e., the diamonds in Figure 4) were presented in the card sorting task for both concepts. Cards were presented to the coders in a randomised order. A task was considered to be completed by a coder if all cards in the task were organised into groups. Finally, coders could suggest names for the groups they created.

A grouping of constructs was accepted if at least 50% of coders agree to place these constructs together. The agreement between coders is considered as the strength of these groups.

3.1.2 Defining Unified Construct Set. The flowchart in Figure 5 shows the process of defining the unified construct set. Here, we investigate whether groups have 'overlapping constructs'. This means that in these groups at least one construct is placed also in another group (remember that constructs related to two concepts, 'diamonds' in Figure 4, are presented in two card sorting tasks).

Three coders analyse each accepted group using the following steps: Firstly, for each group they create an initial title (consisting of keywords describing the group) by looking at the included constructs' names and their definitions and by the titles suggested by the coders. In the next step they combine groups based on the

[3] Results of Study 2: Defining Constructs, https://osf.io/ysfmx/wiki/home/

Table 3: 19 measurement constructs and dimensions for evaluation human interaction with an ASA

No	Construct/Dimension	Definition
1.	Agent's Believability	The extent to which a user believes that the artefact is a social agent
1.1.	Human-Like Appearance	The extent to which a user believes that the social agent appears like a human
1.2.	Human-Like Behaviour	The extent to which a user believes that the social agent behaves like a human
1.3.	Natural Appearance	The extent to which a user believes that the social agent's appearance could exist in or be derived from nature
1.4.	Natural Behaviour	The extent to which a user believes that the social agent's behaviour could exist in or be derived from nature
1.5.	Agent's Appearance Suitability	The extent to which the agent's appearance is suitable for its role
2.	Agent's Usability	The extent to which a user believes that using an agent will be free from effort (future process)
3.	Performance	The extent to which a task was well performed (past performance)
3.1.	Agent's Performance	How well an agent does a task
3.2.	User's Performance	How well the user does a task
4.	Agent's Likeability	The agent's qualities that bring about a favourable regard
5.	Agent's Sociability	The agent's quality or state of being sociable
6.	Agent's Personality	The combination of characteristics or qualities that form an individual's distinctive character
6.1.	Agent's Personality Presence	To what extent the user believes that the agent has a personality
6.2.	Agent's Personality Type	The particular personality of the agent
7.	User Acceptance of the Agent	The willingness of the user to interact with the agent
8.	Agent's Enjoyability	The extent to which a user finds interacting with the agent enjoyable
9.	User's Engagement	The extent to which the user feels involved in the interaction with the agent
10.	User's Trust	The extent to which a user believes in the reliability, truthfulness, and ability of the agent (for future interactions)
11.	User-Agent Alliance	The extent to which a beneficial association formed
11.1	Task Alliance	The extent to which an association formed for performing a common goal
11.1	Social Alliance	The extent to which a close and harmonious association formed in which both understand each other's feelings and ideas
12.	Agent's Attentiveness	The extent to which the user believes that the agent is aware of and has attention for the user
13.	Agent's Coherence	The extent to which the agent is perceived as being logical and consistent
14.	Agent's Intentionality	The extent to which the agent is perceived as being deliberate and has deliberations
15.	Attitude	A favourable or unfavourable evaluation toward the interaction with the agent
16.	Social Presence	The degree to which the user perceives the presence of a social entity in the interaction
17.	Interaction Impact on Self-Image	How the user believes others perceive the user because of the interaction with the agent
18.	Emotional Experience	A self-contained phenomenal experience. They are subjective, evaluative, and independent of the sensations, thoughts, or images evoking them
18.1	Agent's Emotional Intelligence Presence	To what extent the user believes that the agent has an emotional experience and can convey its emotions
18.2	Agent's Emotional Intelligence Type	The particular emotional state of the agent
18.3	User's Emotion Presence	To what extent the user believes that his/her emotional state is caused by the interaction or the agent
18.4	User's Emotion Type	The particular emotional state of the user during or after the interaction with the agent
19.	User-Agent Interplay	The extent to which the user and the agent have an effect on each other

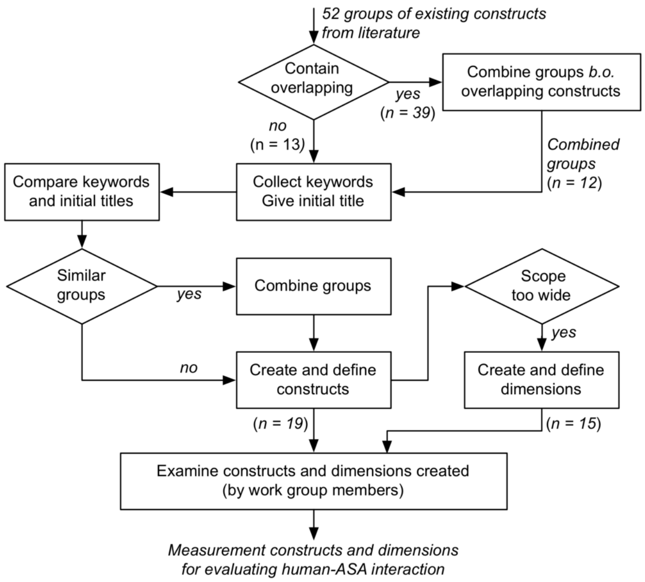


Figure 5: Analysis process of the card-sorting results.

similarity of the domain of the keywords and suggested titles. The coders looked at the semantics of the constructs in the groups and at the names for groups given by coders. When there is only one set of unique domain groups left, the coders create and define measurement constructs based on these groups. The initial titles are used as the title of the resulted constructs. Further, they re-examine all the resulting constructs. If the scope of a construct is too wide, for example because it contains keywords that encompass different measurement domains, they divide the construct into dimensions. Finally, the coders create titles and definitions for those resulting dimensions in collaboration with the work group members.

3.2 Results

First we discuss the results of the grouping of constructs and then using these groupings to define the unifying construct set.

3.2.1 Construct Grouping. In the study, 23 independent coders participated with 12 coders (52.2%) completing all seven tasks (on average tasks were completed by 17 coders). The 177 constructs that, in step 1, were attributed to the human-ASA interaction were included in the card sorting tasks. More than 50% of coders agreed on sorting 152 constructs (out of 189, 80.4%) into 52 card-sorting groups (see Table 2). There were 25 constructs excluded because less than 50% of the coders agreed on how to group these (see #*Excluded* in Table 2).

3.2.2 Defining Unified Constructs. In total, 39 groups (out of 52, 75%) contained overlapping constructs. To resolve this redundancy, three coders discussed till they reached consensus to re-arrange and combine these groups based on their overlapping constructs. This resulted in 12 groups of combinations. Combined with the 13 groups without overlapping constructs, there are 25 individual groups (see Table A2). For each of the 25 groups the three coders created a title (consisting of keywords describing the group) by looking at the included constructs' names and their definitions and by the titles suggested by the coders (see Table A2). Subsequently, the same three coders determined the uniqueness of each group: groups that contained similarities were combined and groups that were too broad in scope were split. The final measurement construct that each group is represented in can be seen in column **Const.** in Table A2. This resulted in 19 final measurement constructs for evaluating human interaction with ASA. Note that five constructs consist of multiple dimensions (in total there are 15 dimensions). Work group members were invited to discuss the 19 constructs. Eight members of the work group finalised the names and descriptions of the 19 constructs (see Table 3). The notes of the discussion can be found in the Open Science Foundation repository.[4].

4 DISCUSSION AND CONCLUSION

The work presented in this paper is part of a community effort that aims at developing a validated standardised questionnaire instrument for evaluating human interaction with ASAs. Grouping all the individual measurement constructs used in recent years by IVA researchers yielded 19 measurement constructs. These unified measurement instruments can be used as a set of shared constructs to describe the interaction experience with an artificial social agent.

Obviously, these efforts do not provide the definitive list of constructs that the reflect the interest of the community for all time. In the future, this construct set might be updated or new constructs might be included to adapt to the evolving interests in the field. This could be achieved, for example, by looking at IVA conference papers over the next five years to find new construct specific questionnaires. Finally, it is unavoidable that the names and definitions of the current set of constructs might not always match how researchers refer to that construct. Still we hope that researchers refer to the construct set in this work as a reference point to create a common language about constructs studied in our community.

The next steps to achieve the goal of a unified community supported questionnaire are: (1) Determine an initial set of questionnaire items for the 19 constructs and their dimensions; (2) Confirmatory factor analysis to examine the validity of the constructs; (3) Establish the final item set with the provision to create a long and short questionnaire version; (4) Determine criteria validity and concurrent validity; (5) Translate the questionnaire; and (6) Develop a normative data set. We welcome more researchers to join our group-effort.

The findings presented in this paper show that the community is interested in 19 measurement constructs (according to the work group members' expert opinion), see Table 3. These 19 constructs represent more than 80% of the studies evaluating human-ASA interactions in the IVA community (between 2013-2018). A clear indication that we are indeed measuring the same things.

REFERENCES

[1] Icek Ajzen and Martin Fishbein. 1980. *Understanding attitudes and predicting social behavior.* Englewood Cliffs, NJ: Prentice Hall.
[2] M. Dragone, T. Holz, B. R. Duffy, and G. M. P. O'Hare. 2005. Social Situated Agents in Virtual, Real and Mixed Reality Environments. In *Proc. of IVA'05*, Themis Panayiotopoulos, Jonathan Gratch, Ruth Aylett, Daniel Ballin, Patrick Olivier, and Thomas Rist (Eds.). Springer Berlin Heidelberg, 166–177.

[4] Examining19Constructs.xlsx (read-only), https://osf.io/c8p3e/

[3] 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.

[4] Maaike Harbers, Jeffrey M. Bradshaw, Matthew Johnson, Paul Feltovich, Karel van den Bosch, and John-Jules Meyer. 2012. Explanation in Human-Agent Teamwork. In *Coordination, Organizations, Institutions, and Norms in Agent System VII*. Springer Berlin Heidelberg, 21–37.

[5] Michael Luck and Ruth Aylett. 2000. Applying artificial intelligence to virtual reality: Intelligent virtual environments. *Applied Artificial Intelligence* 14, 1 (2000), 3–32.

APPENDIX

Table A1: Examples of constructs classified into concepts in Figure 2

Concept	Example Constructs
AProp	-*Competence*: Measuring the user's beliefs of the agent's ability to do something successfully or having sufficient knowledge/skill. -*Perceived intelligence*: Measuring user's perception of how the agent is intelligent.
ASoc	-*Perceived sociable*: Measuring the perceived ability of the system to perform sociable behaviour. -*Personality*: User's perception of the personality characteristics of the agent.
ARole	-*Perceived quality*: Measuring the user's perception of the ability of the agent performing its role. -*Perceived professionalism*: Measuring the role-appropriateness of the agent based on its attire and environmental setting.
AImpr	-*Bothered*: Measuring the user's perception of the agent's being worried, disturbed or upset. -*Perceived positiveness*: Measuring the user's perception of the agent being pleased with him/her.
IntQ	-*Bonding*: Measuring the degree of collaboration and trust between a user and the agent. -*Perceived behavioral interdependence*: The extent to which a user's behaviour affects and is affected by the agent's behaviour.
HImpr	-*Perceived enjoyment*: Measuring feelings of joy/pleasure associated by the user with the use of the system. -*Trust*: Measuring whether the user finds the agent can be relied on.
HAttr	-*Performance rating*: User's self-rating of his/her own performance. -*Self-efficacy*: Measuring the user's capability to perform task.

Refer to Table 1 for an explanation of the abbreviations of the concepts.

Table A2: List of keywords collected in resulted groups

No	OV	Keywords	Const.
1.	OV	Beliefs, human-like appearance, human-like behaviours, naturalness, user model, believable, appropriateness	1.1, 1.2, 1.3, 1.4
2.		Appearance, suitability, role	1.5
3.		System, usable	2
4.		System, free of efforts, ease of use	2
5.		Helpfulness, task performance, usefulness	3.1
6.		User performance, self-efficacy	3.2
7.		Likeable, evoking empathy/sympathy	4
8.		Social attitude, social richness	5
9.		Sociable, social behaviour	5
10.		Personality, characteristic	6.1
11.		Friendly, agreeable, caring	6.2
12.		Accepting, use intention	7
13.		Enjoy, pleasure, interacting	8
14.		Feeling connected, engaging	9
15.	OV	Involvement, sustained attention	9
16.	OV	Agent's attributes, performance, quality, trustworthiness, competence, reliability, credibility, goal-oriented, cooperation, working alliance, social relationship, rapport, warmth, intimacy	3.1, 10, 11.1, 11.2
17.	OV	Awareness, attentiveness	12
18.	OV	Coherence, consistency, clarity	13
19.	OV	Deliberate, conveying, expressive	14
20.	OV	Satisfaction, interaction, evoking user's attitude, evoking user's emotion	15, 18.3, 18.4
21.	OV	Social presence, presence of an artificial being, spatial presence, interactive	16
22.	OV	Social influence, self-image	17
23.	OV	Emotional intelligence, understand user's emotion, conveying emotion, adapting behaviour accordingly	18.1
24.	OV	Expressing emotion, emotion type	18.2
25.	OV	Interplay, reciprocity, interdependence	19

OV = containing overlapping construct(s); Const. = The measurement construct this group is represented in, the numbering refers to Table 3