The Artificial Social Agent Questionnaire (ASAQ) — Development and evaluation of a validated instrument for capturing human interaction experiences with artificial social agents

Fitrianie, Siska; Bruijnes, Merijn; Abdulrahman, Amal; Brinkman, Willem Paul

DOI
10.1016/j.ijhcs.2025.103482

Publication date
2025

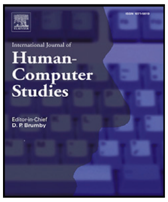
Document Version
Final published version

Published in
International Journal of Human Computer Studies

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

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



The Artificial Social Agent Questionnaire (ASAQ) — Development and evaluation of a validated instrument for capturing human interaction experiences with artificial social agents

Siska Fitrianie [a,*], Merijn Bruijnes [b], Amal Abdulrahman [a], Willem-Paul Brinkman [a]

[a] Delft University of Technology, Van Mourik Broekmanweg 6, Delft, 2628 XE, The Netherlands
[b] Utrecht University, Bijlhouwerstraat 6, Utrecht, 3511 ZC, The Netherlands

ARTICLE INFO

Keywords:
Artificial social agent
Social interactive agent
User study
Evaluation instrument
Questionnaire
Reliability
Validity
Interpretability

ABSTRACT

Validating claims and replicating findings on the impact of artificial social agents (ASA), such as virtual agents, conversational agents, and social robots, requires a standardised measurement instrument that researchers can employ in different settings and for various agents. Such an instrument would allow researchers to evaluate their agents and establish insights beyond their specific study context. Therefore, we present the long and short versions of the ASA questionnaire (ASAQ) for evaluating human-ASA interaction on 19 constructs, such as the agent's believability, sociability, and coherence. It has been developed by an international workgroup with more than 100 ASA-researchers over multiple years who identified community-relevant constructs and associated questionnaire items and examined the questionnaire's reliability, validity, and interpretability. The result is a questionnaire that can capture more than 80% of the constructs that studies in the intelligent virtual agent community investigate, with acceptable levels of reliability, content validity, construct validity, and cross-validity. We suggest that ASA-researchers use the ASAQ short version to report their agent's psychographic information and the ASAQ long version to analyse any constructs in-depth that are specifically relevant to their agent or study. Finally, this paper gives instructions for practical use, such as sample size estimations, and how to interpret and present results.

1. Introduction

Which questionnaire should I use? - A question familiar to researchers studying chatbots, intelligent virtual agents, social robots, or any *artificial social agent* (ASA). The questionnaire they select determines how convincing and useful their research results will be to other researchers. The ideal questionnaire should ensure the measurement is generalisable, reliable and widely accepted by the scientific community. In this paper, we present the ASA questionnaire (ASAQ) for evaluating *human-ASA interaction*. We describe the sizable community effort in establishing it, examine its reliability and validity, and discuss how researchers can use it.

The question of what is a good questionnaire for a study is fundamentally about the validity of research claims and, relatedly, the reproducibility of a study using its code, method and same input data, and the replicability of scientific findings across studies based on their own data (National Academies of Sciences, Engineering, and Medicine, 2019). Recent replication studies in psychological sciences (Open Science Collaboration, 2015), clinical research (Ioannidis, 2005a,b; Mobley et al., 2013; Errington et al., 2021), and economics (Camerer et al., 2016) repeatedly show the difficulty of replicating previously reported results. Researchers in other disciplines, such as chemistry, biology, physics and engineering, also reported to have tried and failed to reproduce reported findings (Baker, 2016), and reproducibility and replicability concerns are also raised in the area of computer science in general (e.g. Moraila et al. (2014)) and human–robot interaction in particular (e.g., Leichtmann et al. (2022), Gunes et al. (2022)). To address the potential causes encouraging remedies are suggested, such as confirmatory tests, large sample sizes, preregistration, methodological transparency (e.g., Protzko et al. (2023)), and authentication of study material for replicating complex experiments (Li et al., 2015). While discussing each of these (and potentially more) replication issues and remedying actions is beyond the scope of this article, it is clear that the replication crisis needs our attention. The latter remedying suggestion to have study material available for replication comes from biomedical research, but it also applies to human-ASA interaction studies where the agent is the study material. Without standardising the study material, supporting, opposing, or null findings are difficult, if not impossible,

[*] Corresponding author.
E-mail address: s.fitrianie@tudelft.nl (S. Fitrianie).

https://doi.org/10.1016/j.ijhcs.2025.103482
Received 11 October 2024; Received in revised form 19 February 2025; Accepted 19 February 2025
Available online 1 March 2025

to interpret as the cause for the observed effect may differ, making a replication study like comparing apples to oranges.

An initial response to solving this is to make ASAs available to other researchers, for example, by making available their code or the runtime version. ASAs, also referred to as *Socially Interactive Agents* (Lugrin et al., 2021), are computer-controlled entities that can autonomously interact with humans in a manner that can be regarded from the perspective of social rules of human-human interaction (Fitrianie et al., 2019). Unlike traditional machines or tools, these agents often interact, learn from, and adapt to human users in ways that simulate human social behaviours. They are deployed across various fields, such as healthcare, education, customer service, and entertainment, frequently taking on roles, such as companions, assistants, or educators. They are typically programmed to perceive and respond to social signals, such as facial expressions, body language, tone of voice, and the context in which interactions occur.

Despite recent advances, the solution of providing their code or runtime version still requires considerable technical knowledge of virtualisation (for example, of virtual environments such as Docker[1]) to ensure that an ASA can function in the same manner between experiments on various hardware and software configurations. This solution is a literal interpretation of the reproduction objective that can be useful when, for example, comparing a new ASA with a specific normative exemplary ASA. However, is a comparison with literal reproduction always intended? Do we want to compare with a specific ASA, or with what that ASA represents? For example, a researcher might want to compare their agent with an agent that uses a particular communication modality, e.g., auditory, symbolic, spoken communication, or even a combination. Such a comparison requires an accurate representative specification of each ASA, which comes with its own challenges, such as what is essential to describe and how to do this. Some efforts in this direction have been made over the years, for instance formalising the description of ASA behaviour generation in a framework that describes the Situation, Agent, Intention, Behaviour, and Animation (SAIBA). Here specific multi-modal ASA behaviour is defined using a standardied "behavioural markup language" (BML) (Kopp et al., 2006) and the function of each behaviour is described in a "functional markup language" (FML) (Cafaro et al., 2014; Heylen et al., 2008). However, this approach proved insufficiently rich for all future use cases. This resulted in lab-specific implementations and additions, which were foreseen and allowed under the initial specification, of the standard framework (e.g., van Welbergen et al. (2009), Cafaro et al. (2017), Kolkmeier et al. (2017), Bevacqua et al. (2010), Kipp et al. (2010), Hartholt et al. (2013), Holroyd et al. (2011)). As such, while the SAIBA framework, BML, and FML are valuable efforts, they still lead to the same issue as incompatible one-off descriptions of ASAs.

Therefore, in this work, we follow a mediated route (Fig. 1). Instead of focusing on the literal reproducing the ASA or what it represents (replicating the independent variable), we seek to replicate the impression the interaction with the ASA made on the individual (replicating the experience), such as its social presence and the human-likeness of its behaviour. This impression, we assume, causes the sought-after effect (the dependent variable). To be specific, we argue that the experience the user has is paramount to the realisation of any effects of the interaction with an agent and, crucially, these effects are not due to the specific agent. This reasoning is analogous to other settings where the experience – not the specific agent – is defining the outcome. For example, different teachers in our school system create a similar learning experience which realises the same knowledge or skills in a group of students. Another example might be the outcome effect of enjoyment that can be realised through the experience of a music concert nearly independent of which musician performs – nearly, because some musicians (or teachers, or agents in our case) might not be able to create the same experience. Consequently, we argue that a successful replication of a human-ASA interaction study aims to show that, although it uses a different ASA, it creates a similar and authentic human experience that, in turn, impacts the dependent variable constituting a desired outcome such as weight reduction.

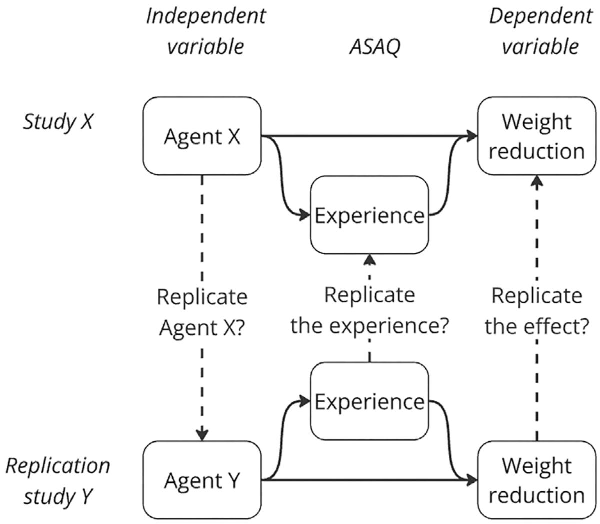


Fig. 1. The ASAQ's approach: replicating human-ASA interaction experience in the context of a study, for example, a study where an agent attempts to impact the weight reduction of a user. Rather than directly replicating Agent X, the focus is on replicating the user experience created by the agent-human interaction, which in turn influences the dependent variable (weight reduction). Study Y attempts to achieve similar outcomes using Agent Y by recreating comparable user experiences.

Thus, instead of specifying the agent, we need to specify the human experience with an agent. Considering the ISO standard for human-centred design for interactive systems (International Organisation for Standardisation, 2019), we define user experience as a person's perceptions and responses that result from the use of an artificial social agent. This includes a person's emotions, beliefs, preferences, perceptions, comfort, behaviours, and accomplishments that occur in relation to the interaction. We can specify such psychological phenomena in various ways: theory-driven, data-driven, or, as we have done, using a community-driven strategy. Let us take measuring personality as an example. The Myers-Briggs Type Indicator (MBTI) (Myers and McCaulley, 1985) questionnaire measures a person's personality according to Jungian personality characteristics (Matthews et al., 2003). An unmistakable limitation of this strategy is its dependence on the validity of the underlying theory. Therefore, Goldberg (1990) asked people to rate themselves on a large set of questionnaire items and identified groups of correlating items as evidence for a latent five-factor underlying structure to specify personality. The data-driven approach works for phenomena that are at the start of a causal chain, as correlation grouping makes no distinction whether items relate because they are indicators of the same underlying construct or because they are indicators of two distinct constructs that have a causal relationship. This would be an issue as ASA researchers tend to be interested in only some constructs, optimise their ASA for those constructs, and study them in isolation. Thus causal relationships between constructs, such as the effect an agent's enjoyability might have on accepting the agent, need to be accounted for in a measurement instrument. Therefore, we took the community's research interests as our main lead for specifying the human experience by measuring aspects we knew researchers would find relevant without grounding it in an overarching theoretical framework and accepting data dependencies between constructs.

Unfortunately, currently, no questionnaire seems to cover the vastly diverse community interest. For example, our literature survey (Fitrianie et al., 2019) of the intelligent virtual agents conference proceedings from 2013 to 2018, found that 81 studies used 89 questionnaires.

[1] https://www.docker.com/

The survey found little reuse, as 76% of questionnaires were only reported in a single paper. Even worse, the studies reported a total of 189 measured constructs, suggesting, at least on the surface, a diverse interest with little attempt to replicate the user experience. Therefore, we needed to look below the surface, see the underlying shared community interest, and capture this in a single standardised questionnaire.

In addition to what a questionnaire measures, researchers looking to select a questionnaire for their study should also consider the research underlying the questionnaire's development. Good methodological quality makes the questionnaire's findings more trustworthy. Several appraisal tools (Rosenkoetter and Tate, 2018) help researchers review the quality of an instrument by considering basic psychometric properties, such as reliability and validity, and the guidelines it provides for reporting findings. We were specifically inspired by the COSMIN initiative (Mokkink et al., 2010), which aims to improve the selection of health measurement instruments. As a community, they put forward reliability, validity, responsiveness and interpretability as quality criteria for a measurement instrument. Therefore, we set out to conduct a series of studies that, besides defining the questionnaire, examine these quality aspects, with the exception of responsiveness. This refers to the ability to detect change over time in a person, i.e., the ability to measure how an individual's illness changes over time, for example, after a treatment (de Vet et al., 2011). Responsiveness is therefore important for a health context, but less so for a human-agent interaction experience context as it is related to the context-dependent outcome variable that we consider out of scope for this questionnaire. Furthermore, we also provide guidelines for the number of participants needed in a study and for reporting the results of the ASAQ. In this paper, we first describe the questionnaire (Section 2) and then dive into the development and subsequent validation steps (Sections 3–4). We end with usage instructions, power analysis and a discussion of the implications of this work (Sections 5–6).

2. The ASA questionnaire

In this section, we introduce the ASAQ. This instrument is developed through a comprehensive creation and validation process, which we describe in detail in the following sections. The ASAQ consists of 19 constructs that represent specific phenomena or aspects we aim to measure. Three of these constructs are split up into a total of eleven dimensions (Table 1). Each construct and dimension is measured by questionnaire items (i.e., statements). Participants rate the extent to which they agree or disagree with these statements on seven-point scales ranging from 'disagree' (value of −3) to 'agree' (value of 3) with the middle point (value of 0) for 'neither agree nor disagree'.

The ASAQ set-up is based on Classical Test Theory (CCT) (Lord and Novick, 2008), in which we gain an understanding of unobservable constructs by measuring observable items that we regard as manifestations of these constructs. In other words, constructs are reflected in the items. Therefore ASAQ can also be referred to as using a reflective model (e.g., Hair et al. (2021)) to describe the relationship between items and construct.

Although the initial ASAQ defines eleven dimensions that are regarded as important to the ASA community, three dimensions are not included in the validation process of ASAQ namely: Agent's Personality Type (6.2), Agent's Emotional Intelligence Type (18.2), and User's Intelligence Type (18.4). The reason for excluding these three dimensions is the existence of well-established specialised questionnaires to measure them (e.g. OCEAN (Goldberg, 1990), Universal Emotions (Ekman, 1999), Valence-Arousal Circumplex Model (Feldman-Barrett and Russell, 1998), Self-Assessment Manikin (Bradley and Lang, 1994), and Geneva Emotion Wheel (Scherer, 2005)). Consequently, that would leave the ASAQ with eight dimensions, with the provision that the community *is* interested in these three additional dimensions. However, excluding the Agent's Personality Type (6.2) dimension left the Agent Personality (6) construct with a single dimension namely, Agent's Personality Presence (6.1). As a result, the ASAQ consists of 19 constructs, with three of these constructs (1, 6, and 18) being measured by eight dimensions, resulting in 24 groups of questionnaire items (the non-italics constructs or dimensions in Table 1).

The ASAQ offers two versions, a long and a short version. The long version consists of 90 items distributed across the 24 constructs and dimensions, with about four items dedicated to measuring each construct or dimension Appendix A. The long version allows researchers to establish a comprehensive evaluation of the human-ASA interaction. In contrast, the short version of the ASAQ includes 24 items where every construct and dimension from the long version is represented by only one item Appendix B. This short version allows researchers to establish a general impression of the human-ASA interaction quickly.

Furthermore, the ASAQ facilitates measuring the experience of people who interacted with an agent themselves (i.e., first-person perspective), or of people who have observed someone else interacting with an agent (i.e., third-person perspective). Take, for example, the item "[I/The user] can rely on [the agent]" from the User's Trust construct. From the first-person perspective, the statement is "*I can rely on...*" and from the third-person perspective, the statement is "*The user can rely on...*". Moreover, instead of referring to the abstract concept "[the agent]", researchers are expected to replace this part in each item with the name of the agent under evaluation. To broaden applicability to *any* ASA, items in ASAQ do not refer to any particular physical part, modality, task, or functionality of an agent. Also, items that are reversed, are indicated by a [*R*]. These scores must be reversed (multiplied by -1) before taking the mean item score for the construct or dimension. Likewise, items indicated with [*R*] in the short version, also need to be reversed like this.

Finally, currently, translations of the ASAQ to several languages are available. Notably, the validated Dutch, German (Albers et al., 2024) and Chinese (Li et al., 2023) versions of the ASAQ have been developed, and online available.[2]

3. Creation of the questionnaire

The development of the ASAQ has been a long and ongoing process in which many ($n = 120+$) people from the ASA community participated. For this we are grateful and their efforts should be acknowledged. In particular, our ASA workgroup, organised on the Open Science Foundation's platform, contributed on many occasions to this work.[3] When in the next sections we speak of experts or judges we refer to members of this group. Finally, data and analysis code of the results presented in this paper are available online (Fitrianie et al., 2025).

3.1. Constructs and dimensions

First, we had to identify which constructs are important to the ASA community by reviewing the literature. For this, it was essential to determine the scope of the ASAQ (Fig. 2). We considered only the experience of human-ASA interaction in scope, whereas what precedes or follows the interaction is out of scope. Preceding the interaction are external entities such as an individual's personality, gender, age, economic status, or the agent's embodiment or interaction method. Although potentially relevant to an interaction, these entities are defined before the interaction, often static during the interaction, and usually more objective. Following the interaction, external entities are linked to process and outcome measures in a specific context. These context-dependent entities are why we design or use the agent, for example, to promote education, health, or entertainment. For instance, in an

[2] See the ASAQ project website for updates: https://asaq.ewi.tudelft.nl

[3] See the OSF ASA workgroup contributors and efforts here: https://osf.io/6duf7/

Table 1
19 constructs of the ASAQ.

No.	Construct/Dimension	Definition
1	*Agent 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)
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 is formed
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

Note: The numbering following <construct no>.<dimension no>. In *italics* are the constructs and dimensions that are not (or not directly) measured.

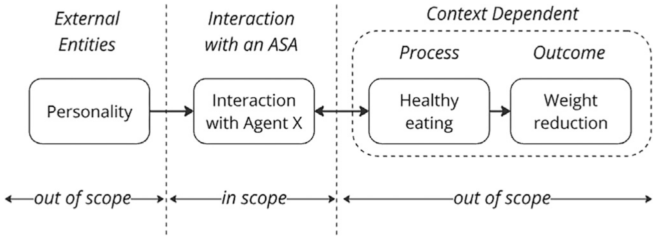


Fig. 2. The scope of the ASAQ: "(human) interaction with an artificial social agent". Out of scope are the more or less static entities that exist before and after the interaction (e.g., personality) and context-dependent process (e.g., healthy eating) and outcome (e.g., weight reduction) measures that are impacted by the interaction.

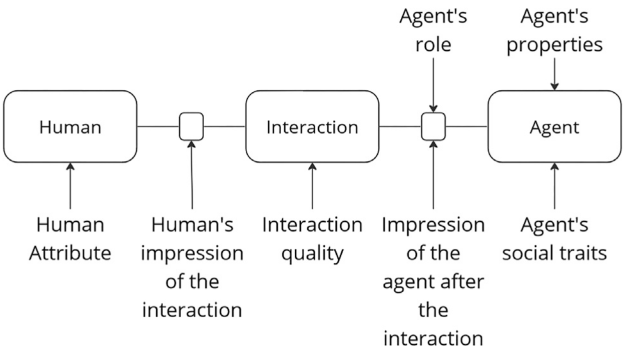


Fig. 3. The relation between the core elements of an ASA-human interaction and the anchor points.

academic learning context, the process measure is the daily hours of studying and the outcome measure is the exam mark at the end of the course. Researchers often construct tailor-made questionnaires, rely on established measurement instruments, or use objective measures for these out-of-scope constructs.

Thus, constructs are relevant to the ASAQ when they represent the fundamental aspects of human-ASAs interaction experiences that are commonly studied by researchers in the ASA community. Therefore, we set out by looking for shared interest in the questionnaire constructs reported over six years (2013–2018) of one of the leading artificial social agent conferences; the international Intelligent Virtual Agent conference (IVA) (Fitrianie et al., 2019, 2020a).

This resulted in 189 identified constructs, of which many appear closely related. Take, for example, the constructs: Bonding (Jaques et al., 2016), Friendship (Grigore et al., 2016), Feeling of Closeness (Pejsa et al., 2017), Perceived Rapport (Zhao et al., 2018), and Intimacy (Bickmore et al., 2013; Kolkmeier et al., 2016). These constructs appear so closely related in their name and description that they might warrant combining them into a single unifying construct. Many such combinations seemed possible, however, performing a direct grouping exercise of all 189 constructs was too cognitively challenging. Consequently, we followed a divide-and-conquer approach, whereby we first split the constructs into large groups and, afterwards, into more detailed groups.

For our first divide-and-conquer step, we identified seven anchor points that experts could use during the large grouping steps (Tabel 2). These anchor points are related to the three core elements in any human-agent interaction: the human, the agent and the interaction, and the interaction between these core elements (Fig. 3 and Fitrianie et al. (2020a) for more details). Our ASA workgroup members (n = 49) volunteered to associate the 189 constructs to one of the seven anchor points or, as a quality check, to one of three out-of-scope anchors (i.e., external entities such as personality, and context-dependent entities such as healthy eating and weight reduction, see Fig. 2). The experts could independently assess as many constructs as they liked. Twelve experts assessed all constructs. On average, each construct was assessed by 13 experts, with a minimum of 12 assessments per construct.

Table 2
The number of constructs classified into the single anchor points and anchor pairs.

Core element	Anchor point	Single Anchor	A pair of Anchors
Agent	Agent's properties	4	24
	Agent's social traits	27	30
Agent-Interaction	Agent's role	9	23
	Impression of the agent after the interaction	1	2
Interaction	Interaction quality	12	32
Human-Interaction	Human's impression of the interaction	23	49
Human	Human's attributes	2	6

We used this expert data to determine the initial estimated grouping of constructs. A majority rule criterion was followed where a construct is associated with an anchor if at least 50% of the experts agreed on this association. Table 2 shows that 89 constructs (47% of the constructs) were associated with a single anchor. An additional 99 (52%) constructs required the combination of the votes for two anchor points, which we call an "anchor pair", to result in a majority. We found no majority for anchor or anchor pair association for the construct "Perceived Similarity" (Diehl et al., 2017) and as such we considered this construct as *unclassified*. Finally, the experts classified 11 constructs as out of scope (e.g., (the user's) Preferred Role in Decision Making (Zhang and Bickmore, 2018)). This left us with 177 (93.7%) for detailed grouping (189 starting - 1 unclassified - 11 out of scope = 177).

For the second divide-and-conquer step (Fitrianie et al., 2020a), the detailed grouping, we asked experts to use a card-sorting method (Nielsen, 2018). We created groups of constructs associated with each single anchor and the constructs associated with that anchor's pair(s). This resulted in seven card-sorting tasks, related to the seven anchor points, in which 23 experts independently participated. Experts could participate in as many card-sorting tasks as they liked. Twelve (52.2%) experts decided to participate in all tasks and on average each task was completed by 17 experts, with a minimum of 12 experts for each task. Finally, experts suggested a name for each group of constructs they created in each card-sorting task.

In analysing the card-sorting task (Fitrianie et al., 2020a), we followed a majority rule approach to group constructs, resulting in 52 groups. This means that constructs were included in a group only when a majority had placed these constructs together. These 52 groups included 152 constructs, representing a coverage of 86% of the total set of 177 considered constructs. Unsurprisingly, many ($n = 39$) groups exhibited overlapping constructs with another group. The overlap was expected as we deliberately assigned constructs associated with an anchor pair to two card-sorting tasks. This strategy reduced the potential bias caused by the initial – somewhat arbitrarily chosen – anchor points.

A panel of three expert judges examined the 52 card-sorted groups to combine all constructs into 19 unifying constructs. For this, they considered the constructs included in the card-sorted groups, the identified overlap between groupings, and the expert-proposed names for each group. The panel defined an initial name and description for each unified construct. In the cases of Agent's Believability (1), Agent's Personality (6), Performance (3), User-Agent Alliance (11), and Emotional Experience (18), the panel concluded the scope of the construct was too broad and divided the constructs into dimensions. Next, members of our ASA workgroup were invited to discuss the initial names and descriptions of the panel. Eight work-group members took up this invitation and finalised the names and descriptions, see Table 1. Note, however, that as a result of the construct validity study, which we will discuss later, we eventually dropped the dimensions of the constructs Performance (3) and User-Agent Alliance (11).

3.2. Questionnaire items

With the unified constructs identified, we needed to create items that could measure each construct. Our ASA workgroup members were invited to propose as many items as they wished for each construct. Eight experts put forward 431 items, on average 17 per construct or dimension. A panel of three judges checked and improved these items to address grammar or formulation issues.

To reduce these 431 items to the final 90 items included in the ASAQ we took three steps (Fitrianie et al., 2021a). First, twenty workgroup members checked the items in a content validity analysis. Four judges then selected the items these experts assessed to best measure the construct or dimension, aiming for eight items per construct/dimension. The judges kept items that experts associated with the intended construct, and removed items that experts associated with unintended constructs or with multiple constructs at the same time. When more than eight items remained, they also examined the results of a similarity test using a combination of the Word2Vec embedding (Mikolov et al., 2013), smooth inverse frequency (Arora et al., 2017), and cosine similarity methods (Sieg, 2018). With the goal of capturing the various relevant aspects of a construct or dimension, the judges selected items most dissimilar from each other, while also considering the items' semantic, lexical, and pragmatic sides. This reduced the set to 207 items. Second, we conducted a reliability analysis (Section 4.1) on data from 192 crowd-workers who used the items to rate a human-ASA interaction of the robot ASIMO (Advanced Step in Innovative Mobility by Honda) they saw in a video. Based on this analysis we removed items with 1) substantial rating differences (absolute standardised mean difference ($> Q_3 + 1.5 \times IQR$) when the item was formulated in first-person or third-person perspective, or 2) substantial correlation with other constructs ($> .50$). This reduced the set further to 131 items. Third, based on data from the construct validity study (Section 4.3), we removed a further 41 items, resulting in the final 90 items of the long version of the ASAQ. In the next sections, we describe the characteristics of these 90 items.

Additionally, we developed a short version of the ASAQ by selecting one representative item for each construct or dimension. Four judges chose these representative items based on their factor loadings, which reflect how strongly they correlate with their respective constructs, as well as their theoretical relevance to the underlying dimension (Fitrianie et al., 2022a). In total, the judges selected 24 items Appendix B to serve as the short version of the ASAQ.

4. Characteristics of the ASAQ

In this chapter, we describe the (statistical) characteristics of the ASAQ. Where possible, we combine datasets from the various validation studies that we performed during the creation of the ASAQ to efficiently reach a large sample size for the descriptives in this chapter. We describe the reliability, validity, and interpretability of the ASAQ, as these are important quality criteria for a measurement instrument (Rosenkoetter and Tate, 2018; Mokkink et al., 2010).

4.1. Reliability of the ASAQ

The ASAQ consists of multiple items, each representing a latent construct or dimension. For the questionnaire to be reliable, it is necessary for the ratings on items of one construct or dimension to correlate with each other. High correlation means that the instrument has high internal consistency.

Table 3
Reliability analysis results (Cronbach's α) of three studies.

No.	Construct/Dimension	Study Early '21 $n = 192$			Study Mid '21 $n = 532$	Study '22 $n = 534$	Combined Mid '21 & '22 $n = 1066$
		All	1st	3rd			
1.1	Human-Like Appearance	0.86	0.88	0.84	0.83	0.86	0.84
1.2	Human-Like Behaviour	0.82	0.83	0.80	0.84	0.85	0.84
1.3	Natural Appearance	0.70	0.66	0.73	0.73	0.73	0.73
1.4	Natural Behaviour	0.56	0.58	0.55	0.60	0.60	0.60
1.5	Agent's Appearance Suitability	0.77	0.74	0.80	0.73	0.72	0.72
2	Agent's Usability	0.80	0.77	0.83	0.72	0.81	0.77
3	Performance	0.53	0.72	0.30	0.68	0.64	0.66
4	Agent's Likeability	0.84	0.84	0.84	0.80	0.80	0.80
5	Agent's Sociability	0.56	0.66	0.47	0.71	0.66	0.69
6.1	Agent's Personality Presence	0.55	0.38	0.64	0.60	0.61	0.60
7	User Acceptance of the Agent	0.65	0.74	0.51	0.65	0.67	0.66
8	Agent's Enjoyability	0.76	0.79	0.74	0.70	0.77	0.74
9	User's Engagement	0.65	0.88	0.84	0.70	0.65	0.68
10	User's Trust	0.62	0.83	0.80	0.66	0.67	0.67
11	User Agent Alliance	0.72	0.66	0.73	0.77	0.78	0.77
12	Agent's Attentiveness	0.79	0.58	0.55	0.72	0.70	0.71
13	Agent's Coherence	0.75	0.74	0.80	0.67	0.69	0.68
14	Agent's Intentionality	0.70	0.77	0.83	0.68	0.73	0.70
15	Attitude	0.74	0.72	0.30	0.78	0.81	0.79
16	Social Presence	0.71	0.84	0.84	0.71	0.62	0.67
17	Interaction Impact on Self-Image	0.79	0.66	0.47	0.72	0.73	0.73
18.1	Agent's Emotional Intelligence Presence	0.87	0.38	0.64	0.86	0.85	0.86
18.3	User's Emotion Presence	0.81	0.74	0.51	0.64	0.69	0.66
19	User Agent Interplay	0.70	0.79	0.74	0.67	0.61	0.64

Note: The columns refer to data that originates from three studies: Study Early '21 originates from the initial Reliability Study, Study Mid '21 from the Construct Validation study, and Study '22 from the Cross-Validation Study. The Combined Data column refers to the combination of the Mid '21 and '22 datasets.

To this end, we collected data for a reliability analysis from 192 crowd-workers from the platform Prolific Academic on 11-02-2021 using Qualtrics as the questionnaire platform. We asked the participants to use the questionnaire items, which we offered in a random order, to rate the human interacting with the robot ASIMO based on a 30-second video. We randomly assigned half of the participants to a questionnaire with items formulated in the third-person perspective and the other half to one with items formulated in the first-person perspective. In the latter, we asked participants to rate the items from the perspective of the person in the video who was interacting with the ASA. For the analysis, we included only participants who passed at least 12 of the 15 attention-check questions we randomly added to the questionnaire. This study was approved by the Human Research Ethics Committee TUDelft (no. 1402, date 18-12-2020) and preregistered (Fitrianie et al., 2021b). For more details about this study, see Fitrianie et al. (2021a).

To determine the reliability we calculated Cronbach's α for each ASAQ construct and dimension. For this paper we report the combined data from three studies: the reliability study in early 2021 and two follow-up studies (aimed at construct- and cross-validation). We combine the data of these last two to efficiently reach a large n, see Table 3. The combined studies have values ranging from .60 to .86, with an average of .72, which is classified as a respectable reliability (DeVellis and Thorpe, 2021). The differences in reliability between the first-person and third-person perspectives are not substantial, see Table 3 (column Study Early '21). The mean absolute difference between the two perspectives is small ($M = .12$, $SD = .12$) and the difference between perspectives is stable across the constructs and dimensions, except for Agent's Personality (6.1) and Agent's Emotional Intelligence Presence (18.1). We speculate that for these constructs some items might not reliably transfer between perspectives, for instance, observing someone having a reaction might not convince participants that 'they would also have that reaction' despite that in reality they might react the same – a bias similar the Dunning–Kruger effect (Kruger and Dunning, 1999).

4.2. Content validity of the ASAQ

To determine whether the ASAQ items adequately reflect their intended constructs (i.e., content validity (Fitrianie et al., 2021a)), we were inspired by the approach put forward by Lawshe (1975). This involves asking experts to independently assess whether or not an item effectively measures the construct for which we had initially proposed it.

We broke down this assessment into a series of small tasks for each construct and dimension. Each task showed the definition of a construct or dimension, and four items in a random order. These items were randomly selected such that two items were written for that construct or dimension (target items), and two were written for another construct or dimension (distractor items). The expert's task was to identify which two items would adequately measure the construct and which would not. In total, 20 experts from our workgroup volunteered to participate, with an average of 10 experts per task, ranging from 8 to 15 experts. The study was approved by the Human Research Ethics Committee TUDelft (no. 1402, date 18-12-2020), and preregistered (Fitrianie et al., 2020b).

For each item, we calculated the standardised, chance-corrected ($p = .50$, i.e., 2 out of 4) True Positive Rate (TPR_s), defined as:

$$TPR_s = \frac{TP - p(TP + FN)}{TP + FN} + p,$$

where True Positive (TP) stands for the times an item is intended and identified as a target, and False Negative (FN) stands for the times an item is intended as a target but identified as a distractor. In theory, TPR_s can range from 0 to 1. We define an item's TPR_s value above .70 as an acceptable level of correctly identifying as measuring a construct. To give this some intuition consider ten experts; a target item where 8 out of 10 experts agree that it measures the construct is acceptable, while 7 out of 10 is not. Next, we calculated the standardised, chance-corrected False Positive Rate (FPR_s), defined as:

Table 4
The average TPR_s and FPR_s for each ASAQ construct or dimension.

No.	Construct/Dimension	TPR_s Mean (SD)	TPR_s Mean (SD)
1.1	Human-Like Appearance	0.96 (0.05)	0.06 (0.08)
1.2	Human-Like Behaviour	0.98 (0.04)	0.21 (0.07)
1.3	Natural Appearance	0.92 (0.05)	0.08 (0.08)
1.4	Natural Behaviour	0.97 (0.05)	0.22 (0.03)
1.5	Agent's Appearance Suitability	0.93 (0.12)	0.07 (0.06)
2	Agent's Usability	0.94 (0.05)	0.03 (0.06)
3	Performance	0.91 (0.09)	0.13 (0.12)
4	Agent's Likeability	0.81 (0.05)	0.11 (0.07)
5	Agent's Sociability	0.95 (0.05)	0.13 (0.12)
6.1	Agent's Personality Presence	0.97 (0.05)	0.14 (0.10)
7	User Acceptance of the Agent	0.96 (0.07)	0.08 (0.08)
8	Agent's Enjoyability	0.96 (0.05)	0.17 (0.05)
9	User's Engagement	0.89 (0.11)	0.12 (0.06)
10	User's Trust	0.97 (0.06)	0.04 (0.07)
11	User Agent Alliance	0.97 (0.05)	0.19 (0.07)
12	Agent's Attentiveness	1.00 (0.00)	0.20 (0.05)
13	Agent's Coherence	0.96 (0.05)	0.14 (0.10)
14	Agent's Intentionality	0.95 (0.06)	0.10 (0.07)
15	Attitude	1.00 (0.00)	0.15 (0.09)
16	Social Presence	0.93 (0.06)	0.25 (0.04)
17	Interaction Impact on Self-Image	0.95 (0.06)	0.00 (0.00)
18.1	Agent's Emotional Intelligence Presence	1.00 (0.00)	0.10 (0.07)
18.3	User's Emotion Presence	0.98 (0.04)	0.11 (0.04)
19	User Agent Interplay	0.95 (0.05)	0.04 (0.07)

$$FPR_s = \frac{FP - p(TN + FP)}{TN + FP} + p,$$

whereby False Positive (FP) stands for the times an item is intended as a distractor but identified as a target, and True Negative (TN) stands for the times an item is intended and identified as a distractor. Likewise, FPR_s ranges from 0 to 1. We took as a threshold for unacceptable confusion $FPR_s > .30$, as this is an indication that an item was too often associated with an unintended construct. This means that 7 out of 10 experts correctly identifying an item as a distractor is acceptable, while less is not.

Table 4 shows the average TPR_s and FPR_s of the ASAQ constructs. The TPR_s ranges from .81 to 1, with an average of .95 ($SD = .04$), while FPR_s ranges from 0 to .25, with an average of .12 ($SD = .06$). This can be interpreted as that a substantial majority of experts agree that each of the 90 ASAQ items measures the intended construct or dimension and not unintended constructs or dimensions.

4.3. Construct validity of the ASAQ

As ASAQ includes multiple constructs and dimensions, we examined whether the items operate consistently (i.e., construct validity). More specifically, we examined whether an item converges with the items of the same construct or dimension (i.e., convergent validity) and diverges from items of other constructs or dimensions (i.e., discriminant validity) (Lawrence, 2014). In other words, are the relationships between the ASAQ items' scores consistent with the hypothesised construct and dimension structure?

As we wanted our results to generalise across agents,[4] we used videos of different types of agents. The videos demonstrated an interaction between an agent and a human, for example, a user asking an agent for help with some task. In gathering this stimulus set, we set out to select agents keeping three aims in mind: (i) to ensure that all constructs/dimensions were relevant in at least some of the interactions; (ii) to ensure some variation in the ratings among the items observed within each construct/dimension; and (iii) to maintain some degree of independence between the ratings of items from different constructs/dimensions. Nine experts from the work group volunteered to collect the stimulus set, resulting in video clips of 56 different agents. The agents vary in the physical type (e.g., robots, chatbots, voice assistants, virtual agents, and real animals), application domain (e.g., education, healthcare, personal assistant, and entertainment), interaction environment (i.e., reality, mixed reality, virtual reality, and augmented reality), and production stages (i.e., high or low fidelity prototypes, partially or fully functional systems). To ensure that the stimuli would elicit a wide range of ratings across the different constructs/dimensions, three experts examined all agents and predicted the ratings of each agent on each construct as high, medium, or low. Using these predicted ratings, we selected 14 agents Appendix C with the lowest correlation and the biggest predicted spread of ratings.

To evaluate the construct validity, we collected ASAQ data from 532 crowd-workers on the online crowd-sourcing platform Prolific from 12–14 July 2021. Participants rated the interaction between an individual and an agent shown in a 30-second video from a third-person perspective. The Human Research Ethics Committee TUDelft (no. 1402, date 18-12-2020) approved the study, which we also preregistered (Fitrianie et al., 2021c). The videos used are available[5] and for a more detailed description of the study, see Fitrianie et al. (2022a,b).

Each participant evaluated one randomly selected agent using the questionnaire items, which were presented in a random order. As we conducted this study during the development of the ASAQ, we were still working with 131 questionnaire items. Thus, for the sample size estimation, the 4 to 10 participants per item rule-of-thumb for factor analysis (de Vet et al., 2005) suggests a minimum sample size of (131 items × 4) 524 participants. We increased this to 532 to have an equal distribution of participants across the 14 selected agents (14 × 38 = 532). We included 15 attention-check questions (pass ≥ 12 correct) and a check on whether people could watch the video. Of the 567 recruited participants, 33 failed the video-compatibility check, and two failed the attention check.

[4] Note that here we use the term 'agent' and not ASA, as some videos include an interaction with a 'natural' social agent, such as a dog. However, most videos show an interaction with an artificial social agent, such as a social robot.

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

To examine the convergent validity, we conducted a Confirmatory Factor Analysis (CFA) on each construct and its dimensions separately[6]. The models had Comparative Fit Index (CFI) scores ranging from .96 to 1 (CFI $M = .99$, $SD = .02$), where a score close to one indicates a good fit between data and model (Blunch, 2013).

To examine discriminant validity we conducted CFA on the items of multiple constructs at the same time. However, we failed to run an admissible second-order model based on a single conceptual model that included all ASAQ constructs and dimensions. We ended up with a negative variance and a non-positive definite matrix, as our model was likely too complex (Blunch, 2013). Therefore, we instead ran four CFA with first-order models, grouping the highest correlated constructs and dimensions. This way, we aimed to maximise the likelihood of spotting discriminant validity weaknesses. To determine the grouping, we first ran an Exploratory Factor Analysis (EFA) on the predicted latent score of constructs we obtained in CFA during the convergent validity analysis. The CFI score of these four models ranged from .95 to .98 ($M = .96$, $SD = .01$), suggesting that the ASAQ model fitted well with the collected data.

4.4. Cross-validity of the ASAQ

In the description of the ASAQ characteristics so far, we have used data that was also used to create or modify the ASAQ. This might lead to quality performance scores suffering from overfitting as we merged dimensions or removed items to optimise the scores based on that same data. To address this concern, we conducted a repetition of the construct validity analysis with a new sample of participants and agents. We obtained approval for this study from the Human Research Ethics Committee TUDelft (no. 1963, date 24-01-2022), preregistered the study (Fitrianie et al., 2022c), and made the analysis script and data online available (Fitrianie et al., 2025).

We recruited 544 new Prolific Academic crowd-workers (> 18 years, fluent in English). Out of the 544 participants, eight failed the video-compatibility check, one failed the attention check, and one completed the questionnaire multiple times, leaving us with a sample of 534 participants for the analysis. The data collection took place between 5–15 September 2022.

From the original set of 54 agents, we selected a new set of agents, minimising the correlation between the agents and maximising the spread and coverage of predicted ratings (Section 4.3), to ensure diverse agents. These 15 agents consisted of 13 ASAs, one zombie, and one fish Appendix C. The 24 CFAs (for 16 constructs and 8 dimensions) as part of the convergence analysis, showed good model fit with, on average, a CFI score of .99 ($SD = .02$, range [.94 .. 1]). For the discriminant validity analysis, we conducted six CFA on groups of the most related constructs and dimensions. The grouping was again based on an EFA. The constructs and dimensions that loaded highly on multiple factors were included in multiple CFAs.

In all six analyses, we only considered the relation between items and their own construct or dimension, and the relation among constructs and dimensions. In other words, we followed the assumption that the constructs and dimensions are aspects in their own right and that we did not have to include associations between items from other constructs. Here we test whether this assumption was correct. The factor loadings of the items ranged from .42 to .86 ($M = .64$, $SD = .10$) in the six CFAs, while the CFI scores ranged from .93 to .97 ($M = .94$, $SD = .02$), see Fig. 4. All CFIs are above the .90 threshold indicating a good fit of the model to the data (Stevens, 2009; Blunch, 2013). Thus, the results support the hypothesis that the constructs and dimensions are aspects in their own right, as we did not have to include associations between items and other constructs. We allow such associations on the construct-dimension level, acknowledging their interlinking.

4.5. Predictive validity of the ASAQ

We consider predictive validity a measure of the relation between expert predictions of ASA performance and the construct ratings of ASAs in the ASAQ. As outlined in Section 4.3, we selected ASAs for the construct and cross-validity study based on the predictions of three experts. They predicted for a set of 54 agents the ratings for each construct on a scale from low, medium and high.[7] We correlate the median of these expert predictions with the mean ASAQ rating of participants for the 29 agents. Table 5 shows this correlation, giving insight into ASAQ's predictive validity on a construct level. Note that these predictions were made during the construction of the ASAQ where constructs were still in flux, therefore we can only show the correlations for the ASAQ constructs for which we have predictions. Thus, excluding the constructs Performance and User Agent Alliance, the Spearman correlations for the long ASAQ version range from .10 to .92, with a median of .50. According to Hinkle (2003) we can classify this as a moderate correlation (Hinkle et al., 2003).

4.6. Concurrent validity of the long and short ASAQ versions

To determine the relation between the long and short versions of the ASAQ, we calculated the correlations and differences between these two versions using the data that we have available from the Mid '21 and '22 studies, see Table 6. Combining these datasets shows an average correlation of .81 over all constructs (range [.70 .. .92]). This can be interpreted as high concurrent validity (Hinkle et al., 2003). Likewise, the absolute mean differences between the means of the constructs and dimensions of the long version and the representative item of the short version shows an acceptable variation, with a mean of the absolute mean differences of .21 (range [0 .. .58], $SD = .17$).

4.7. Interpretability of the ASAQ

We examined the ASAQ interpretability focusing on three aspects that helped us to interpret the scores of the ASAQ. We investigated (1) the extent to which participants used the full range of a scale, (2) potential floor or ceiling effects, (3) and the availability of normative data (Mokkink et al., 2010). We again reused the data from both the construct- and cross-validity studies for this analysis. This combined dataset provides a representative set of 29 agents and 1066 ($532 + 534$) unique participant ratings.

First, we assessed the coverage of the answer-range of the scale. Table D.1 shows the relative frequency of how often the participants used each of the seven points on the answer scale. Note that the scores of reverse items were adjusted accordingly. For some constructs, such as User's Engagement (9), the score distribution appears asymmetrical, suggesting that the stimulus set was not balanced for these constructs: the range of such a construct was not covered. For example, perhaps only ASAs were included that evoke strong user engagement. Alternatively, the experiences for these constructs might be polarised. For example, participants perceived some agents as having human-like appearance while others did not, which limited the middle ground.

Second, we moved from analysing individual item-level scores to construct-level scores to investigate ceiling or floor effects. To achieve this, we calculated the average score across the items of each construct or dimension to investigate extreme scoring. Table 7 shows the percentages of participants who gave the lowest or highest possible construct or dimension score. Specifically, this lowest score means that a participant gave a −3 score on all items of a construct, whereas the highest means a +3 rating on all items. Summing these percentages gives the percentage of extreme construct scores, see the last column

[6] For the construct validity analyses we used R (v4.0.4) with factor analysis libraries from the packages psych (v2.1.3) and lavaan (v0.6–8), and we used the maximum likelihood method with promax rotation, which allows for factors to be correlated.

[7] Dataset pertaining to 54 agents on the OSF: https://osf.io/q2xur/wiki/home/

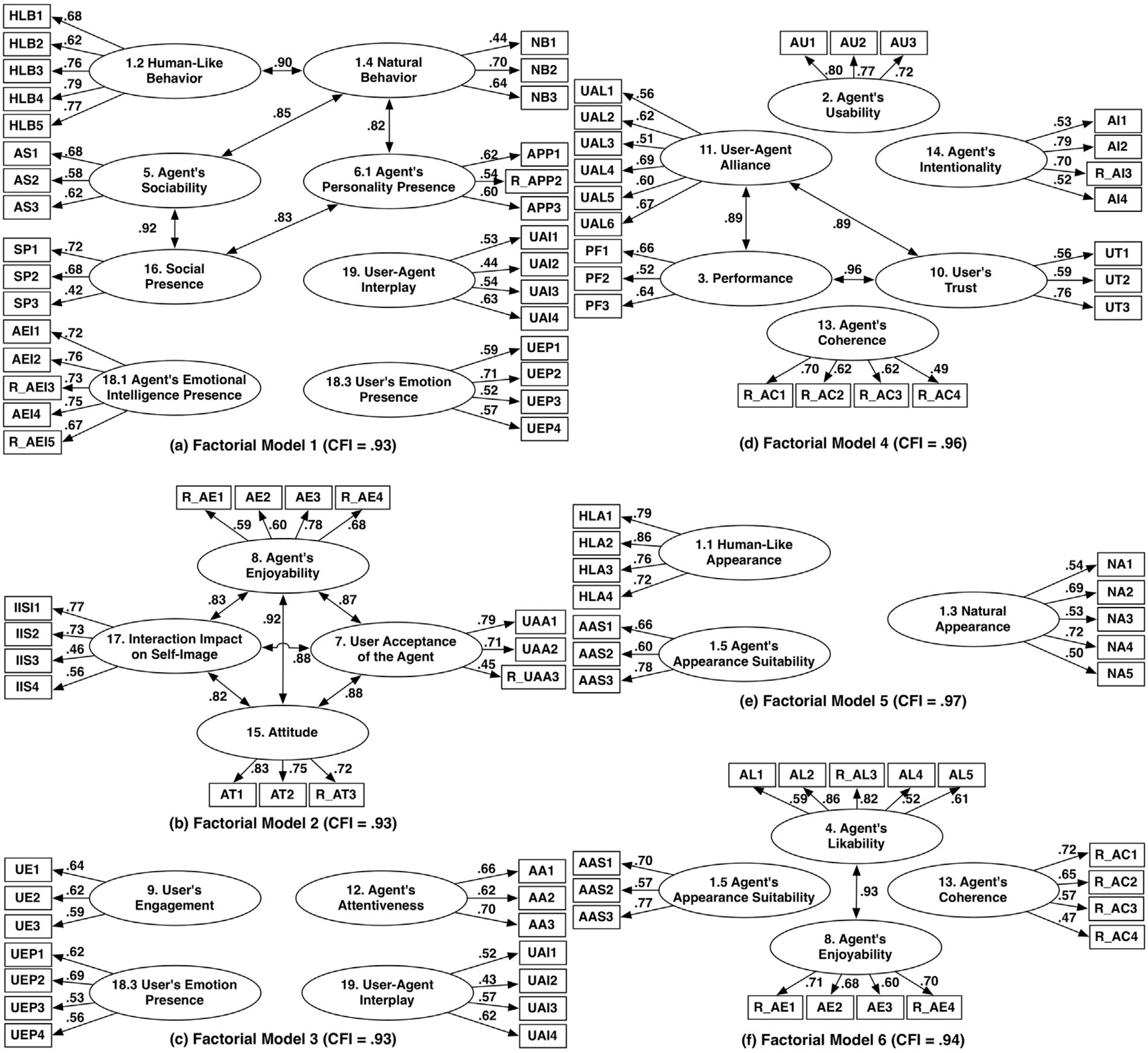


Fig. 4. Confirmatory factor analysis diagrams. Links between constructs that are of marginal and moderate concern $\rho \geq .80$ (Rönkkö and Cho, 2022) are shown.

in the table. The percentage ranges from 2% to 28% with an average of 8.50%, which is lower than 15%, a cut-off point suggested to indicate a floor or ceiling effect (de Vet et al., 2011). These results suggest that, on average, participants can fully express their experience in the ASAQ, unlimited by the scale ends. Additionally, for the individual constructs and dimensions, most percentages for extreme scores are also below or around 15%, except for Human-like appearance (1.1). The relative ease of rating this construct might explain the larger extremes, specifically it might be somewhat binary whether or not something looks like a human, resulting in more pronounced and extreme ratings.

Further, we conducted a similar analysis for the short ASAQ, see Table D.2. However, with single-item constructs, there is less room for nuance and extreme scoring can be expected. Consequently, this is reflected in an average of (mean of score −3 (10%) + score 3 (19.21%) =) 29.21% extreme scores for the short ASAQ. Consequently, the short version is less capable of capturing extreme experiences because of potential floor and ceiling effects.

However, to investigate the impact of potential floor and ceiling effects on the total ASAQ performance, we computed the range of sum of the item scores for both the long and short versions. Where, in theory, the total score of the long version can range from −270 to 270, the scores of our 1066 participants range from −225 to 255 ($M = 50.50$, $SD = 76.76$), ruling out floor or ceiling effects. Similarly, for the short version of the ASAQ, the sum of item scores can range from −72 to 72, while the score of our participants ranges from −67 to 69 ($M = 15.75$, $SD = 22.15$). Furthermore, the bell-shaped curves suggest that there are no holes in the distribution of the sum of item scores for both the long and short versions, see Fig. 5.

Third and finally, we explored how to assign meaning to ASAQ results. As the ASAQ is a set of statements on which participants rate their agreement, it is not self-evident how to interpret a score. Fortunately, we can again use our representative dataset to ground and interpret future ASAQ results. To this end, we present our normative set and name it *ASAQ representative set 2024*.

One strategy future researchers could follow is to report agents that received similar scores as their ASA to give some context. For this, they can use Tables D.3 to D.10 in Appendix D, which show each agent's mean scores and standard deviations for each construct and dimension. For example, if we have an ASA with a mean Agent's Usability (2) of .5, we could report that from the representative agent set, agents Dog (as a pet) and DeepBlue, with scores of .42 and .8, received a relatively similar score as our ASA.

Another strategy for researchers is to report how well their ASA performs on a particular construct compared to the entire normative agent set. For this purpose, Tables D.11 and D.12 show the percentile score for each construct and dimension based on the long and short

Table 5
The correlation (ρ) between the median of expert predictions and the mean ratings obtained with ASAQ of 29 agents.

No.	Construct/Dimension	ρ Prediction and Long ASAQ Version	ρ Prediction and Short ASAQ Version
1.1	Human-Like Appearance	0.92	0.91
1.2	Human-Like Behaviour	0.58	0.57
1.3	Natural Appearance	0.74	0.72
1.4	Natural Behaviour	0.50	0.48
1.5	Agent's Appearance Suitability	0.58	0.52
2	Agent's Usability	0.35	0.45
3	Performance		
4	Agent's Likeability	0.64	0.66
5	Agent's Sociability	0.50	0.54
6.1	Agent's Personality Presence	0.61	0.38
7	User Acceptance of the Agent	0.41	0.32
8	Agent's Enjoyability	0.67	0.46
9	User's Engagement	0.41	0.43
10	User's Trust	0.32	0.34
11	User Agent Alliance		
12	Agent's Attentiveness	0.39	0.54
13	Agent's Coherence	0.10	0.09
14	Agent's Intentionality	0.53	0.34
15	Attitude	0.45	0.49
16	Social Presence	0.71	0.74
17	Interaction Impact on Self-Image	0.25	0.18
18.1	Agent's Emotional Intelligence Presence	0.41	0.46
18.3	User's Emotion Presence	0.51	0.46
19	User Agent Interplay	0.47	0.50

Table 6
Correlation (ρ) and absolute standardised mean difference (Δ M) between the long and short versions of ASAQ.

No.	Construct/Dimension	Study Mid '21		Study '22		Combined Mid '21 & '22	
		$n = 532$		$n = 534$		$n = 1066$	
		ρ	ΔM	ρ	ΔM	ρ	ΔM
1.1	Human-Like Appearance	0.92	0.03	0.93	0.03	0.92	0.03
1.2	Human-Like Behaviour	0.87	0.15	0.82	0.21	0.85	0.18
1.3	Natural Appearance	0.83	0.09	0.82	0.02	0.82	0.05
1.4	Natural Behaviour	0.85	0.61	0.85	0.53	0.85	0.57
1.5	Agent's Appearance Suitability	0.85	0.08	0.85	0.00	0.85	0.04
2	Agent's Usability	0.83	0.11	0.85	0.20	0.84	0.16
3	Performance	0.78	0.13	0.73	0.35	0.75	0.24
4	Agent's Likeability	0.87	0.10	0.86	0.19	0.87	0.15
5	Agent's Sociability	0.86	0.53	0.84	0.64	0.85	0.58
6.1	Agent's Personality Presence	0.74	0.39	0.79	0.32	0.77	0.35
7	User Acceptance of the Agent	0.80	0.05	0.82	0.12	0.81	0.09
8	Agent's Enjoyability	0.80	0.48	0.77	0.10	0.78	0.29
9	User's Engagement	0.83	0.07	0.72	0.05	0.77	0.01
10	User's Trust	0.82	0.09	0.80	0.16	0.81	0.13
11	User Agent Alliance	0.69	0.43	0.71	0.56	0.70	0.50
12	Agent's Attentiveness	0.79	0.19	0.76	0.18	0.78	0.18
13	Agent's Coherence	0.73	0.01	0.79	0.04	0.76	0.03
14	Agent's Intentionality	0.75	0.25	0.76	0.35	0.76	0.30
15	Attitude	0.88	0.00	0.90	0.00	0.89	0.00
16	Social Presence	0.85	0.07	0.81	0.05	0.83	0.06
17	Interaction Impact on Self-Image	0.77	0.24	0.77	0.26	0.77	0.25
18.1	Agent's Emotional Intelligence Presence	0.86	0.38	0.83	0.24	0.85	0.31
18.3	User's Emotion Presence	0.74	0.07	0.77	0.10	0.76	0.08
19	User Agent Interplay	0.78	0.36	0.73	0.37	0.76	0.36

version of the ASAQ, a scale score where k% of the 29 agents had a lower mean score. Future researchers, therefore, can report the fraction of representative agents with a lower score than their ASA. This is the percentile rank of their score. For example, the ASA Furby from our ASAQ representative set 2024 has a score of 1.92 on the Agent's Usability (2) construct (based on the long version of the ASAQ, Table D.4). This is larger than 1.76 (80th percentile) and smaller than 1.94 (90th percentile) (Table D.11), meaning it has a mean score in the top 80%–90% of the Usability ratings of the entire ASAQ representative set 2024. We propose to interpret such percentile rank scores, following Tran et al. (2024), as is shown in Table 8. This means that Furby's Usability score can be classified as "high" and reported as in the top 80%–90% scores of the ASAQ representative set 2024.

In the same way, Table 9 also allow us to report the percentile rank of the ASAQ score. The ASAQ score is calculated by adding up the mean score of the constructs and dimensions. For the short version, this means simply adding all up items. Therefore, in the future we would calculate the mean ASAQ score of our ASA and use the table to look up its percentile rank. For example, the character Samantha from the movie Her has an ASAQ score of 25 on the long ASAQ version (Table D.6). This means this character has a percentile rank of above 90%, which can be classified as "very high". In other words, it 'beats' at least 90% of ASAs in the ASAQ representative set 2024.

Besides the interpretation of the absolute scores, the data from the 29 agents also allow us to interpret the differences in the scores we obtained between two different ASAs, specifically, the size of the (absolute) differences. Tables D.13 and D.14 show the percentile difference scores observed between the mean scores of all combinations of the 29 agents for each construct and dimension based on the short and long versions of the ASAQ. For example, take an ASA A with a score

Table 7
The relative frequency (RF) of the lowest, highest and sum of extreme scores for each ASAQ construct and dimension, based on the long ASAQ version.

		Number	RF (n = 1066)		Sum RF
No.	Construct/Dimension	of Items	Lowest	Highest	Extremes
1.1	Human-Like Appearance	4	0.23	0.05	0.28
1.2	Human-Like Behaviour	5	0.03	0.02	0.05
1.3	Natural Appearance	5	0.02	0.01	0.03
1.4	Natural Behaviour	3	0.06	0.02	0.08
1.5	Agent's Appearance Suitability	3	0.00	0.12	0.12
2	Agent's Usability	3	0.01	0.11	0.12
3	Performance	3	0.00	0.06	0.06
4	Agent's Likeability	5	0.00	0.05	0.05
5	Agent's Sociability	3	0.02	0.03	0.05
6.1	Agent's Personality Presence	3	0.03	0.02	0.05
7	User Acceptance of the Agent	3	0.01	0.10	0.11
8	Agent's Enjoyability	4	0.00	0.09	0.09
9	User's Engagement	3	0.00	0.17	0.17
10	User's Trust	3	0.01	0.02	0.03
11	User Agent Alliance	6	0.01	0.01	0.02
12	Agent's Attentiveness	3	0.00	0.16	0.16
13	Agent's Coherence	4	0.00	0.11	0.11
14	Agent's Intentionality	4	0.01	0.03	0.04
15	Attitude	3	0.01	0.15	0.16
16	Social Presence	3	0.05	0.01	0.06
17	Interaction Impact on Self-Image	4	0.01	0.02	0.03
18.1	Agent's Emotional Intelligence Presence	5	0.12	0.00	0.12
18.3	User's Emotion Presence	4	0.00	0.03	0.03
19	User Agent Interplay	4	0.00	0.02	0.02

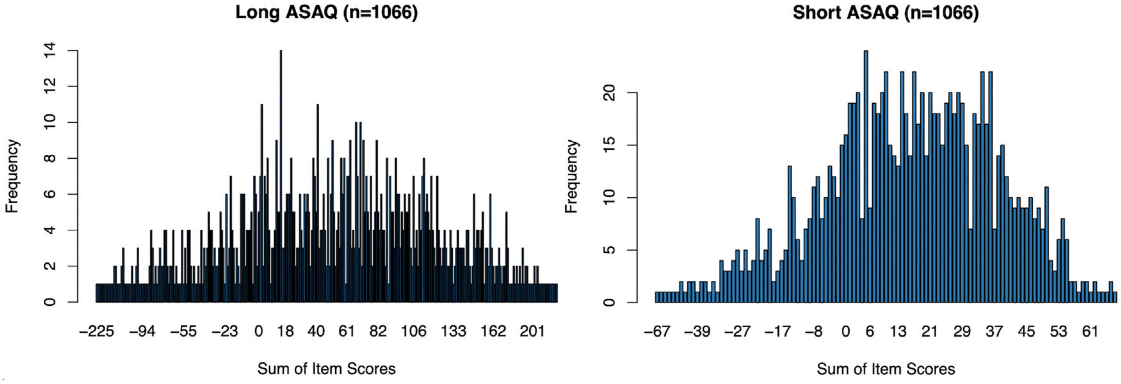


Fig. 5. Range sum of scores of each item of the long and short versions of ASAQ.

Table 8
Interpretation of the percentile rank scores (Tran et al., 2024).

Percentile	Interpretation
0 – 50%	Low
50% – 75%	Moderate
75% – 90%	High
90% – 95%	Very High
95% – 1	Exceptional

of 1.6 and an ASA B with a score of 0.1 on the Agent's Usability (2) scale of the long version; this would give us a 1.5 difference score and, consequently, the 90th percentile rank score (> 1.33 score). In other words, we can report that the observed difference between ASAs A and B is larger than 90% of differences observed in the representative data set regarding Agent's Usability (2).

These strategies provide meaning or interpretability to ASAQ results. Future studies, reporting the ASAQ scores of novel ASAs, contribute to this representative set of agents - provided these studies are well performed and the ASA is well documented. For example, a groundbreaking ASA that everyone becomes familiar with (e.g., ChatGPT) could be evaluated with ASAQ and its high-scoring constructs can become the new "target to beat" for the ASA community. In this way, future usage of the ASAQ allows for continuous calibration of the interpretation of ASAQ scores.

5. Using the ASAQ

In this section, we outline three main considerations when using the ASAQ: choosing the ASAQ version (long or short version), determining the appropriate sample size, and reporting and visualising the results.

5.1. Choosing the ASAQ version

The choice of using either the long or short version of the ASAQ depends on the purpose of the study. There are three scenarios: (1) researchers aiming for a quick first impression of the interaction experience their ASA evokes can consider using the short ASAQ version; (2) researchers aiming for a comprehensive and thorough understanding of their ASA's performance should consider opting for the long ASAQ version; and (3) researchers interested in only some of the constructs have to consider using the long ASAQ for those constructs provided they *also* include the short ASAQ for the constructs in which they are less interested. This way, their study is more relevant as it contributes to the overall community's ability to replicate and understand the human-ASA interaction experience. We recommend option three for most researchers, *so probably you (!)*, as this is economical and helps

Table 9
The percentile of ASAQ scores of the ASAQ representative set 2024 ($n = 29$).

	Percentile												
Version	5%	10%	20%	25%	30%	40%	50%	60%	70%	75%	80%	90%	95%
Long ASAQ	−0.72	2.68	9.36	10.37	12.54	13.58	14.32	16.41	17.82	19.28	20.45	24.91	27.69
Short ASAQ	−0.05	6.21	9.81	11.57	12.79	14.07	14.95	17.88	19.67	20.85	21.81	25.86	30.20

Table 10
Most conservative sample size estimations for studies comparing two ASAs with .8 power and .05 alpha level.

	Percentile:	25%	50%	75%
Version	Effect size:	Small	Medium	Large
Long ASAQ		485	116	41
Short ASAQ		614	154	53

the community understand the general experience of your ASA, while allowing you to perform comparisons on the specific constructs your work focuses on.

5.2. Sample size

The number of participants to be recruited is an important question when planning a study, especially if a frequentist approach is taken. Here, we offer suggestions for two situations: (1) when comparing between two ASAs or (2) when comparing a single ASA to the 29 agents in the ASAQ representative set 2024.

Let us start with the first situation, which concerns the Null Hypothesis Significance Testing (NHST). For this, researchers often conduct a power analysis to determine sample size, taking key parameters as the statistical significance criterion (alpha level), the power of a test, and the expected effect size. Whereas we usually follow conventions for setting power and alpha level (e.g., power of .80 and alpha of .05), the ASAQ representative set 2024 can guide us in choosing an appropriate effect size. Following the idea of Tables D.13–D.14 with percentile difference scores, Tables D.15–D.16 show the associated effect sizes using the following formula (Cohen, 2013):

$$d = \frac{M_1 - M_2}{SD_{pooled}}$$

These tables show the effect sizes ordered by percentiles of differences observed between all combinations of two agents, ranked from the smallest to largest. As a convention, we propose to classify the effect sizes of the 25th, 50th, and 75th percentile as small, medium, and large, respectively. For example, for Agent's Usability (2) in the long ASAQ version, we consider $d = .25$ a small effect, $d = .46$ a medium effect, and $d = .81$ a large effect. Consequently, a power analysis for an independent t-test, with a power of .80, alpha = .05, would indicate the need for 252, 75, or 25 participants, respectively.

When we do not target any specific construct or dimension, Tables D.15–D.16 also give an idea of what to do. In that case, if we want to be conservative, we should use the smallest effect size value of any individual construct to detect differences also for that targeted construct. A less conservative approach would be taking the median values across constructs and dimensions, which is also presented at the bottom of the tables. Additionally, for convenience, Table D.17 shows the sample sizes for each construct at the various effect sizes, for an independent t-test with .80 power and .05 alpha level. Finally, Table 10 shows the most conservative sample sizes (i.e., using the smallest effect size for any construct) for small, medium, and large effects.

Having discussed comparing two ASAs, let us now determine the sample size for a study involving a single ASA. In the previous section, we discussed how a score could be compared with the percentile rank score of the 29 agents in a representative data set. However, the sample mean ($\bar{X}$) obtained in the study is only an estimate of the population mean (μ), so researchers often report a Confidence Interval (CI) using the following formula (Montgomery and Runger, 2003):

$$\bar{X} - E < \mu < \bar{X} + E,$$

whereby $\bar{X}$ is the sample mean of ASAQ score, and E is the *maximum error of estimation*. The formula for this is (Montgomery and Runger, 2003):

$$E = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

From this, we derive the formula for the sample size, which is:

$$n = \left(\frac{z_{\alpha/2} \times \sigma}{E}\right)^2$$

The values for $z_{\alpha/2}$ are known for common α-level CIs (i.e., $100(1-\alpha)$ CI), as for a 90% CI, a 95% CI, a 97.5% CI and a 99% CI the $z_{\alpha/2}$-values are 1.64, 1.96, 2.24 and 2.58 respectively. For σ we take the pooled *SD* of each agent's construct in the representative set. Therefore, researchers only have to determine what error margin they find acceptable for their CI. For convenience, we have expressed this in the size of an interval of the ranked percentile score observed in the representative data set for the 5th, 10th, 20th, and 25th percentile sizes. For this, we have calculated the mean interval size observed in rank percentiles scores for these four interval sizes, which we then used to calculate the associated sample sizes for each construct, see Tables D.18–D.21.

Researchers can use these tables as follows. If they want to use a 95% CI to estimate the Agent's Usability (2) with the long version of ASAQ and are satisfied with an error margin associated with a 25th percentile interval, they would need at least 27 participants (Table D.21). Intuitively this means that these researchers can claim whether their score falls within the first, second, third or fourth quartile of the representative data set scores. If they want more precision, say, an error margin of a 20th percentile interval, they would need a sample size of 52 participants (Table D.20). If they are not focusing on a specific construct, they should take to maximum sample size across the constructs, in the 20th percentile case this is the construct Agent's Attentiveness (12) with a sample size of 100. Apparently, people's experience with this construct was more diverse, while this was less when it came to rating agents' human-like appearance, as that would require only 6 participants. For convenience, we have included the sample size for studies with a single ASA that use either version of ASAQ in Table 11.

5.3. Reporting and visualising your results

In the previous section, we provided some suggestions on reporting ASAQ findings that fit with the goals of the ASAQ. Specifically, we suggested contrasting ASAs with the ASAQ representative set 2024, which increases the usefulness of the ASAQ results of one study for the broader community. Besides evaluating that one agent, it adds to the body of comparisons of the human-ASA interaction experiences of the community. To aid in easy comparisons, we suggest researchers report two key aspects in their study (or in supplementary material accompanying their study): the ASAQ score and a visualisation of the ASAQ construct scores.

Firstly, the ASAQ score serves as a singular number representing the overall experience with the ASA. The ASAQ score is calculated by adding up the mean scores of all constructs and dimensions. For the

Table 11
Sample size for studies investigating a single ASA (and comparing it with the ASAQ representative set 2024) that are interested in all ASAQ constructs, for the various error margins (percentiles), specific confidence intervals (CI), and for both ASAQ versions. In bold are the suggested sample sizes considered appropriate in most studies.

Version	Error Margin	90% CI	95% CI	97.5% CI	99% CI
Long ASAQ	5%	689	979	1280	1691
	10%	201	286	374	493
	20%	70	100	130	172
	25%	46	**66**	86	114
Short ASAQ	5%	871	1237	1618	2137
	10%	350	497	650	859
	20%	76	108	141	186
	25%	68	**97**	126	167

short version, this means adding up all item scores after flipping the scores of reverse-scoring items. The ASAQ scores of all the ASAs used in our studies can be seen in . To ensure that other researchers can compare their ASAs, regardless of which construct they might be most interested in, we strongly recommend that researchers *always* report the ASAQ score in their studies *and* all constructs' scores.

Secondly, we suggest researchers generate an ASAQ chart to visualise the profile of the human-ASA interaction experience of their ASA (Fitrianie et al., 2022a). These charts show the ASA's score on 24 constructs and dimensions, on the original −3 to +3 scale, or in percentiles based on a representative set 2024 of 29 agents. The centre of these charts shows the ASAQ scores or percentile scores of the ASAs. For example, Fig. 6 shows the profile of two agents from the representative set, Siri and NAO. We provide a script with examples to generate ASAQ charts (Fitrianie et al., 2025), which expects the data of an ASAQ study of each agent in a comma-separated values (CSV) format (i.e., for the long ASAQ version the data is the mean of 24 ASAQ constructs/dimensions, while for the short version is the mean of its 24 items) and organised following the numbering of ASAQ's constructs/dimensions (from HLA to UAI). This chart can be added to the paper or report of a study or, if that is not possible, as supplementary material. Additionally, the chart is also useful in (poster) presentations to offer an at-a-glance overview of the ASAQ results.

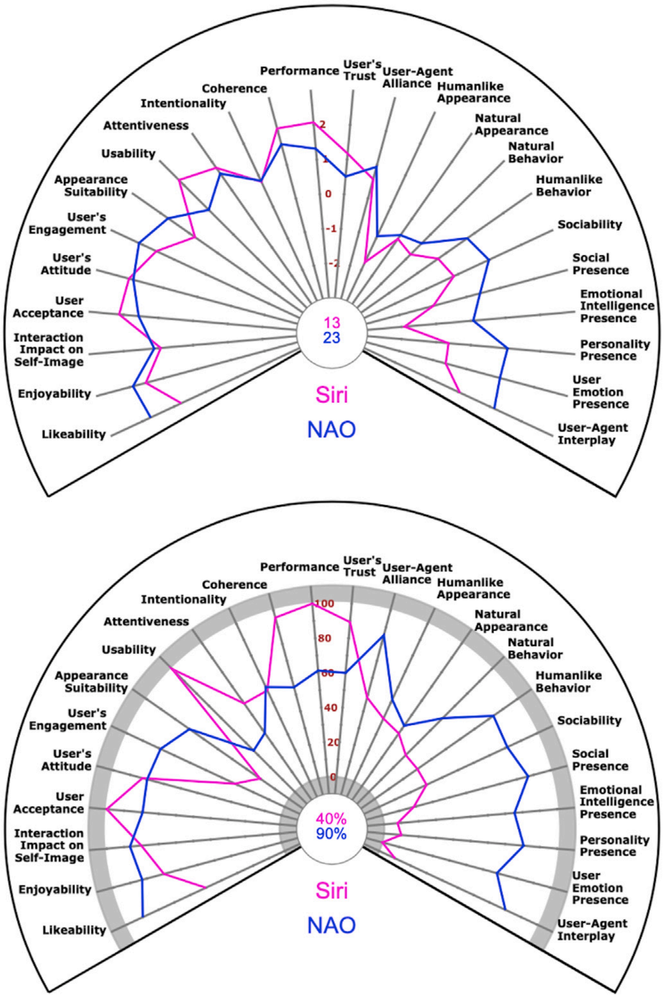


Fig. 6. The performance of Siri and NAO. Above: the ASAQ chart. Below: the percentile ASAQ chart comparing the representative set 2024. Here, the grey area indicates scores below or above the representative set.

6. Discussion and conclusion

The analysis of the characteristics of the ASAQ allows us to draw several conclusions. First, the experience measured by ASAQ extensively covers the community's interest, as its 19 constructs are associated with more than 80% of the questionnaire constructs identified in studies presented at IVA between 2013 and 2018. Therefore, reporting ASAQ results helps the community establish insights on various ASA aspects they value; hence, ASAQ allows individual studies to contribute beyond their specific research interest. Second, three separate studies demonstrate a respectable level of reliability, indicating internal consistency between measurement results of construct items, and indicating they at least align with capturing their latent construct. Third, the different studies also suggest a positive outlook on the validity of ASAQ, specifically content validity, construct validity, cross-validity, and predictive validity. In other words, the content validity shows that experts confirm the ASAQ items' association with the construct they are measuring. Furthermore, the construct validity shows that a model where ASAQ items are only associated with their intended construct but where constructs are allowed to correlate can accurately describe the ASAQ data collected. During the creation of the ASAQ, in the construct validity study, we also removed questionnaire items to optimise the model's fit. Therefore, repeating the analysis on a new data set in a cross-validity study was necessary and the results confirmed that we could reproduce this good fit of such a model. Additionally, the predictive validity shows that ASAQ outcomes correlate with experts' predictions. Further, the high correlation between data from the long and short ASAQ versions demonstrates the concurrent validity of both versions. Fourth, the interpretability of the ASAQ results is promising as the scale provides adequate room for people to express their extent of agreement or disagreement. Furthermore, researchers can compare their ASA's score with the scores of the representative 29 agents, using them as anchors for their findings. Additionally, they can rank their ASA's score against those 29 agents from the ASAQ representative set 2024. This approach allows researchers to make statements such as the ASA's performance score falling in the top 10% of the ASAQ representative agent set. Fifth, we provide a guide on selecting the long or short ASAQ version and determining the sample size for studies using ASAQ. For their power analysis to compare two ASAs, researchers can look up the effect size equal to or larger than, for instance, the bottom 25% differences between agents in the representative data set, ranked from smallest to largest. Alternatively, researchers, who want to base their sample size on a desired level of precision about a single ASA, can look up the required sample sizes associated with the error margin expressed in percentile accuracy of scores observed in ASAQ representative set 2024. Finally, we provide suggestions on how and what to report from a study using ASAQ, including a visualisation in the form of an ASAQ chart, which illustrates the ASA's performance on all 24 constructs/dimensions at-a-glance in a single figure.

These conclusions also have limitations. First, we asked the participants to rate their experience after watching a video without directly interacting with an ASA (third-person perspective). Although this was pragmatic and allowed us to conduct studies with many participants in different countries with various ASAs, scores based on vicarious experiences might differ from those based on direct experiences with ASAs (first-person perspective). Here, how well participants identified with the person in the video might influence the impact of their vicarious experience. For example, Kang et al. (2021) found that the degree of self-identification moderated the impact of vicarious experience on self-efficacy. Therefore, the scores in the representative data set might be limited in how well they generalise to experience based on actual interaction and, consequently, the absolute score, the effect size, and the sample size derived from them. Hence, as we write this paper, the workgroup is collecting ASAQ data on people's direct experiences with ASAs Fitrianie et al. (2023). Despite this limitation, we do not expect these vicarious base experiences to limit the generalisation of the conclusions drawn concerning reliability and validity analyses as we primarily base them on correlation analyses.

Second, besides the expert data, we collected all other data from participants registered on the platform Prolific Academic, which is limited to people in Croatia and South Africa and in countries represented in the Organisation for Economic Co-Operation and Development (OECD) countries[8] (with the exceptions of Turkey, Lithuania, Colombia, and Costa Rica where Prolific is not available). Further, Prolific populations might have a different demographic distribution than the general public. Although we refrained from collecting demographic data such as gender and age, the gender distribution of participants registered on this platform[9] was 30.3% (Men), 43.6% (Women), and 1.5% (non-Binary), and 24.6% (not reported), and the age distribution was 27.6% (18–25 years), 36.5% (26–35 years), 18.7% (36–45 years), 10.3% (46–55 years), 5% (46–55 years), 1.9% (older than 65) and 0.1% (not reported). Therefore, researchers should exercise caution when generalising the reported ASAQ scores, effect sizes, and sample sizes to populations with a different demographic distribution or from countries not included in this platform.

Third, since the ASAQ measures interaction experience, which is useful for determining what is working or needs improvement, it might not be directly clear which part of the agent is responsible for the observed outcome. To investigate this, researchers could compare ASAQ scores across multiple versions of an agent, or seek a questionnaire that specifically assesses parts or modalities of the agent. For example, the Mean Opinion Scale (Polkosky and Lewis, 2003) measures only the speech quality of an agent. To maintain the general applicability of the ASAQ, we avoided references to such specifics. Future researchers might also explore whether a component-based ASAQ version would be feasible, similar to what has been proposed for measuring the usability of interaction components of a system (Brinkman et al., 2009).

Fourth, the epistemological starting point of what an ASA is can vary between studies, ASAs, users, and researchers: is an ASA a tool (warranting words like 'use') or a social actor (making words like 'interact' more appropriate)? The ASAQ mixes these as it is a reflection of these various standpoints in the research community. Therefore, a researcher strictly following one of these epistemological stances might not appreciate all constructs, dimensions, or item formulations in the ASAQ.

These differing perspectives suggest that any questionnaire, including the ASAQ, must evolve continuously - not only to incorporate new insights into social sensitivities within specific contexts but also to reflect broader shifts in social norms. The need for evolution was highlighted in our experience translating the ASAQ into German, a gendered language, where we had to address the agent according to its gender. This challenge also applies to the ongoing debate about whether agents should exhibit human-like characteristics. For example, the European Union AI Act (2024) (European Union, 2024) prohibits AI entities from pretending to be humans. Shneiderman (2020) takes this even further and argues that ASAs should not be designed to mimic humans at all.

6.1. Future work

The work also provides opportunities for future research, such as examining the validity of the ASAQ criteria; in other words, how does ASAQ compare with 'gold' standards or criteria (Neuman, 2013)? Traditionally, this is split up into subtypes: (1) concurrent validity, how does it compare with the existing measures?, and (2) predictive validity, how does the ASAQ score predict a standard in the future? We already see an example of concurrent validity when comparing the short and long ASAQ versions; in that case, we considered the long version the gold standard. However, a future task is comparing constructs measured in the ASAQ, such as Agent's Usability, with existing measures that either align well on content (e.g., the Bot Usability Scale (Borsci et al., 2022)) or are widely used (e.g., UTAUT (Venkatesh et al., 2003) or the Godspeed questionnaire (Bartneck et al., 2009)). As we write this paper, the workgroup is studying the ASAQ's concurrent validity by comparing it with various existing questionnaires (Fitrianie et al., 2023). Additionally, we can examine predictive validity by selecting a group of ASAs and predicting their scores, as we have done for the 29 representative agents. Expanding on this would mean selecting ASAs where insight about, for example, their usability has been confirmed empirically and then investigating whether the ASAQ would reconfirm this.

As shown in Fig. 2, ASAQ focuses on a person's interaction experience with a specific agent, leaving out factors preceding and succeeding the interaction. Future research, however, could specifically study the relationship, for example, between demographic factors, such as gender or age, and ASAQ constructs. And also, for instance, which ASAQ constructs predict specific context-dependent process or outcome factors, such as enhancing healthy eating or someone's body mass index.

Future work can also focus on strengthening researchers' adoption of ASAQ measures. We have worked with the community to establish the measure and reported about it at various conferences (Fitrianie et al., 2019, 2020a, 2021a, 2022a; Albers et al., 2024) and workshops, which means that an extensive group already recognises the work and considers it relevant to some degree. This is also reflected in the large number of ASA researchers involved in the OSF workgroup. Nevertheless, developing tutorials and educational material such as manuals, instruction videos and statistical packages (e.g., in R) to analyse and visualise results, could further facilitate and encourage researchers to use the ASAQ. Although the ASAQ score of a representative data set of 29 agents is available online, offering researchers a repository to share and find ASAQ results allows researchers to more easily compare their findings.

Searching in this repository also requires a taxonomy for describing objective properties of an ASA, such as, as mentioned before, their communication modality, communication language, embodiment, and

[8] See the list in https://www.oecd.org/en/countries.html

[9] Data from participants active in the past 90 days before the moment of collecting this data, 17-07-2024.

mobility. Likewise, taxonomies are needed to describe participants and study setups. Thus, well-executed future research that reports the ASAQ scores of new ASAs can enhance the ASAQ representative set 2024 of agents presented in this paper. The next widely recognised, revolutionary ASA like ChatGPT could set a new benchmark for the ASA community when assessed using the ASAQ. This ongoing use of ASAQ allows for a continuous re-calibration of score interpretations, and that should be reflected in a periodic systematic update of the representative set of agents.

Furthermore, researchers could use a representative set of agents to examine different types, embodiments, domains, user perspectives, and settings of ASAs. The ASAQ can also be continuously refined. For instance, a new short version of the ASAQ could be developed, featuring questionnaire items that capture the complete definition of each construct or dimension, rather than selecting representative items from each. This refinement can also take the form of differentiating between the types of ASAs, for instance, by creating partial ASAQ scores derived from selected constructs relevant to a particular type of ASA.

To conclude, we hope that the ASAQ will help the community to make valid claims, replicate findings, and establish more insight into how people experience their interaction with ASAs, especially as more and more of these ASAs enter our daily lives.

Glossary

- Artificial social agent: A computer-controlled entity that can autonomously interact with humans following the social rules of human-human interactions.
- ASAQ Chart: Visualizing the ASAQ analysis result of an ASA on the −3 to 3 scale.
- ASAQ Percentile Chart: Contrasting the ASAQ analysis result of an ASA with the ASAQ representative set.
- ASAQ-Score: A number calculated by adding up the mean score of the ASAQ constructs and dimensions of an ASA. For the short version, this means simply adding all up items.
- ASAQ representative set 2024: Dataset of 29 agents collected in the Studies Mid '21 and '22.

Abbreviations

- ASA: Artificial Social Agent
- ASAQ: Artificial-Social-Agent Questionnaire
- ASIMO: Advanced Step in Innovative Mobility
- BML: Behavioural Markup Language
- CCT: Classical Test Theory
- CFA: Confirmatory Factor Analysis
- CFI: Comparative Fit Index
- CI: Confidence Interval
- CSV: Comma-Separated Values
- EFA: Exploratory Factor Analysis
- FML: Functional Markup Language
- FN: False Negative
- FP: False Positive
- FPR: False Postive Rate
- IVA: Intelligent Virtual Agent
- MBTI: The Myers-Briggs Type Indicator
- NHST: Null Hypothesis Significance Testing
- OCEAN: Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism
- OECD: Organisation for Economic Co-Operation and Development
- RF: Relative Frequency
- SAIBA: Situation, Agent, Intention, Behaviour, and Animation
- TN: True Negative
- TP: True Positive
- TPR: True Positive Rate
- TUDelft: Delft University of Technology

CRediT authorship contribution statement

Siska Fitrianie: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation. **Merijn Bruijnes:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Conceptualization. **Amal Abdulrahman:** Writing – review & editing, Writing – original draft, Validation, Formal analysis. **Willem-Paul Brinkman:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used Grammarly and GenAI in order to receive feedback on (parts of) English spelling, grammar and formulation. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge the efforts of all OSF workgroup members, and in particular mention the large recurring efforts of Nele Albers, Andrea Bönsch, Jonathan Ehret, Fengxian Li, and Deborah Richards. Parts of this work were funded by the 4TU Pride and Prejudice project.

Appendix A. The artificial-social-agent questionnaire

Note:

- [R] refers to reverse-scoring questionnaire item,
- [The agent] can be replaced with the ASA's name,
- [.. / ..], e.g. [I am/The user is], means to use either one, and
- Items labelled with $*$ are representative items used in the short version of the questionnaire.
- The construct names, codes, and definitions should not be included in your questionnaire as presented to your participants. For template-examples of how to present the questionnaire, see the project website: https://asaq.ewi.tudelft.nl.

Seven-point rating scale [-3, +3]:

- -3 label: disagree
- 0 label: neither agree nor disagree
- 3 label: agree

A.1. Agent's believability

The extent to which a user believes that the artefact is a social agent.

A.1.1. Human-like appearance

The extent to which a user believes that the social agent appears like a human.

(HLA1) [The agent]'s appearance is human
(HLA2) [The agent] has the appearance of a human $*$
(HLA3) [The agent] has a human-like outside
(HLA4) [The agent]'s appearance makes me think of a human

A.1.2. Human-like behaviour

The extent to which a user believes that the social agent behaves like a human.

(HLB1) A human would behave like [the agent]
(HLB2) [The agent]'s manners are consistent with that of people
(HLB3) [The agent] behaviour makes me think of human behaviour
(HLB4) [The agent] behaves like a real person
(HLB5) [The agent] has a human-like manner ∗

A.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.

(NA1) [The agent] appears like something that could exist in nature
(NA2) [The agent] has a natural physique
(NA3) [The agent]'s resemblance has an organic origin
(NA4) [The agent] seems natural from the outward appearance ∗
(NA5) How [the agent] is represented is realistic

A.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.

(NB1) [The agent] is alive
(NB2) [The agent] acts naturally
(NB3) [The agent] reacts like a living organism*

A.1.5. Agent's appearance suitability

The extent to which the agent's appearance is suitable for its role.

(AAS1) [The agent]'s appearance is appropriate ∗
(AAS2) [The agent]'s physique is suitable for its role
(AAS3) [The agent]'s appearance was suitable

A.2. Agent's usability

The extent to which a user believes that using an agent will be free from effort (future process).

(AU1) [The agent] is easy to use ∗
(AU2) Learning to work with [the agent] is easy
(AU3) Learning how to communicate with [the agent] is quick

A.3. Performance

The extent to which a task was well performed (past performance).

(PF1) [The agent] does its task well ∗
(PF2) [The agent] does not hinder [me/the user]
(PF3) [I am/The user is] capable of succeeding with [the agent]

A.4. Agent's likeability

The agent's qualities that bring about a favourable regard.

(AL1) [The agent]'s appearance is pleasing
(AL2) I like [the agent] ∗
(AL3) [R] I dislike [the agent]
(AL4) [The agent] is cooperative
(AL5) I want to hang out with [the agent]

A.5. Agent's sociability

The agent's quality or state of being sociable.

(AS1) [The agent] can easily mix socially ∗
(AS2) It is easy to mingle with [the agent]
(AS3) [The agent] interacts socially with [me/the user]

A.6. Agent's personality

The combination of characteristics or qualities that form an individual's distinctive character.

A.6.1. Agent's personality presence

To what extent the user believes that the agent has a personality.

(APP1) [The agent] has a distinctive character ∗
(APP2) [R] [The agent] is characterless
(APP3) [The agent] is an individual

A.6.2. Agent's personality type

The particular personality of the agent.

A.7. User acceptance of the agent

The willingness of the user to interact with the agent.

(UAA1) [I/The user] will use [the agent] again in the future ∗
(UAA2) [I/The user] can see themselves using [the agent] in the future
(UAA3) [R] [I oppose/The user opposes] further interaction with [the agent]

A.8. Agent's enjoyability

The extent to which a user finds interacting with the agent enjoyable.

(AE1) [R] [The agent] is boring ∗
(AE2) It is interesting to interact with [the agent]
(AE3) [I enjoy/The user enjoys] interacting with [the agent]
(AE4) [R] [The agent] is unpleasant to deal with

A.9. User's engagement

The extent to which the user feels involved in the interaction with the agent.

(UE1) [I/The user] was concentrated during the interaction with [the agent]
(UE2) The interaction captured [my/the user's] attention ∗
(UE3) [I/The user] was alert during the interaction with [the agent]

A.10. User's trust

The extent to which a user believes in the reliability, truthfulness, and ability of the agent (for future interactions).

(UT1) [The agent] always gives good advice
(UT2) [The agent] acts truthfully
(UT3) [I/The user] can rely on [the agent] ∗

A.11. User-agent alliance

The extent to which a beneficial association is formed.

(UAL1) [The agent] and [I/the user] have a strategic alliance ∗
(UAL2) Collaborating with [the agent] is like a joint venture
(UAL3) [The agent] joins [me/the user] for mutual benefit
(UAL4) [The agent] can collaborate in a productive way
(UAL5) [The agent] and [I/the user] are in sync with each other
(UAL6) [The agent] understands [me/the user]

A.12. Agent's attentiveness

The extent to which the user believes that the agent is aware of and has attention for the user.

(AA1) [The agent] remains focused on [me/the user] throughout the interaction
(AA2) [The agent] is attentive ∗
(AA3) [I/The user] receives [the agent]'s full attention throughout the interaction

A.13. Agent's coherence

The extent to which the agent is perceived as being logical and consistent.

(AC1) [R] [The agent]'s behaviour does not make sense ∗
(AC2) [R] [The agent]'s behaviour is irrational
(AC3) [R] [The agent] is inconsistent
(AC4) [R] [The agent] appears confused

A.14. Agent's intentionality

The extent to which the agent is perceived as being deliberate and has deliberations.

(AI1) [The agent] acts intentionally
(AI2) [The agent] knows what it is doing
(AI3) [R] [The agent] has no clue of what it is doing ∗
(AI4) [The agent] can make its own decision

A.15. Attitude

A favourable or unfavourable evaluation toward the interaction with the agent.

(AT1) [I see/The user sees] the interaction with [the agent] as something positive ∗
(AT2) [I view/The user views] the interaction as something favourable
(AT3) [R] [I think/The user thinks] negatively of the interaction with [the agent]

A.16. Social presence

The degree to which the user perceives the presence of a social entity in the interaction.

(SP1) [The agent] has a social presence
(SP2) [The agent] is a social entity ∗
(SP3) [I have/The user has] the same social presence as [the agent]

A.17. Interaction impact on self-image

How the user believes others perceive the user because of the interaction with the agent.

(IIS1) [My/The user's] friends would recommend [me/them] to use [the agent]
(IIS2) Others would encourage [me/the user] to use [the agent] ∗
(IIS3) [The agent] makes [me/the user] look good
(IIS4) People would look favourably at [me/the user] because of [my/their] interaction with [the agent]

A.18. Emotional experience

A self-contained phenomenal experience. They are subjective, evaluative, and independent of the sensations, thoughts, or images evoking them.

A.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.

(AEI1) [The agent] is emotional
(AEI2) [The agent] experiences emotions
(AEI3) [R] [The agent] is emotionless ∗
(AEI4) [The agent] can express its feelings
(AEI5) [R] [The agent] cannot experience emotions

A.18.2. Agent's emotional intelligence type

The particular emotional state of the agent.

A.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.

(UEP1) [The agent]'s attitude influences how [I feel/the user feels]
(UEP2) [I am/The user is] influenced by [the agent]'s moods
(UEP3) The emotions [I feel/the user feels] during the interaction are caused by [the agent] ∗
(UEP4) [My/The user's] interaction with [the agent] gives [me/them] an emotional sensation

A.18.4. User's emotion type

The particular emotional state of the user during or after the interaction with the agent.

A.19. User-agent interplay

The extent to which the user and the agent have an effect on each other.

(UAI1) [My/The user's] emotions influence the mood of the interaction
(UAI2) [The agent] reciprocates [my/the user's] actions
(UAI3) [The agent]'s and [my/the user's] behaviours are in direct response to each other's behaviour
(UAI4) [The agent]'s and [my/the user's] emotions change to what [we/they] do to each other ∗

Appendix B. The short version of the artificial-social-agent questionnaire

Note:

- [R] refers to 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.

Seven-point rating scale [-3, +3]:

- -3 label: disagree
- 0 label: neither agree nor disagree
- 3 label: agree

24 items of the short version of the ASAQ:

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

Table C.1
The stimuli used in studies: 26 ASAs, a dog, a fish and a zombie.

Study	Agent	Description	Modality	Communication Language	Embodiment	Mobility	ASAQ Score
Mid '21	Aibo	Robotic dog developed by Sony	auditory, visual, tactile	body-, symbolic- & non-language	physical	physical	19
'22	Alexa	Virtual assistant developed by Amazon	auditory	spoken	disembodied	not applicable	10
'22	Alice	Virtual human part of the ARIA framework (Valstar et al., 2016)	auditory, visual	spoken, body language	virtual	virtual	12
Mid '21	Amy	Virtual healthcare agent (Lisetti et al., 2013)	auditory, visual	spoken, body language	virtual	virtual	9
Mid '21	CHAPPiE	Robot character in CHAPPiE	auditory, visual, tactile	spoken, body language	physical	physical	18
'22	C3PO	Fictional character from Star Wars - Saga	auditory, visual, tactile	spoken, body language	physical	physical	18
Mid '21	DeepBlue	Chess playing computer developed by IBM	visual	symbolic language	disembodied	not applicable	7
Mid '21	Dog	A real dog - a domesticated carnivore mammal	auditory, visual, tactile	spoken, body- & non-language	physical	physical	29
'22	Effie	A virtual human therapist developed by (Ranjbartabar and Richards, 2016)	auditory, visual	spoken, textual, body language	virtual	stationary	-1
'22	Fish	A real fish - an aquatic animal	visual	body language	body language	physical	32
Mid '21	Furby	Toy resembling a hamster or owl-like creature developed by Tiger Electronics	auditory, visual, tactile	spoken, body- & non-language	physical	physical (limited)	13
'22	Furhat	Physical, stationary robot developed by Furhat Robotics	auditory, visual	spoken, body language	physical	stationary	14
'22	Geminoid	Human android developed by Hiroshi Ishiguro Laboratories (Sakamoto and Ishiguro, 2009)	auditory, visual, tactile	spoken, body language	physical	stationary	18
Mid '21	Hal 9000	Fictional character in A Space Odyssey	auditory	spoken	disembodied	not applicable	14
Mid '21	iCAT	Cat-like robot developed by Philips	auditory, visual, tactile	spoken, body language	physical	stationary	-2
'22	Kitt	High-tech car in the TV series Knight Rider	auditory, visual	spoken, non language	physical	physical	16
'22	Lola	Virtual human therapist developed by (Lisetti et al., 2013)	auditory, visual	spoken, body language	virtual	stationary	16
Mid '21	Marcus	Cyborg character in Terminator	auditory, visual, tactile	spoken, body language	physical	physical	25
Mid '21	Nao	Humanoid robot from Aldebaran Robotics	auditory, visual, tactile	spoken, body language	physical	physical	23
'22	Paro	A therapeutic robot baby harp seal developed by AIST	auditory, visual, tactile	body- & non-language	physical	stationary	19
Mid '21	Poppy	Virtual human from SEMAINE (Mckeown et al., 2013)	auditory, visual	spoken, body language	virtual	stationary	14
'22	Robot Boss	Self-driving, two-wheeled videoconferencing robot developed by Double Robotics	auditory, visual, tactile	spoken	physical	physical	0
'22	Samantha	Virtual fictional character in HER	auditory	spoken	disembodied	not applicable	25
Mid '21	Sarah	Customer service from Digital Humans	auditory, visual	spoken, body language	virtual	stationary	22
Mid '21	Sim Sensei	Virtual healthcare agent (DeVault et al., 2014)	auditory, visual	spoken, body language	virtual	stationary	17
Mid '21	Siri	Virtual assistant developed by Apple	auditory	spoken	disembodied	not applicable	13
'22	The Ambient Light TV	Lighting system actively adjusted both color and brightness upon the TV content developed by Philips	visual	Symbolic language	disembodied	not applicable	3
'22	The Negotiator	Virtual agent developed by USC ICT (Nazari et al., 2015)	auditory, visual	spoken, body language	virtual	stationary	14
'22	Zombie	Reanimated corpse of a human being that has developed hunger for flesh	auditory, visual, tactile	body- & non-language	physical	physical	10

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 behaviour 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

Appendix C. The ASAQ representative set 2024

Note:

- Study = the ASA questionnaire validation study where the ASA used as stimulus in the Study Mid '21 (Construct Validity of the ASAQ) and the Study '22 (Cross-Validity of the ASAQ).
- Modality = communication modalities.
- Communication Language = language used (by human and/or the agent) to communicate, i.e., spoken, body language (i.e., facial expression, head-, limbs- or body motion), symbolic (e.g. buzzers, lights, cards), and non-language vocalisation (e.g. vocal sounds without words, bark).
- ASAQ-scores are calculated based on the long ASAQ version.

See Table C.1.

Table D.1
The relative frequency of how often a score was used (on items, $n = 1066$) based on the **long** version of the ASAQ.

No.	Construct/Dimension	Relative Frequency						
		−3	−2	−1	**0**	1	2	3
1.1	Human-Like Appearance	0.37	0.09	0.08	0.07	0.15	0.12	0.12
1.2	Human-Like Behaviour	0.16	0.11	0.12	0.11	0.24	0.16	0.11
1.3	Natural Appearance	0.22	0.10	0.11	0.13	0.19	0.13	0.11
1.4	Natural Behaviour	0.26	0.10	0.10	0.12	0.18	0.12	0.12
1.5	Agent's Appearance Suit.	0.03	0.03	0.04	0.14	0.23	0.25	0.27
2	Agent's Usability	0.03	0.03	0.05	0.16	0.23	0.26	0.24
3	Performance	0.02	0.03	0.06	0.19	0.22	0.26	0.22
4	Agent's Likeability	0.11	0.06	0.07	0.16	0.17	0.18	0.25
5	Agent's Sociability	0.11	0.10	0.11	0.17	0.22	0.17	0.12
6.1	Agent's Personality Pr.	0.15	0.09	0.12	0.15	0.19	0.16	0.13
7	User Acceptance of the A.	0.04	0.04	0.07	0.20	0.18	0.21	0.26
8	Agent's Enjoyability	0.05	0.05	0.08	0.13	0.19	0.22	0.29
9	User's Engagement	0.02	0.02	0.04	0.08	0.20	0.29	0.36
10	User's Trust	0.08	0.06	0.07	0.31	0.19	0.17	0.12
11	User Agent Alliance	0.08	0.07	0.09	0.21	0.25	0.18	0.12
12	Agent's Attentiveness	0.03	0.02	0.05	0.08	0.21	0.25	0.36
13	Agent's Coherence	0.02	0.03	0.06	0.14	0.16	0.24	0.34
14	Agent's Intentionality	0.09	0.06	0.08	0.16	0.21	0.20	0.19
15	Attitude	0.04	0.05	0.08	0.15	0.19	0.21	0.28
16	Social Presence	0.18	0.12	0.14	0.18	0.19	0.12	0.07
17	Interaction Impact on Self.	0.06	0.06	0.08	0.30	0.20	0.18	0.12
18.1	Agent's Emotional Int. Pr.	0.31	0.12	0.11	0.14	0.15	0.10	0.07
18.3	User's Emotion Presence	0.08	0.07	0.08	0.16	0.24	0.20	0.16
19	User Agent Interplay	0.07	0.06	0.08	0.18	0.25	0.20	0.16
	Mean:	0.11	0.07	0.08	0.16	0.20	0.19	0.19
	SD:	0.10	0.03	0.03	0.06	0.03	0.05	0.09
	Median:	0.08	0.06	0.08	0.16	0.20	0.19	0.16
	Min:	0.02	0.02	0.04	0.07	0.15	0.10	0.07
	Max:	0.37	0.12	0.14	0.31	0.25	0.29	0.36

Table D.2
The relative frequency of how often a score was used (n = 1066) based on the **short** version of the ASAQ.

No.	Construct/Dimension	Relative Frequency						
		−3	−2	−1	0	1	2	3
1.1	Human-Like Appearance	0.39	0.08	0.08	0.06	0.14	0.11	0.13
1.2	Human-Like Behaviour	0.14	0.10	0.12	0.09	0.26	0.18	0.12
1.3	Natural Appearance	0.21	0.12	0.14	0.11	0.18	0.12	0.12
1.4	Natural Behaviour	0.15	0.09	0.11	0.10	0.23	0.17	0.15
1.5	Agent's Appearance Suit.	0.04	0.03	0.04	0.13	0.22	0.25	0.29
2	Agent's Usability	0.02	0.03	0.04	0.16	0.21	0.27	0.27
3	Performance	0.02	0.02	0.04	0.15	0.20	0.30	0.27
4	Agent's Likeability	0.09	0.05	0.06	0.19	0.16	0.19	0.26
5	Agent's Sociability	0.17	0.15	0.13	0.20	0.18	0.11	0.07
6.1	Agent's Personality Pr.	0.09	0.06	0.11	0.17	0.25	0.19	0.13
7	User Acceptance of the A.	0.04	0.03	0.04	0.23	0.19	0.21	0.26
8	Agent's Enjoyability	0.06	0.07	0.11	0.16	0.17	0.19	0.25
9	User's Engagement	0.01	0.02	0.03	0.08	0.21	0.32	0.33
10	User's Trust	0.06	0.07	0.10	0.22	0.24	0.19	0.13
11	User Agent Alliance	0.12	0.10	0.09	0.30	0.20	0.12	0.07
12	Agent's Attentiveness	0.03	0.03	0.04	0.10	0.26	0.26	0.28
13	Agent's Coherence	0.02	0.02	0.06	0.14	0.16	0.26	0.34
14	Agent's Intentionality	0.07	0.05	0.08	0.15	0.16	0.22	0.27
15	Attitude	0.03	0.05	0.08	0.14	0.23	0.20	0.27
16	Social Presence	0.17	0.10	0.13	0.21	0.19	0.13	0.07
17	Interaction Impact on Self.	0.05	0.05	0.06	0.25	0.24	0.20	0.15
18.1	Agent's Emotional Int. Pr.	0.25	0.12	0.12	0.13	0.16	0.12	0.10
18.3	User's Emotion Presence	0.08	0.07	0.08	0.15	0.23	0.20	0.18
19	User Agent Interplay	0.09	0.08	0.09	0.23	0.23	0.17	0.10
	Mean:	0.10	0.07	0.08	0.16	0.20	0.20	0.19
	SD:	0.09	0.04	0.03	0.06	0.04	0.06	0.09
	Median:	0.08	0.06	0.08	0.16	0.20	0.19	0.16
	Min:	0.01	0.02	0.03	0.06	0.14	0.11	0.07
	Max:	0.39	0.15	0.14	0.30	0.26	0.32	0.34

Appendix D. Statistical analysis based on the combined studies mid-2021 and 2022

See Tables D.1–D.21.

Table D.3
The ASAQ constructs or dimension scores of AIBO, ALEXA, ALICE, AMY, CHAPPIE, C3PO, and DEEPBLUE based on the **long** version of ASAQ. The mean (standard deviation) values of each construct and dimension are calculated based on the mean scores of the construct's or dimension's items.

No.	Construct/Dimension	AIBO	ALEXA	ALICE	AMY	CHAPPIE	C3PO	DEEPBLUE
		n =39	*n* = 37	*n* = 35	*n* = 39	*n* = 38	*n* = 36	*n* = 39
1.1	Human-Like Appearance	−2.54 (0.78)	−2.47 (0.90)	0.96 (1.58)	0.88 (1.38)	−1.09 (1.54)	−0.31 (1.72)	−1.96 (1.56)
1.2	Human-Like Behaviour	−1.59 (1.11)	−0.34 (1.35)	0.21 (1.34)	0.09 (1.51)	0.31 (1.09)	0.42 (1.29)	−0.55 (1.61)
1.3	Natural Appearance	0.04 (1.47)	−1.30 (1.02)	0.11 (1.18)	−0.25 (1.18)	−0.92 (1.16)	−0.64 (1.18)	−1.10 (1.20)
1.4	Natural Behaviour	−0.30 (1.30)	−1.06 (1.23)	−1.09 (1.14)	−0.85 (1.40)	0.16 (1.36)	0.04 (1.28)	−1.14 (1.55)
1.5	Agent's Appearance Suit.	2.06 (0.81)	1.13 (1.36)	1.44 (1.08)	1.21 (1.27)	0.95 (1.28)	1.61 (0.84)	1.09 (1.27)
2	Agent's Usability	1.78 (0.77)	2.02 (0.97)	1.62 (1.25)	1.28 (0.83)	0.75 (1.09)	1.07 (1.31)	0.80 (1.25)
3	Performance	1.56 (0.95)	1.78 (0.77)	1.05 (1.38)	0.87 (1.08)	1.13 (0.85)	1.15 (1.28)	1.92 (0.97)
4	Agent's Likeability	1.48 (1.23)	0.77 (1.19)	0.77 (1.06)	0.08 (1.28)	1.27 (0.98)	1.44 (1.26)	0.33 (1.07)
5	Agent's Sociability	0.93 (1.36)	−0.03 (1.24)	0.01 (1.42)	−0.28 (1.25)	0.63 (1.07)	0.42 (1.20)	−1.23 (1.62)
6.1	Agent's Personality Pr.	0.32 (1.14)	−0.53 (1.40)	−0.64 (1.57)	−0.44 (1.22)	0.77 (1.42)	1.31 (1.38)	−1.31 (1.43)
7	User Acceptance of the A.	1.82 (0.98)	1.94 (1.00)	0.98 (1.09)	0.63 (0.98)	1.39 (1.13)	0.48 (1.51)	1.34 (1.06)
8	Agent's Enjoyability	1.96 (0.97)	1.30 (1.13)	0.95 (1.12)	0.31 (1.09)	1.82 (0.88)	1.10 (1.38)	1.18 (0.98)
9	User's Engagement	2.19 (0.97)	1.54 (1.06)	1.93 (0.92)	1.31 (0.93)	2.04 (0.77)	1.43 (0.94)	1.80 (1.01)
10	User's Trust	−0.09 (1.12)	1.01 (1.12)	0.73 (1.20)	0.61 (0.93)	0.07 (1.14)	1.20 (1.12)	1.18 (1.23)
11	User Agent Alliance	0.36 (1.17)	0.60 (0.99)	0.48 (1.06)	0.10 (1.04)	0.65 (0.89)	0.70 (1.06)	0.63 (1.25)
12	Agent's Attentiveness	1.85 (0.96)	1.80 (0.98)	2.13 (1.04)	1.66 (1.11)	1.15 (1.19)	1.73 (0.91)	1.32 (1.49)
13	Agent's Coherence	1.76 (0.89)	1.89 (0.89)	1.65 (0.94)	1.50 (1.01)	1.05 (0.95)	1.33 (1.07)	2.11 (0.88)
14	Agent's Intentionality	0.13 (1.42)	0.61 (1.06)	0.44 (1.35)	0.33 (1.44)	0.18 (1.39)	1.41 (0.97)	1.54 (1.31)
15	Attitude	2.37 (0.67)	1.71 (0.86)	1.08 (1.33)	0.36 (1.05)	1.86 (1.04)	0.03 (1.72)	1.47 (1.06)
16	Social Presence	0.03 (1.45)	−0.77 (1.41)	−0.98 (1.32)	−0.46 (1.41)	0.25 (1.21)	−0.27 (1.21)	−1.21 (1.58)
17	Interaction Impact on Self.	1.06 (1.16)	0.74 (0.92)	0.16 (0.97)	−0.10 (0.81)	0.88 (0.98)	0.48 (1.15)	0.83 (1.07)
18.1	Agent's Emotional Int. Pr.	−0.69 (1.58)	−1.71 (1.45)	−1.46 (1.36)	−0.92 (1.40)	0.26 (1.37)	0.86 (1.20)	−2.17 (1.27)
18.3	User's Emotion Presence	1.51 (0.90)	−0.38 (1.15)	−0.33 (1.27)	0.30 (1.03)	1.18 (1.20)	0.13 (0.96)	−0.40 (1.35)
19	User Agent Interplay	1.30 (1.22)	0.13 (1.09)	0.25 (0.91)	0.69 (0.95)	1.02 (1.12)	0.61 (0.89)	0.17 (1.07)
	ASAQ Score:	19	10	12	9	18	18	7

Table D.4
The ASAQ constructs or dimension scores of DOG, EFFIE, FISH, FURBY, FURHAT, GEMINOID, HAL 9000 and iCAT based on the **long** version of ASAQ. The mean (standard deviation) values of each construct and dimension are calculated based on the mean scores of the construct's or dimension's items.

No.	Construct/Dimension	DOG	EFFIE	FISH	FURBY	FURHAT	GEMINOID	HAL9000	iCAT
		n =39	*n* = 34	*n* = 37	*n* = 39	*n* = 34	*n* = 35	*n* = 37	*n* = 36
1.1	Human-Like Appearance	−1.91 (1.28)	0.62 (1.68)	−1.87 (1.24)	−2.14 (1.22)	−0.76 (1.77)	2.09 (1.21)	−2.08 (1.15)	−2.21 (1.23)
1.2	Human-Like Behaviour	−0.62 (1.47)	−0.19 (1.65)	−0.16 (1.61)	−0.14 (1.71)	−0.29 (1.53)	0.14 (1.26)	0.23 (1.49)	−1.28 (1.31)
1.3	Natural Appearance	1.70 (0.82)	−0.35 (1.51)	1.85 (1.06)	−0.85 (1.28)	−0.82 (1.40)	0.75 (1.07)	−1.20 (1.16)	−1.50 (1.07)
1.4	Natural Behaviour	1.89 (1.00)	−1.41 (1.37)	2.21 (0.84)	−0.61 (1.41)	−1.09 (1.19)	−0.74 (1.37)	−0.36 (1.35)	−1.88 (0.97)
1.5	Agent's Appearance Suit.	1.85 (0.96)	0.67 (1.49)	2.14 (0.86)	1.80 (1.14)	0.79 (1.68)	1.90 (0.87)	1.39 (0.99)	0.49 (1.49)
2	Agent's Usability	0.42 (1.07)	0.85 (1.44)	0.81 (1.26)	1.92 (0.98)	1.71 (1.00)	1.27 (1.13)	1.39 (1.00)	1.19 (1.25)
3	Performance	1.28 (1.50)	0.58 (1.33)	1.69 (0.76)	1.16 (1.07)	1.50 (0.79)	1.37 (0.98)	1.18 (1.29)	0.94 (1.04)
4	Agent's Likeability	2.31 (0.79)	−0.32 (1.51)	2.10 (0.74)	0.91 (1.59)	0.69 (1.26)	0.87 (1.40)	0.36 (1.45)	−0.57 (1.28)
5	Agent's Sociability	1.48 (1.04)	−0.53 (1.52)	1.22 (0.82)	0.73 (1.43)	0.15 (1.21)	0.59 (1.20)	0.05 (1.39)	−0.35 (1.56)
6.1	Agent's Personality Pr.	1.47 (1.04)	−0.60 (1.42)	1.39 (0.86)	0.74 (1.43)	−0.54 (1.04)	−0.13 (1.54)	0.29 (1.34)	−0.52 (1.45)
7	User Acceptance of the A.	1.99 (0.89)	0.03 (1.35)	1.58 (0.93)	1.02 (1.34)	1.05 (1.48)	1.10 (1.13)	1.36 (1.11)	0.66 (1.27)
8	Agent's Enjoyability	2.29 (0.76)	0.06 (1.46)	2.32 (0.75)	1.44 (1.37)	1.23 (1.04)	1.29 (1.31)	0.86 (1.26)	0.47 (1.18)
9	User's Engagement	2.33 (0.60)	0.81 (1.49)	1.78 (0.93)	1.50 (1.11)	2.01 (0.81)	1.72 (0.94)	2.12 (1.02)	1.67 (1.02)
10	User's Trust	0.21 (1.12)	0.11 (1.12)	0.05 (1.10)	−0.46 (1.37)	1.04 (1.00)	0.30 (1.37)	0.41 (1.40)	−0.03 (1.01)
11	User Agent Alliance	1.03 (1.17)	−0.17 (1.33)	1.14 (0.83)	−0.06 (1.32)	0.83 (0.95)	0.59 (0.92)	0.84 (1.08)	−0.26 (1.08)
12	Agent's Attentiveness	1.91 (1.18)	0.53 (1.48)	1.98 (0.94)	1.14 (1.30)	2.13 (0.86)	1.75 (0.90)	2.10 (0.79)	1.38 (1.23)
13	Agent's Coherence	1.07 (1.10)	0.68 (1.43)	1.74 (1.16)	1.41 (1.27)	2.04 (0.82)	1.25 (0.98)	1.61 (1.21)	1.68 (0.90)
14	Agent's Intentionality	0.21 (1.05)	−0.11 (1.38)	1.42 (0.93)	−0.49 (1.32)	0.76 (1.14)	0.52 (1.53)	1.80 (0.98)	0.07 (1.22)
15	Attitude	2.50 (0.82)	0.31 (1.26)	2.32 (0.82)	1.56 (1.28)	1.46 (1.40)	1.28 (1.04)	0.41 (1.71)	1.10 (1.43)
16	Social Presence	0.48 (1.05)	−0.97 (1.48)	0.54 (1.22)	−0.12 (1.75)	−0.27 (1.38)	−0.02 (1.20)	−0.26 (1.17)	−1.12 (1.35)
17	Interaction Impact on Self.	1.41 (1.04)	−0.15 (1.26)	1.22 (1.05)	0.63 (1.33)	0.29 (0.95)	0.54 (1.02)	0.75 (1.15)	−0.08 (1.03)
18.1	Agent's Emotional Int. Pr.	1.10 (1.27)	−1.21 (1.69)	1.15 (0.95)	−0.15 (1.76)	−1.48 (1.15)	−1.18 (1.39)	−1.06 (1.33)	−1.84 (1.27)
18.3	User's Emotion Presence	1.51 (0.91)	−0.19 (1.43)	1.55 (1.02)	1.03 (1.24)	0.81 (1.24)	0.06 (1.22)	0.80 (1.21)	0.26 (1.22)
19	User Agent Interplay	1.58 (1.08)	−0.18 (1.41)	1.68 (0.71)	1.03 (1.10)	0.55 (1.07)	0.55 (1.07)	0.68 (1.00)	0.06 (1.24)
	ASAQ Score:	29	−1	32	13	14	18	14	−2

Table D.5
The ASAQ constructs or dimension scores of KITT, LOLA, MARCUS, NAO, PARO, POPPY, and ROBOT BOSS based on the **long** version of ASAQ. The mean (standard deviation) values of each construct and dimension are calculated based on the mean scores of the construct's or dimension's items.

No.	Construct/Dimension	KITT	LOLA	MARCUS	NAO	PARO	POPPY	ROBOT BOSS
		n = 36	n = 37	n = 36	n = 36	n = 37	n = 38	n = 34
1.1	Human-Like Appearance	−2.01 (1.31)	1.14 (1.55)	1.69 (1.13)	−0.93 (1.55)	−2.41 (1.02)	0.64 (1.42)	−1.18 (1.58)
1.2	Human-Like Behaviour	0.52 (1.38)	0.58 (1.52)	1.70 (0.97)	0.74 (1.25)	−1.09 (1.25)	0.51 (1.50)	−0.06 (1.42)
1.3	Natural Appearance	−0.94 (1.31)	0.62 (1.21)	0.67 (1.21)	−0.57 (1.33)	0.38 (1.35)	−0.31 (1.35)	−1.24 (1.12)
1.4	Natural Behaviour	−0.22 (1.37)	−0.57 (1.25)	1.21 (1.13)	−0.36 (1.51)	−0.48 (1.16)	−0.38 (1.37)	−0.68 (1.46)
1.5	Agent's Appearance Suit.	1.31 (0.98)	1.45 (1.24)	1.41 (1.02)	1.74 (0.94)	1.99 (0.89)	0.70 (1.21)	0.24 (1.66)
2	Agent's Usability	1.60 (1.18)	1.75 (1.18)	0.65 (1.25)	1.00 (1.15)	1.77 (0.93)	0.84 (1.53)	1.25 (1.43)
3	Performance	1.11 (1.17)	1.22 (1.10)	1.09 (0.87)	1.31 (1.01)	1.62 (1.00)	0.94 (1.19)	0.41 (1.28)
4	Agent's Likeability	1.21 (1.22)	0.59 (1.55)	0.41 (1.09)	1.74 (1.33)	1.57 (1.22)	0.22 (1.24)	−0.71 (1.45)
5	Agent's Sociability	0.44 (1.42)	0.45 (1.32)	0.75 (1.07)	0.97 (1.13)	1.25 (1.00)	0.75 (1.24)	−0.25 (1.59)
6.1	Agent's Personality Pr.	0.89 (1.39)	−0.20 (1.40)	1.35 (1.10)	1.06 (1.26)	0.04 (1.29)	−0.04 (1.22)	0.05 (1.38)
7	User Acceptance of the A.	0.92 (1.44)	0.46 (1.25)	0.76 (1.09)	1.57 (1.23)	2.05 (0.95)	1.35 (1.23)	0.01 (1.15)
8	Agent's Enjoyability	1.34 (0.99)	0.64 (1.47)	0.71 (1.19)	1.91 (1.06)	2.03 (0.93)	1.03 (1.26)	−0.29 (1.48)
9	User's Engagement	1.44 (1.08)	1.28 (1.12)	2.08 (0.89)	2.11 (0.73)	1.84 (0.93)	1.99 (1.04)	0.67 (1.68)
10	User's Trust	0.99 (1.36)	1.08 (1.06)	0.31 (0.97)	0.51 (1.17)	−0.04 (0.96)	0.32 (1.15)	0.28 (1.17)
11	User Agent Alliance	0.98 (1.04)	0.68 (1.11)	1.07 (0.84)	0.94 (1.14)	0.56 (0.99)	0.22 (1.13)	0.15 (1.37)
12	Agent's Attentiveness	1.72 (1.08)	2.22 (0.82)	1.96 (0.94)	1.59 (1.16)	1.40 (1.11)	1.50 (1.24)	0.84 (1.24)
13	Agent's Coherence	1.76 (1.06)	1.95 (0.90)	1.04 (1.08)	1.61 (0.86)	1.46 (1.02)	1.22 (1.34)	0.85 (1.00)
14	Agent's Intentionality	1.50 (0.98)	0.68 (1.40)	1.63 (0.96)	0.82 (1.05)	−0.66 (1.14)	0.44 (1.24)	0.51 (1.40)
15	Attitude	0.21 (1.48)	0.33 (1.12)	−0.01 (1.24)	1.90 (0.99)	2.41 (0.93)	1.68 (1.36)	−0.63 (1.52)
16	Social Presence	−0.41 (1.37)	0.05 (1.19)	0.75 (1.50)	0.40 (1.46)	0.23 (0.91)	0.04 (1.45)	−0.59 (1.37)
17	Interaction Impact on Self.	0.80 (1.27)	−0.06 (1.22)	0.24 (0.84)	1.12 (1.34)	1.45 (1.09)	0.21 (1.18)	−0.68 (1.47)
18.1	Agent's Emotional Int. Pr.	−1.07 (1.67)	−1.42 (1.32)	0.78 (1.36)	0.07 (1.80)	−0.87 (1.40)	−0.73 (1.60)	−0.26 (1.54)
18.3	User's Emotion Presence	0.72 (1.12)	0.33 (1.14)	1.24 (0.93)	0.99 (1.05)	1.88 (0.92)	0.49 (1.19)	0.49 (1.44)
19	User Agent Interplay	1.06 (1.06)	0.54 (0.97)	1.40 (0.88)	1.14 (1.27)	0.91 (1.00)	0.68 (1.34)	0.68 (1.14)
	ASAQ Score:	16	16	25	23	19	14	0

Table D.6
The ASAQ constructs or dimension scores of SAMANTHA, SARAH, SIM SENSEI, SIRI, THE AMBIENT LIGHT TV, THE NEGOTIATOR, and ZOMBIE based on the **long** version of ASAQ. The mean (standard deviation) values of each construct and dimension are calculated based on the mean scores of the construct's or dimension's items.

No.	Construct/Dimension	SAMANTHA	SARAH	SIM SENSEI	SIRI	AMBIENT TV	NEGOTIATOR	ZOMBIE
		n = 36	n = 39	n = 38	n = 39	n = 35	n = 36	n = 35
1.1	Human-Like Appearance	−1.31 (1.69)	1.66 (1.14)	1.28 (1.42)	−1.76 (1.23)	−2.19 (1.07)	0.99 (1.35)	1.79 (1.25)
1.2	Human-Like Behaviour	1.58 (1.18)	0.71 (1.21)	0.92 (1.40)	−0.28 (1.42)	−1.37 (1.17)	0.65 (1.26)	0.52 (1.12)
1.3	Natural Appearance	−0.41 (1.43)	0.87 (1.21)	0.61 (1.29)	−0.70 (1.07)	−0.84 (1.18)	0.39 (1.22)	0.41 (1.11)
1.4	Natural Behaviour	0.42 (1.22)	−0.32 (1.35)	−0.36 (1.10)	−0.81 (1.08)	−0.96 (1.39)	−0.35 (1.40)	0.75 (1.23)
1.5	Agent's Appearance Suit.	0.35 (1.32)	1.74 (0.92)	1.55 (1.29)	0.80 (1.23)	1.81 (1.05)	1.53 (0.92)	1.01 (1.27)
2	Agent's Usability	1.99 (0.92)	1.71 (1.10)	1.25 (1.01)	2.21 (0.73)	1.29 (1.05)	0.89 (1.31)	−0.62 (1.21)
3	Performance	1.54 (0.88)	1.72 (0.88)	1.04 (1.21)	2.07 (0.90)	0.95 (1.17)	0.53 (1.11)	0.18 (0.92)
4	Agent's Likeability	1.38 (1.20)	0.77 (1.18)	0.61 (1.27)	0.77 (1.08)	1.06 (1.17)	0.04 (1.19)	−0.60 (1.37)
5	Agent's Sociability	1.51 (1.20)	0.04 (1.37)	0.13 (1.43)	−0.14 (1.59)	−0.95 (1.14)	−0.24 (1.35)	−0.64 (1.31)
6.1	Agent's Personality Pr.	0.41 (1.27)	−0.17 (1.34)	0.01 (1.17)	−0.64 (1.32)	−0.80 (1.44)	0.08 (1.23)	1.04 (1.18)
7	User Acceptance of the A.	1.44 (1.22)	1.51 (0.96)	0.74 (1.20)	2.13 (0.78)	1.33 (1.26)	0.52 (1.06)	−0.26 (1.50)
8	Agent's Enjoyability	1.86 (0.94)	1.11 (1.08)	0.86 (1.06)	1.53 (1.04)	1.63 (1.16)	0.63 (1.31)	−0.35 (1.17)
9	User's Engagement	2.25 (0.65)	1.67 (0.70)	1.04 (1.35)	1.56 (0.99)	1.56 (1.32)	2.25 (0.75)	2.18 (0.86)
10	User's Trust	0.79 (1.02)	1.30 (0.86)	0.48 (1.02)	1.17 (1.04)	−0.30 (1.38)	0.40 (1.13)	−0.41 (1.18)
11	User Agent Alliance	0.77 (1.01)	0.79 (0.82)	0.32 (1.00)	0.57 (1.18)	−0.20 (1.46)	0.35 (1.05)	−0.14 (1.22)
12	Agent's Attentiveness	2.16 (0.90)	2.04 (0.91)	1.76 (1.14)	1.79 (0.95)	0.25 (1.60)	1.40 (1.22)	1.87 (0.91)
13	Agent's Coherence	1.87 (0.98)	1.92 (0.84)	1.58 (1.00)	2.08 (0.80)	1.56 (1.12)	1.48 (1.22)	0.20 (1.04)
14	Agent's Intentionality	1.13 (1.20)	1.36 (0.89)	0.85 (1.15)	0.79 (1.05)	−0.06 (1.51)	1.49 (0.99)	1.16 (1.29)
15	Attitude	1.41 (1.35)	1.73 (0.85)	0.92 (1.07)	2.03 (0.80)	1.67 (1.25)	0.11 (1.35)	−0.54 (1.52)
16	Social Presence	0.10 (1.21)	0.06 (1.32)	−0.06 (1.50)	−0.99 (1.40)	−1.43 (1.25)	−0.32 (1.25)	−0.18 (1.20)
17	Interaction Impact on Self.	0.90 (1.20)	0.78 (0.85)	0.36 (1.03)	0.93 (1.03)	0.93 (1.25)	0.09 (0.88)	−0.68 (1.34)
18.1	Agent's Emotional Int. Pr.	−0.06 (1.60)	−1.36 (1.47)	−0.65 (1.24)	−1.91 (1.11)	−1.50 (1.41)	−1.11 (1.33)	0.35 (1.41)
18.3	User's Emotion Presence	1.72 (0.70)	0.15 (1.03)	0.36 (1.01)	−0.62 (1.34)	0.44 (1.42)	0.98 (1.24)	1.76 (0.90)
19	User Agent Interplay	1.22 (1.23)	0.40 (1.08)	0.95 (1.15)	0.05 (1.05)	−0.51 (1.55)	0.80 (0.79)	0.86 (1.00)
	ASAQ Score:	25	22	17	13	3	14	10

Table D.7
The ASAQ constructs or dimension scores of AIBO, ALEXA, ALICE, AMY, CHAPPIE, C3PO, and DEEPBLUE based on the **short** version of ASAQ. The mean (standard deviation) values of each construct and dimension are calculated based on the mean scores of the construct's or dimension's items.

No.	Construct/Dimension	AIBO	ALEXA	ALICE	AMY	CHAPPIE	C3PO	DEEPBLUE
		n = 39	*n* = 37	*n* = 35	*n* = 39	*n* = 38	*n* = 36	*n* = 39
1.1	Human-Like Appearance	−2.59 (0.88)	−2.70 (0.81)	1.34 (1.68)	1.13 (1.67)	−1.21 (1.76)	−0.19 (2.10)	−2.10 (1.60)
1.2	Human-Like Behaviour	−1.77 (1.42)	0.03 (1.80)	0.77 (1.42)	0.44 (1.82)	0.42 (1.69)	0.64 (1.84)	−0.90 (2.05)
1.3	Natural Appearance	−0.56 (2.21)	−1.03 (1.98)	−0.20 (1.62)	−0.46 (1.82)	−1.13 (1.60)	−0.97 (1.78)	−1.15 (1.83)
1.4	Natural Behaviour	0.69 (1.88)	−0.89 (1.87)	−0.40 (2.05)	−0.18 (1.92)	0.89 (1.59)	0.33 (1.62)	−0.87 (2.30)
1.5	Agent's Appearance Suit.	2.00 (1.19)	1.49 (1.66)	1.51 (1.67)	1.13 (1.59)	0.87 (1.70)	1.19 (1.60)	1.18 (1.54)
2	Agent's Usability	1.95 (0.86)	2.22 (0.98)	1.60 (1.26)	1.51 (1.10)	0.71 (1.37)	1.28 (1.56)	1.28 (1.41)
3	Performance	1.90 (1.07)	2.30 (0.88)	0.66 (1.85)	0.90 (1.48)	1.29 (1.04)	1.53 (1.34)	2.51 (0.79)
4	Agent's Likeability	1.44 (1.65)	1.30 (1.63)	0.71 (1.62)	−0.05 (1.97)	1.47 (1.50)	1.92 (1.42)	0.85 (1.37)
5	Agent's Sociability	0.54 (1.82)	−1.00 (1.67)	−0.71 (1.93)	−1.05 (1.54)	0.13 (1.47)	−0.31 (1.69)	−1.54 (1.73)
6.1	Agent's Personality Pr.	0.82 (1.65)	0.03 (1.88)	−0.37 (1.90)	0.03 (1.56)	0.68 (1.54)	1.39 (1.71)	−0.56 (1.96)
7	User Acceptance of the A.	1.74 (1.09)	2.05 (1.25)	0.77 (1.55)	0.36 (1.20)	1.71 (1.27)	1.19 (1.56)	1.59 (1.09)
8	Agent's Enjoyability	1.41 (1.45)	0.54 (1.76)	0.03 (1.74)	−0.26 (1.79)	1.47 (1.47)	1.42 (1.70)	0.33 (1.66)
9	User's Engagement	2.31 (1.26)	1.54 (1.24)	1.51 (1.22)	1.10 (1.23)	2.13 (0.91)	0.78 (1.64)	1.69 (1.28)
10	User's Trust	0.21 (1.59)	1.30 (1.24)	0.60 (1.82)	0.33 (1.36)	−0.08 (1.51)	1.17 (1.83)	1.56 (1.65)
11	User Agent Alliance	−0.13 (1.59)	−0.08 (1.64)	−0.37 (1.61)	−0.23 (1.51)	−0.03 (1.53)	1.00 (1.53)	0.46 (1.79)
12	Agent's Attentiveness	1.69 (1.08)	1.62 (1.26)	1.97 (1.29)	1.23 (1.55)	1.11 (1.43)	2.22 (0.93)	1.49 (1.67)
13	Agent's Coherence	1.92 (1.18)	2.22 (1.00)	1.54 (1.44)	1.31 (1.56)	1.45 (1.33)	1.56 (1.42)	2.00 (1.19)
14	Agent's Intentionality	0.59 (2.09)	1.14 (1.90)	1.11 (1.79)	0.82 (1.68)	0.53 (1.93)	1.78 (1.53)	1.72 (1.90)
15	Attitude	2.33 (0.77)	1.51 (1.35)	0.97 (1.32)	0.44 (1.25)	2.00 (1.29)	0.00 (1.93)	1.38 (1.46)
16	Social Presence	−0.13 (2.00)	−0.81 (1.85)	−1.14 (1.48)	−0.18 (1.86)	0.42 (1.59)	−0.11 (1.75)	−1.38 (1.90)
17	Interaction Impact on Self.	1.28 (1.45)	1.22 (1.40)	0.57 (1.31)	0.51 (1.39)	1.21 (1.34)	0.94 (1.49)	1.10 (1.59)
18.1	Agent's Emotional Int. Pr.	0.05 (2.26)	−1.49 (1.85)	−0.86 (1.96)	−0.36 (2.05)	0.53 (1.77)	1.14 (1.59)	−2.41 (1.14)
18.3	User's Emotion Presence	1.82 (1.10)	−0.27 (1.64)	−0.34 (1.53)	0.03 (1.46)	1.55 (1.52)	−0.53 (1.83)	−0.03 (2.08)
19	User Agent Interplay	1.33 (1.56)	−0.65 (1.62)	−0.86 (1.80)	0.38 (1.33)	0.79 (1.32)	0.31 (1.58)	−1.00 (1.86)
	ASAQ Score:	21	12	10	9	19	20	7

Table D.8
The ASAQ constructs or dimension scores of DOG, EFFIE, FISH, FURBY, FURHAT, GEMINOID, HAL 9000 and iCAT based on the **short** version of ASAQ. The mean (standard deviation) values of each construct and dimension are calculated based on the mean scores of the construct's or dimension's items.

No.	Construct/Dimension	DOG	EFFIE	FISH	FURBY	FURHAT	GEMINOID	HAL9000	iCAT
		n = 39	*n* = 34	*n* = 37	*n* = 39	*n* = 34	*n* = 35	*n* = 37	*n* = 36
1.1	Human-Like Appearance	−2.13 (1.36)	0.85 (2.03)	−2.19 (1.56)	−2.69 (0.80)	−0.65 (2.01)	2.20 (1.39)	−2.14 (1.58)	−2.19 (1.51)
1.2	Human-Like Behaviour	−0.51 (2.00)	0.18 (1.82)	−0.41 (2.03)	0.00 (2.20)	−0.35 (1.67)	0.77 (1.52)	0.54 (1.97)	−1.22 (1.81)
1.3	Natural Appearance	2.10 (1.39)	−0.56 (2.05)	2.16 (1.01)	−1.10 (1.92)	−0.97 (1.88)	0.71 (1.89)	−1.62 (1.69)	−1.81 (1.62)
1.4	Natural Behaviour	2.28 (0.92)	−1.00 (1.71)	2.57 (0.87)	0.21 (2.00)	−0.82 (1.80)	−0.09 (1.69)	0.00 (2.00)	−1.64 (1.73)
1.5	Agent's Appearance Suit.	2.28 (1.23)	0.44 (1.91)	2.14 (1.23)	1.87 (1.26)	0.74 (1.91)	1.94 (0.97)	1.70 (1.10)	0.64 (1.66)
2	Agent's Usability	0.74 (1.53)	1.03 (1.45)	1.24 (1.40)	2.18 (1.21)	1.74 (1.05)	1.43 (1.29)	1.65 (1.16)	1.25 (1.52)
3	Performance	1.23 (1.80)	0.71 (1.51)	1.62 (1.11)	1.51 (1.23)	1.97 (0.94)	1.54 (1.17)	1.43 (1.39)	0.75 (1.42)
4	Agent's Likeability	2.62 (0.75)	−0.38 (2.15)	2.51 (0.84)	0.85 (2.06)	0.97 (1.51)	0.43 (2.08)	0.41 (1.79)	−0.75 (1.75)
5	Agent's Sociability	1.18 (1.30)	−1.09 (1.75)	0.43 (1.32)	0.00 (2.09)	−0.47 (1.46)	−0.40 (1.75)	−0.38 (1.80)	−1.22 (1.91)
6.1	Agent's Personality Pr.	1.21 (1.44)	−0.38 (1.84)	1.43 (1.07)	1.33 (1.74)	−0.15 (1.64)	0.14 (1.78)	0.73 (1.74)	0.28 (1.99)
7	User Acceptance of the A.	2.08 (1.06)	−0.15 (1.64)	1.62 (1.42)	1.00 (1.69)	0.74 (1.78)	1.03 (1.27)	1.92 (1.23)	0.50 (1.48)
8	Agent's Enjoyability	2.03 (1.39)	−0.15 (2.16)	2.27 (1.12)	0.82 (1.99)	1.06 (1.43)	0.89 (1.76)	0.54 (1.92)	−0.50 (1.73)
9	User's Engagement	2.36 (0.87)	0.79 (1.74)	2.08 (1.04)	1.74 (1.31)	1.97 (1.36)	1.57 (1.12)	1.97 (1.28)	1.39 (1.34)
10	User's Trust	1.05 (1.49)	0.24 (1.52)	0.57 (1.64)	−0.59 (1.85)	0.82 (1.38)	0.40 (1.79)	0.76 (1.80)	−0.11 (1.62)
11	User Agent Alliance	0.82 (1.70)	−0.56 (1.46)	0.41 (1.54)	−0.54 (1.85)	0.35 (1.70)	0.00 (1.24)	0.51 (1.98)	−0.86 (1.71)
12	Agent's Attentiveness	1.87 (1.28)	0.29 (1.92)	1.68 (1.25)	0.90 (1.73)	1.82 (1.03)	1.66 (1.14)	2.24 (1.06)	0.92 (1.75)
13	Agent's Coherence	1.33 (1.81)	0.62 (1.97)	1.65 (1.44)	1.13 (1.73)	2.09 (1.06)	1.57 (1.31)	1.27 (1.50)	1.31 (1.58)
14	Agent's Intentionality	−0.54 (1.82)	0.09 (2.09)	1.57 (1.26)	−0.62 (2.21)	1.35 (1.76)	0.97 (1.71)	2.30 (1.08)	−0.08 (1.87)
15	Attitude	2.49 (0.97)	0.47 (1.50)	2.19 (1.31)	1.54 (1.50)	1.41 (1.60)	1.26 (1.24)	0.49 (1.77)	1.11 (1.53)
16	Social Presence	1.03 (1.61)	−1.09 (1.86)	0.81 (1.41)	0.00 (2.27)	−0.26 (1.54)	0.20 (1.45)	−0.14 (1.51)	−1.14 (1.90)
17	Interaction Impact on Self.	1.31 (1.42)	−0.03 (1.82)	0.76 (1.42)	0.90 (1.65)	0.56 (1.56)	0.69 (1.23)	0.95 (1.51)	−0.06 (1.35)
18.1	Agent's Emotional Int. Pr.	1.51 (1.54)	−0.94 (1.98)	1.43 (1.63)	0.79 (2.09)	−1.26 (1.69)	−1.29 (1.62)	−0.92 (1.79)	−1.83 (1.54)
18.3	User's Emotion Presence	2.26 (0.97)	−0.15 (1.88)	1.70 (1.08)	1.10 (1.73)	1.00 (1.81)	0.23 (1.65)	0.65 (1.64)	0.44 (1.46)
19	User Agent Interplay	1.26 (1.58)	−0.79 (1.57)	1.38 (1.01)	1.05 (1.82)	−0.03 (1.77)	0.17 (1.38)	0.14 (1.46)	−0.22 (1.73)
	ASAQ Score:	32	−2	32	13	14	18	15	−5

Table D.9
The ASAQ constructs or dimension scores of KITT, LOLA, MARCUS, NAO, PARO, POPPY, and ROBOT BOSS based on the **short** version of ASAQ. The mean (standard deviation) values of each construct and dimension are calculated based on the mean scores of the construct's or dimension's items.

No.	Construct/Dimension	KITT	LOLA	MARCUS	NAO	PARO	POPPY	ROBOT BOSS
		n = 36	*n* = 37	*n* = 36	*n* = 36	*n* = 37	*n* = 38	*n* = 34
1.1	Human-Like Appearance	−2.28 (1.58)	1.11 (1.76)	1.67 (1.45)	−1.17 (1.75)	−2.62 (1.01)	0.95 (1.54)	−0.97 (1.99)
1.2	Human-Like Behaviour	1.06 (1.51)	0.65 (1.64)	1.97 (1.13)	0.97 (1.59)	−1.59 (1.74)	0.87 (1.80)	0.12 (1.89)
1.3	Natural Appearance	−0.92 (2.13)	0.49 (1.84)	0.61 (2.11)	−0.89 (1.89)	0.54 (1.79)	−0.18 (1.72)	−1.24 (1.69)
1.4	Natural Behaviour	0.53 (1.86)	0.14 (1.93)	1.67 (1.24)	0.50 (2.05)	0.54 (1.61)	0.45 (1.81)	−0.47 (1.81)
1.5	Agent's Appearance Suit.	1.28 (1.50)	1.62 (1.40)	0.97 (1.52)	2.00 (1.15)	1.95 (1.56)	0.87 (1.55)	0.24 (1.79)
2	Agent's Usability	1.83 (1.30)	1.92 (1.14)	0.25 (1.68)	1.06 (1.15)	2.05 (1.03)	0.68 (1.86)	1.06 (1.76)
3	Performance	1.64 (1.29)	1.54 (1.46)	1.42 (1.08)	0.69 (1.60)	1.76 (1.42)	0.92 (1.65)	0.88 (1.79)
4	Agent's Likeability	1.72 (1.45)	0.41 (2.15)	0.58 (1.57)	2.00 (1.60)	1.62 (1.53)	0.13 (1.77)	−0.65 (2.12)
5	Agent's Sociability	−0.25 (2.01)	−0.19 (1.49)	0.08 (1.68)	0.50 (1.48)	0.84 (1.57)	0.42 (1.65)	−0.62 (1.94)
6.1	Agent's Personality Pr.	1.42 (1.46)	0.30 (1.76)	1.56 (1.32)	1.31 (1.33)	0.73 (1.71)	0.16 (1.42)	0.21 (1.79)
7	User Acceptance of the A.	1.50 (1.87)	0.35 (1.69)	0.72 (1.49)	1.56 (1.56)	2.03 (1.19)	1.53 (1.45)	0.47 (1.71)
8	Agent's Enjoyability	1.69 (1.12)	0.27 (1.81)	1.31 (1.88)	1.75 (1.61)	1.30 (1.70)	−0.13 (2.20)	−0.24 (1.88)
9	User's Engagement	1.61 (1.36)	1.32 (1.27)	2.06 (1.04)	1.78 (1.12)	2.27 (0.90)	1.84 (1.48)	1.44 (1.85)
10	User's Trust	1.31 (1.56)	0.84 (1.59)	0.42 (1.18)	0.17 (1.48)	0.68 (1.65)	0.18 (1.57)	0.32 (1.82)
11	User Agent Alliance	0.83 (1.48)	−0.22 (1.69)	1.14 (1.33)	0.67 (1.53)	−0.11 (1.74)	−0.26 (1.50)	−0.41 (1.83)
12	Agent's Attentiveness	1.75 (1.30)	1.95 (1.05)	1.78 (1.10)	1.56 (1.27)	0.95 (1.39)	1.00 (1.61)	0.59 (1.60)
13	Agent's Coherence	1.69 (1.39)	1.76 (1.32)	1.17 (1.48)	2.00 (1.15)	1.73 (1.17)	1.18 (1.75)	0.91 (1.68)
14	Agent's Intentionality	1.69 (1.49)	1.11 (1.93)	1.83 (1.30)	1.14 (1.69)	−0.38 (1.38)	0.61 (1.76)	0.79 (1.82)
15	Attitude	0.67 (1.79)	0.24 (1.59)	−0.06 (1.47)	2.03 (1.11)	2.49 (1.02)	1.63 (1.57)	−0.59 (1.78)
16	Social Presence	−0.19 (1.69)	−0.08 (1.86)	0.53 (1.65)	0.50 (1.90)	0.65 (1.38)	0.05 (1.68)	−0.50 (1.81)
17	Interaction Impact on Self.	1.03 (1.75)	0.49 (1.66)	0.33 (1.51)	1.39 (1.42)	1.65 (1.48)	0.37 (1.76)	−0.32 (1.87)
18.1	Agent's Emotional Int. Pr.	−1.28 (1.80)	−1.22 (1.58)	0.69 (1.88)	0.47 (2.25)	−0.30 (1.98)	−0.21 (1.85)	−0.06 (2.01)
18.3	User's Emotion Presence	0.78 (1.84)	0.54 (1.71)	1.39 (1.38)	1.00 (1.47)	2.05 (0.97)	0.61 (1.62)	0.62 (1.81)
19	User Agent Interplay	0.56 (1.78)	−0.38 (1.53)	1.22 (1.42)	1.19 (1.62)	0.62 (1.57)	0.37 (1.73)	0.62 (1.72)
	ASAQ Score:	20	15	25	24	21	14	2

Table D.10
The ASAQ constructs or dimension scores of SAMANTHA, SARAH, SIM SENSEI, SIRI, THE AMBIENT LIGHT TV, THE NEGOTIATOR, and ZOMBIE based on the **short** version of ASAQ. The mean (standard deviation) values of each construct and dimension are calculated based on the mean scores of the construct's or dimension's items.

No.	Construct/Dimension	SAMANTHA	SARAH	SIM SENSEI	SIRI	AMBIENT TV	NEGOTIATOR	ZOMBIE
		n = 36	*n* = 39	*n* = 38	*n* = 39	*n* = 35	*n* = 36	*n* = 35
1.1	Human-Like Appearance	−1.75 (1.86)	1.87 (1.26)	1.42 (1.57)	−1.72 (1.61)	−2.20 (1.57)	0.81 (1.79)	2.00 (1.35)
1.2	Human-Like Behaviour	1.78 (1.40)	0.87 (1.40)	1.21 (1.61)	−0.03 (1.77)	−1.43 (1.72)	0.72 (1.54)	1.37 (1.35)
1.3	Natural Appearance	−0.42 (2.41)	1.00 (1.56)	0.71 (1.74)	−0.28 (1.65)	−0.60 (1.82)	0.33 (1.77)	0.29 (1.58)
1.4	Natural Behaviour	1.11 (1.94)	0.36 (1.75)	0.50 (1.80)	−0.46 (1.85)	−0.86 (2.03)	0.25 (1.79)	1.74 (1.31)
1.5	Agent's Appearance Suit.	0.42 (1.89)	1.85 (1.01)	1.66 (1.49)	0.90 (1.68)	2.03 (0.89)	1.72 (1.26)	0.60 (1.72)
2	Agent's Usability	2.22 (0.96)	1.72 (1.30)	1.45 (1.33)	2.26 (0.88)	2.00 (1.19)	1.03 (1.50)	−0.34 (1.53)
3	Performance	1.97 (1.16)	1.87 (1.10)	1.26 (1.37)	2.28 (0.89)	2.06 (1.30)	1.17 (1.44)	0.63 (1.21)
4	Agent's Likeability	1.72 (1.34)	0.59 (1.73)	0.53 (1.57)	1.38 (1.29)	2.00 (1.28)	0.03 (1.61)	−0.54 (2.06)
5	Agent's Sociability	1.06 (1.67)	−0.46 (1.76)	−0.26 (1.73)	−0.87 (2.10)	−1.69 (1.51)	−0.64 (1.57)	−1.11 (1.83)
6.1	Agent's Personality Pr.	1.03 (1.72)	0.18 (1.37)	0.32 (1.56)	0.33 (1.59)	−0.49 (2.21)	0.33 (1.47)	0.86 (1.77)
7	User Acceptance of the A.	1.89 (1.28)	1.31 (1.15)	0.58 (1.33)	2.36 (0.96)	1.57 (1.48)	0.56 (1.08)	−0.20 (1.80)
8	Agent's Enjoyability	1.75 (1.36)	0.41 (1.85)	0.66 (1.76)	0.97 (1.74)	1.86 (1.44)	0.67 (1.84)	1.00 (1.39)
9	User's Engagement	2.36 (0.83)	1.59 (0.91)	1.11 (1.27)	1.38 (1.33)	1.74 (1.54)	2.22 (0.93)	2.20 (0.90)
10	User's Trust	0.86 (1.48)	1.49 (1.00)	0.37 (1.75)	1.49 (1.05)	0.66 (1.81)	0.25 (1.40)	−0.37 (1.68)
11	User Agent Alliance	−0.06 (1.62)	0.41 (1.35)	−0.79 (1.23)	0.03 (1.72)	−0.77 (1.85)	−0.28 (1.65)	−0.80 (1.71)
12	Agent's Attentiveness	1.75 (1.11)	1.79 (1.15)	1.61 (1.31)	1.36 (1.58)	0.17 (1.98)	1.50 (1.38)	1.29 (1.23)
13	Agent's Coherence	2.25 (0.97)	2.08 (1.09)	1.61 (1.35)	2.10 (1.14)	1.69 (1.53)	1.14 (1.78)	−0.09 (1.56)
14	Agent's Intentionality	1.53 (1.58)	1.69 (1.28)	1.24 (1.75)	1.87 (1.40)	0.69 (1.79)	1.42 (1.42)	1.17 (1.52)
15	Attitude	1.47 (1.44)	1.62 (1.04)	0.84 (1.42)	2.05 (0.97)	1.46 (1.40)	0.03 (1.30)	−0.40 (1.70)
16	Social Presence	0.58 (1.65)	0.03 (1.71)	0.05 (1.80)	−0.95 (1.78)	−1.69 (1.66)	−0.72 (1.65)	−0.23 (1.55)
17	Interaction Impact on Self.	1.17 (1.58)	1.08 (1.18)	0.71 (1.52)	1.28 (1.30)	1.34 (1.55)	0.39 (1.46)	−0.51 (1.90)
18.1	Agent's Emotional Int. Pr.	0.42 (2.06)	−0.90 (1.82)	0.16 (1.60)	−1.67 (1.56)	−1.26 (1.74)	−0.67 (1.96)	0.23 (2.06)
18.3	User's Emotion Presence	1.69 (1.37)	0.05 (1.62)	−0.45 (1.74)	−0.64 (1.84)	0.66 (1.85)	1.17 (1.83)	2.31 (0.93)
19	User Agent Interplay	1.25 (1.20)	−0.10 (1.19)	0.79 (1.61)	−1.03 (1.56)	−0.66 (1.88)	0.81 (1.51)	1.31 (1.18)
	ASAQ Score:	28	22	17	14	8	14	12

Table D.11
The percentile scores of the ASAQ constructs/dimensions based on the **long** version of the ASAQ representative set 2024 ($n = 1066$).

No.	Construct/Dimension	Percentile												
		5%	10%	20%	25%	30%	40%	50%	60%	70%	75%	80%	90%	95%
1.1	Human-Like Appearance	−2.45	−2.25	−2.11	−2.01	−1.94	−1.67	−1.09	−0.40	0.78	0.96	1.05	1.67	1.75
1.2	Human-Like Behaviour	−1.34	−1.13	−0.42	−0.29	−0.24	−0.12	0.14	0.29	0.51	0.52	0.61	0.78	1.31
1.3	Natural Appearance	−1.28	−1.21	−0.93	−0.85	−0.83	−0.63	−0.35	−0.02	0.39	0.41	0.61	0.78	1.37
1.4	Natural Behaviour	−1.30	−1.10	−1.00	−0.85	−0.78	−0.60	−0.38	−0.36	−0.31	−0.22	0.09	0.84	1.62
1.5	Agent’s Appearance Suit.	0.41	0.63	0.80	0.95	1.04	1.23	1.41	1.51	1.69	1.74	1.81	1.92	2.03
2	Agent’s Usability	0.51	0.73	0.83	0.85	0.93	1.21	1.27	1.37	1.67	1.71	1.76	1.94	2.01
3	Performance	0.46	0.57	0.94	0.95	1.04	1.12	1.16	1.27	1.45	1.54	1.58	1.73	1.87
4	Agent’s Likeability	−0.59	−0.37	0.16	0.33	0.38	0.63	0.77	0.85	1.15	1.27	1.41	1.61	1.96
5	Agent’s Sociability	−0.83	−0.55	−0.26	−0.24	−0.09	0.04	0.15	0.45	0.69	0.75	0.83	1.22	1.39
6.1	Agent’s Personality Pr.	−0.74	−0.64	−0.53	−0.52	−0.35	−0.12	0.04	0.25	0.61	0.77	0.95	1.32	1.37
7	User Acceptance of the A.	0.02	0.37	0.59	0.66	0.75	0.99	1.10	1.35	1.42	1.51	1.58	1.95	2.02
8	Agent’s Enjoyability	−0.15	0.26	0.64	0.71	0.86	1.04	1.18	1.30	1.49	1.63	1.84	1.97	2.19
9	User’s Engagement	0.91	1.23	1.44	1.50	1.55	1.67	1.78	1.91	2.02	2.08	2.11	2.20	2.25
10	User’s Trust	−0.37	−0.13	0.02	0.07	0.15	0.30	0.40	0.50	0.77	0.99	1.02	1.17	1.19
11	User Agent Alliance	−0.18	−0.15	0.13	0.22	0.33	0.50	0.59	0.65	0.74	0.79	0.83	0.99	1.05
12	Agent’s Attentiveness	0.65	1.08	1.35	1.40	1.44	1.67	1.75	1.80	1.90	1.96	2.01	2.13	2.15
13	Agent’s Coherence	0.75	1.00	1.16	1.25	1.36	1.48	1.58	1.64	1.75	1.76	1.87	1.97	2.06
14	Agent’s Intentionality	−0.34	−0.07	0.16	0.21	0.38	0.52	0.68	0.81	1.15	1.36	1.41	1.51	1.59
15	Attitude	−0.33	0.02	0.27	0.33	0.38	1.08	1.41	1.55	1.70	1.73	1.88	2.33	2.40
16	Social Presence	−1.18	−1.02	−0.85	−0.59	−0.44	−0.27	−0.18	−0.03	0.04	0.06	0.15	0.41	0.52
17	Interaction Impact on Self.	−0.47	−0.11	0.03	0.16	0.22	0.38	0.63	0.78	0.86	0.90	0.93	1.14	1.33
18.1	Agent’s Emotional Int. Pr.	−1.88	−1.74	−1.46	−1.42	−1.30	−1.10	−0.92	−0.70	−0.20	−0.06	0.14	0.79	1.00
18.3	User’s Emotion Presence	−0.39	−0.34	0.10	0.15	0.28	0.37	0.49	0.81	1.01	1.18	1.35	1.58	1.74
19	User Agent Interplay	−0.09	0.06	0.22	0.40	0.54	0.62	0.68	0.85	0.99	1.03	1.09	1.32	1.51
	Mean:	−0.42	−0.20	0.05	0.14	0.24	0.43	0.60	0.77	1.01	1.10	1.20	1.47	1.65
	SD:	0.85	0.89	0.89	0.87	0.87	0.87	0.80	0.73	0.65	0.64	0.61	0.53	0.46
	Median:	−0.36	−0.12	0.15	0.22	0.36	0.51	0.66	0.81	1.00	1.10	1.22	1.54	1.68
	Min:	−2.45	−2.25	−2.11	−2.01	−1.94	−1.67	−1.09	−0.70	−0.31	−0.22	0.09	0.41	0.52
	Max:	0.91	1.23	1.44	1.50	1.55	1.67	1.78	1.91	2.02	2.08	2.11	2.33	2.40

Table D.12
The percentile scores of the ASAQ constructs/dimensions based on the **short** version of the ASAQ representative set 2024 ($n = 1066$).

No.	Construct/Dimension	Percentile												
		5%	10%	20%	25%	30%	40%	50%	60%	70%	75%	80%	90%	95%
1.1	Human-Like Appearance	−2.66	−2.60	−2.20	−2.19	−2.13	−1.74	−1.17	−0.28	0.91	1.11	1.21	1.71	1.95
1.2	Human-Like Behaviour	−1.53	−1.26	−0.45	−0.35	−0.02	0.13	0.44	0.65	0.77	0.87	0.91	1.24	1.62
1.3	Natural Appearance	−1.47	−1.17	−1.06	−0.97	−0.95	−0.59	−0.46	−0.22	0.31	0.49	0.57	0.77	1.66
1.4	Natural Behaviour	−0.96	−0.88	−0.61	−0.46	−0.31	0.03	0.25	0.43	0.52	0.54	0.77	1.68	2.07
1.5	Agent’s Appearance Suit.	0.43	0.57	0.82	0.87	0.93	1.18	1.49	1.65	1.80	1.87	1.94	2.01	2.09
2	Agent’s Usability	0.42	0.71	1.03	1.06	1.13	1.28	1.45	1.64	1.79	1.92	1.97	2.19	2.22
3	Performance	0.67	0.70	0.89	0.92	1.19	1.31	1.51	1.54	1.71	1.87	1.93	2.10	2.29
4	Agent’s Likeability	−0.61	−0.41	0.09	0.41	0.41	0.58	0.85	1.23	1.46	1.62	1.72	2.00	2.31
5	Agent’s Sociability	−1.41	−1.14	−1.02	−0.87	−0.68	−0.47	−0.38	−0.25	0.05	0.13	0.43	0.60	0.97
6.1	Agent’s Personality Pr.	−0.44	−0.37	0.03	0.14	0.17	0.28	0.33	0.72	0.84	1.03	1.25	1.39	1.43
7	User Acceptance of the A.	0.05	0.36	0.53	0.58	0.73	1.01	1.31	1.55	1.61	1.71	1.80	2.03	2.07
8	Agent’s Enjoyability	−0.25	−0.16	0.17	0.33	0.46	0.66	0.89	1.05	1.37	1.42	1.56	1.77	1.96
9	User’s Engagement	0.92	1.10	1.39	1.44	1.52	1.59	1.74	1.83	2.02	2.08	2.16	2.28	2.34
10	User’s Trust	−0.27	−0.09	0.20	0.24	0.28	0.37	0.57	0.67	0.83	0.86	1.10	1.34	1.49
11	User Agent Alliance	−0.80	−0.78	−0.46	−0.37	−0.27	−0.20	−0.08	−0.01	0.38	0.41	0.48	0.82	0.93
12	Agent’s Attentiveness	0.41	0.84	0.98	1.11	1.25	1.49	1.61	1.67	1.75	1.78	1.81	1.95	2.12
13	Agent’s Coherence	0.74	1.08	1.18	1.27	1.31	1.47	1.57	1.68	1.75	1.92	2.00	2.09	2.17
14	Agent’s Intentionality	−0.47	−0.14	0.56	0.61	0.73	1.00	1.14	1.22	1.48	1.57	1.69	1.79	1.86
15	Attitude	−0.26	−0.01	0.36	0.47	0.56	1.00	1.38	1.47	1.58	1.63	2.01	2.22	2.43
16	Social Presence	−1.29	−1.14	−0.87	−0.72	−0.41	−0.19	−0.13	−0.02	0.05	0.20	0.45	0.60	0.75
17	Interaction Impact on Self.	−0.22	−0.03	0.38	0.49	0.53	0.69	0.90	1.01	1.14	1.21	1.24	1.31	1.37
18.1	Agent’s Emotional Int. Pr.	−1.77	−1.52	−1.27	−1.26	−1.11	−0.89	−0.36	−0.09	0.20	0.42	0.49	0.86	1.32
18.3	User’s Emotion Presence	−0.50	−0.36	−0.07	0.03	0.12	0.55	0.65	0.96	1.14	1.39	1.61	1.87	2.18
19	User Agent Interplay	−0.94	−0.81	−0.49	−0.22	−0.07	0.20	0.38	0.62	0.80	1.05	1.21	1.27	1.33
	Mean:	−0.51	−0.31	0.00	0.11	0.22	0.45	0.66	0.86	1.09	1.21	1.35	1.58	1.79
	SD:	0.89	0.92	0.89	0.89	0.87	0.84	0.80	0.71	0.62	0.61	0.58	0.54	0.49
	Median:	−0.45	−0.26	0.13	0.29	0.34	0.56	0.75	0.98	1.14	1.30	1.41	1.74	1.96
	Min:	−2.66	−2.60	−2.20	−2.19	−2.13	−1.74	−1.17	−0.28	0.05	0.13	0.43	0.60	0.75
	Max:	0.92	1.10	1.39	1.44	1.52	1.59	1.74	1.83	2.02	2.08	2.16	2.28	2.43

Table D.13
The percentile scores of the difference of the ASAQ construct/dimension scores (based on the representative set 2024 of the **long** version of the ASAQ, $n = 1066$).

No.	Construct/Dimension	Percentile												
		5%	10%	20%	25%	30%	40%	50%	60%	70%	75%	80%	90%	95%
1.1	Human-Like Appearance	0.12	0.21	0.41	0.58	0.73	1.09	1.56	2.21	2.83	3.01	3.17	3.67	3.97
1.2	Human-Like Behaviour	0.08	0.14	0.28	0.34	0.42	0.57	0.77	0.94	1.14	1.28	1.46	1.87	2.11
1.3	Natural Appearance	0.09	0.16	0.30	0.38	0.47	0.66	0.90	1.09	1.31	1.46	1.60	2.02	2.51
1.4	Natural Behaviour	0.05	0.12	0.26	0.32	0.38	0.52	0.72	0.84	1.13	1.42	1.57	2.30	2.76
1.5	Agent's Appearance Suit.	0.07	0.12	0.22	0.28	0.31	0.42	0.53	0.66	0.80	0.91	1.01	1.25	1.43
2	Agent's Usability	0.04	0.09	0.21	0.27	0.34	0.43	0.51	0.66	0.84	0.92	0.97	1.33	1.62
3	Performance	0.05	0.08	0.16	0.19	0.23	0.35	0.44	0.56	0.66	0.74	0.79	1.02	1.21
4	Agent's Likeability	0.09	0.16	0.30	0.37	0.44	0.61	0.74	0.93	1.17	1.32	1.40	1.77	2.08
5	Agent's Sociability	0.08	0.16	0.29	0.33	0.41	0.57	0.71	0.87	1.06	1.17	1.28	1.62	1.86
6.1	Agent's Personality Pr.	0.06	0.12	0.28	0.36	0.46	0.60	0.75	0.95	1.23	1.34	1.43	1.81	1.99
7	User Acceptance of the A.	0.05	0.12	0.25	0.31	0.39	0.50	0.62	0.78	0.92	1.02	1.15	1.47	1.65
8	Agent's Enjoyability	0.08	0.15	0.28	0.34	0.41	0.55	0.69	0.85	1.05	1.17	1.28	1.62	1.92
9	User's Engagement	0.06	0.08	0.15	0.20	0.24	0.32	0.42	0.52	0.63	0.69	0.76	1.00	1.26
10	User's Trust	0.05	0.10	0.20	0.25	0.30	0.41	0.53	0.69	0.80	0.88	0.97	1.19	1.39
11	User Agent Alliance	0.04	0.08	0.15	0.20	0.23	0.31	0.41	0.49	0.62	0.71	0.77	0.96	1.11
12	Agent's Attentiveness	0.04	0.08	0.16	0.19	0.24	0.32	0.40	0.50	0.65	0.74	0.84	1.20	1.50
13	Agent's Coherence	0.04	0.08	0.16	0.19	0.22	0.31	0.40	0.49	0.61	0.69	0.80	1.04	1.28
14	Agent's Intentionality	0.07	0.13	0.26	0.32	0.37	0.55	0.68	0.84	1.00	1.10	1.23	1.48	1.74
15	Attitude	0.08	0.16	0.32	0.41	0.54	0.74	0.94	1.16	1.42	1.55	1.69	2.13	2.35
16	Social Presence	0.06	0.12	0.21	0.26	0.31	0.43	0.53	0.70	0.86	0.96	1.03	1.31	1.52
17	Interaction Impact on Self.	0.05	0.10	0.19	0.27	0.32	0.44	0.57	0.67	0.84	0.91	0.99	1.31	1.54
18.1	Agent's Emotional Int. Pr.	0.09	0.19	0.35	0.42	0.50	0.71	0.91	1.14	1.44	1.59	1.80	2.24	2.56
18.3	User's Emotion Presence	0.07	0.17	0.28	0.36	0.46	0.58	0.73	0.89	1.10	1.21	1.36	1.62	1.91
19	User Agent Interplay	0.07	0.11	0.20	0.26	0.31	0.41	0.51	0.62	0.76	0.86	0.97	1.19	1.44
	Mean:	0.07	0.13	0.24	0.31	0.38	0.52	0.67	0.84	1.04	1.15	1.26	1.60	1.86
	SD:	0.02	0.04	0.07	0.09	0.12	0.18	0.25	0.35	0.46	0.48	0.51	0.59	0.64
	Median:	0.06	0.12	0.26	0.32	0.38	0.51	0.65	0.81	0.96	1.06	1.19	1.48	1.69
	Min:	0.04	0.08	0.15	0.19	0.22	0.31	0.40	0.49	0.61	0.69	0.76	0.96	1.11
	Max:	0.12	0.21	0.41	0.58	0.73	1.09	1.56	2.21	2.83	3.01	3.17	3.67	3.97

Table D.14
The percentile scores of the difference of the ASAQ construct/dimension (based on the representative set 2024 of the **short** version of the ASAQ, $n = 1066$).

No.	Construct/Dimension	Percentile												
		5%	10%	20%	25%	30%	40%	50%	60%	70%	75%	80%	90%	95%
1.1	Human-Like Appearance	0.08	0.22	0.48	0.55	0.78	1.13	1.75	2.51	3.08	3.31	3.53	4.00	4.30
1.2	Human-Like Behaviour	0.10	0.18	0.34	0.43	0.52	0.71	0.92	1.18	1.40	1.62	1.80	2.30	2.63
1.3	Natural Appearance	0.08	0.15	0.33	0.42	0.53	0.71	0.91	1.20	1.46	1.60	1.72	2.35	3.01
1.4	Natural Behaviour	0.08	0.17	0.36	0.42	0.54	0.72	0.96	1.20	1.40	1.55	1.78	2.30	2.74
1.5	Agent's Appearance Suit.	0.06	0.13	0.23	0.29	0.33	0.46	0.62	0.76	0.96	1.05	1.13	1.40	1.56
2	Agent's Usability	0.04	0.14	0.23	0.29	0.35	0.46	0.58	0.73	0.92	0.98	1.12	1.47	1.77
3	Performance	0.05	0.10	0.22	0.26	0.33	0.44	0.56	0.68	0.82	0.89	0.99	1.25	1.40
4	Agent's Likeability	0.10	0.18	0.38	0.46	0.56	0.78	0.97	1.16	1.39	1.52	1.69	2.12	2.48
5	Agent's Sociability	0.08	0.15	0.32	0.38	0.45	0.61	0.75	0.90	1.13	1.24	1.41	1.69	2.06
6.1	Agent's Personality Pr.	0.05	0.11	0.21	0.31	0.40	0.53	0.66	0.79	1.04	1.11	1.21	1.41	1.75
7	User Acceptance of the A.	0.05	0.14	0.26	0.33	0.41	0.54	0.73	0.89	1.08	1.19	1.31	1.63	1.81
8	Agent's Enjoyability	0.10	0.16	0.33	0.40	0.46	0.63	0.79	0.97	1.15	1.28	1.44	1.78	2.00
9	User's Engagement	0.05	0.09	0.17	0.22	0.25	0.35	0.46	0.58	0.67	0.75	0.81	0.99	1.26
10	User's Trust	0.05	0.10	0.21	0.26	0.32	0.44	0.54	0.68	0.84	0.93	1.03	1.26	1.43
11	User Agent Alliance	0.05	0.11	0.21	0.28	0.33	0.46	0.57	0.69	0.83	0.94	1.07	1.28	1.54
12	Agent's Attentiveness	0.05	0.09	0.17	0.21	0.26	0.37	0.48	0.62	0.75	0.83	0.94	1.31	1.50
13	Agent's Coherence	0.04	0.09	0.17	0.22	0.27	0.38	0.44	0.55	0.69	0.78	0.87	1.12	1.48
14	Agent's Intentionality	0.06	0.14	0.28	0.34	0.43	0.56	0.68	0.90	1.13	1.21	1.36	1.78	2.16
15	Attitude	0.08	0.16	0.37	0.44	0.53	0.70	0.92	1.08	1.37	1.51	1.62	2.04	2.38
16	Social Presence	0.05	0.11	0.23	0.30	0.38	0.56	0.70	0.86	1.01	1.14	1.25	1.61	1.87
17	Interaction Impact on Self.	0.05	0.11	0.19	0.24	0.30	0.39	0.53	0.64	0.78	0.88	0.96	1.27	1.54
18.1	Agent's Emotional Int. Pr.	0.08	0.23	0.38	0.52	0.60	0.83	1.05	1.32	1.54	1.72	1.88	2.37	2.73
18.3	User's Emotion Presence	0.10	0.18	0.38	0.45	0.55	0.69	0.92	1.10	1.31	1.47	1.64	2.04	2.34
19	User Agent Interplay	0.06	0.16	0.28	0.40	0.46	0.63	0.81	1.00	1.21	1.36	1.46	1.89	2.11
	Mean:	0.07	0.14	0.28	0.35	0.43	0.59	0.76	0.96	1.17	1.29	1.42	1.78	2.08
	SD:	0.02	0.04	0.08	0.10	0.13	0.18	0.28	0.40	0.48	0.52	0.55	0.63	0.68
	Median:	0.06	0.14	0.27	0.34	0.42	0.56	0.72	0.90	1.10	1.20	1.33	1.66	1.94
	Min:	0.04	0.09	0.17	0.21	0.25	0.35	0.44	0.55	0.67	0.75	0.81	0.99	1.26
	Max:	0.10	0.23	0.48	0.55	0.78	1.13	1.75	2.51	3.08	3.31	3.53	4.00	4.30

Table D.15
The percentile scores of the Cohen's effect sizes (*d*) (based on the difference of the ASAQ construct/dimension scores of the **long** version of the ASAQ representative set 2024, *n* = 1066).

No.	Construct/Dimension	Percentile												
		5%	10%	20%	25%	30%	40%	50%	60%	70%	75%	80%	90%	95%
1.1	Human-Like Appearance	0.09	0.16	0.33	0.43	0.55	0.76	1.06	1.48	2.04	2.17	2.36	2.99	3.32
1.2	Human-Like Behaviour	0.06	0.10	0.20	0.24	0.29	0.41	0.54	0.69	0.84	0.98	1.13	1.46	1.68
1.3	Natural Appearance	0.08	0.14	0.24	0.31	0.39	0.54	0.72	0.92	1.11	1.24	1.36	1.77	2.28
1.4	Natural Behaviour	0.04	0.09	0.20	0.25	0.29	0.40	0.54	0.66	0.92	1.10	1.31	2.03	2.44
1.5	Agent's Appearance Suit.	0.06	0.10	0.19	0.24	0.28	0.37	0.47	0.57	0.70	0.76	0.85	1.02	1.21
2	Agent's Usability	0.03	0.08	0.18	0.25	0.29	0.37	0.46	0.60	0.74	0.81	0.91	1.19	1.52
3	Performance	0.05	0.07	0.15	0.19	0.22	0.32	0.42	0.52	0.62	0.68	0.78	1.01	1.20
4	Agent's Likeability	0.07	0.13	0.24	0.31	0.37	0.49	0.61	0.77	0.94	1.06	1.13	1.54	1.73
5	Agent's Sociability	0.06	0.12	0.22	0.26	0.30	0.44	0.56	0.67	0.81	0.89	1.02	1.27	1.45
6.1	Agent's Personality Pr.	0.05	0.10	0.22	0.28	0.34	0.45	0.57	0.72	0.92	1.02	1.15	1.38	1.62
7	User Acceptance of the A.	0.05	0.11	0.21	0.26	0.33	0.43	0.55	0.68	0.79	0.88	1.02	1.25	1.50
8	Agent's Enjoyability	0.07	0.13	0.24	0.31	0.36	0.48	0.62	0.74	0.95	1.05	1.13	1.44	1.70
9	User's Engagement	0.06	0.09	0.17	0.20	0.24	0.33	0.43	0.52	0.64	0.68	0.78	0.94	1.07
10	User's Trust	0.05	0.09	0.17	0.22	0.27	0.36	0.48	0.58	0.70	0.78	0.87	1.07	1.21
11	User Agent Alliance	0.04	0.08	0.14	0.18	0.23	0.30	0.37	0.46	0.57	0.63	0.69	0.85	0.98
12	Agent's Attentiveness	0.05	0.07	0.15	0.19	0.21	0.31	0.38	0.45	0.57	0.65	0.72	1.00	1.20
13	Agent's Coherence	0.04	0.08	0.15	0.19	0.22	0.30	0.39	0.48	0.60	0.64	0.74	1.03	1.23
14	Agent's Intentionality	0.06	0.11	0.22	0.26	0.31	0.45	0.58	0.70	0.85	0.91	1.02	1.26	1.49
15	Attitude	0.06	0.14	0.27	0.36	0.45	0.61	0.81	0.96	1.14	1.30	1.43	1.85	2.09
16	Social Presence	0.04	0.09	0.15	0.21	0.24	0.32	0.42	0.52	0.64	0.71	0.76	0.98	1.16
17	Interaction Impact on Self.	0.05	0.09	0.20	0.25	0.29	0.40	0.52	0.65	0.77	0.84	0.92	1.16	1.32
18.1	Agent's Emotional Int. Pr.	0.06	0.13	0.25	0.30	0.35	0.49	0.63	0.79	0.97	1.12	1.30	1.71	1.98
18.3	User's Emotion Presence	0.06	0.14	0.26	0.31	0.40	0.53	0.64	0.77	0.95	1.10	1.19	1.56	1.77
19	User Agent Interplay	0.06	0.11	0.19	0.24	0.29	0.38	0.47	0.58	0.71	0.80	0.87	1.06	1.30
	Mean:	0.06	0.11	0.21	0.26	0.31	0.43	0.55	0.69	0.85	0.95	1.06	1.37	1.60
	SD:	0.01	0.02	0.05	0.06	0.08	0.11	0.15	0.22	0.30	0.32	0.35	0.47	0.53
	Median:	0.06	0.10	0.20	0.25	0.29	0.41	0.54	0.66	0.80	0.88	1.02	1.25	1.50
	Min:	0.03	0.07	0.14	0.18	0.21	0.30	0.37	0.45	0.57	0.63	0.69	0.85	0.98
	Max:	0.09	0.16	0.33	0.43	0.55	0.76	1.06	1.48	2.04	2.17	2.36	2.99	3.32

Table D.16
The percentile scores of the Cohen's effect sizes (*d*) (based on the difference of the ASAQ construct/dimension scores of the **short** version of the ASAQ representative set 2024, *n* = 1066).

No.	Construct/Dimension	Percentile												
		5%	10%	20%	25%	30%	40%	50%	60%	70%	75%	80%	90%	95%
1.1	Human-Like Appearance	0.05	0.14	0.32	0.40	0.50	0.69	1.06	1.46	1.89	2.06	2.28	2.88	3.12
1.2	Human-Like Behaviour	0.06	0.11	0.20	0.25	0.29	0.42	0.53	0.69	0.85	0.94	1.08	1.39	1.64
1.3	Natural Appearance	0.05	0.07	0.17	0.23	0.29	0.39	0.52	0.66	0.83	0.89	0.98	1.41	1.83
1.4	Natural Behaviour	0.04	0.09	0.20	0.23	0.30	0.40	0.52	0.67	0.80	0.92	1.07	1.50	1.80
1.5	Agent's Appearance Suit.	0.05	0.09	0.16	0.20	0.24	0.32	0.40	0.52	0.63	0.70	0.76	0.94	1.02
2	Agent's Usability	0.04	0.11	0.18	0.22	0.27	0.35	0.46	0.56	0.68	0.75	0.82	1.05	1.28
3	Performance	0.03	0.08	0.17	0.20	0.26	0.34	0.42	0.53	0.63	0.69	0.75	0.96	1.14
4	Agent's Likeability	0.07	0.11	0.22	0.29	0.35	0.48	0.59	0.72	0.85	0.93	1.05	1.38	1.60
5	Agent's Sociability	0.05	0.08	0.18	0.22	0.26	0.36	0.44	0.53	0.67	0.74	0.83	1.03	1.25
6.1	Agent's Personality Pr.	0.03	0.07	0.12	0.19	0.24	0.32	0.38	0.45	0.62	0.68	0.74	0.92	1.02
7	User Acceptance of the A.	0.04	0.10	0.18	0.25	0.29	0.39	0.51	0.62	0.77	0.86	0.96	1.16	1.38
8	Agent's Enjoyability	0.06	0.09	0.19	0.24	0.27	0.36	0.48	0.57	0.70	0.76	0.86	1.06	1.25
9	User's Engagement	0.05	0.08	0.15	0.18	0.21	0.28	0.36	0.45	0.56	0.61	0.66	0.86	1.01
10	User's Trust	0.03	0.07	0.14	0.17	0.20	0.28	0.34	0.44	0.55	0.59	0.66	0.86	0.99
11	User Agent Alliance	0.03	0.06	0.13	0.17	0.21	0.29	0.35	0.44	0.52	0.58	0.67	0.82	0.96
12	Agent's Attentiveness	0.04	0.07	0.13	0.16	0.19	0.27	0.36	0.46	0.54	0.60	0.66	0.88	0.97
13	Agent's Coherence	0.03	0.07	0.13	0.16	0.20	0.26	0.32	0.40	0.50	0.55	0.60	0.77	0.98
14	Agent's Intentionality	0.04	0.09	0.17	0.21	0.24	0.32	0.40	0.54	0.65	0.72	0.80	1.06	1.28
15	Attitude	0.06	0.12	0.27	0.32	0.39	0.49	0.65	0.79	0.95	1.05	1.19	1.47	1.71
16	Social Presence	0.03	0.06	0.13	0.17	0.23	0.32	0.40	0.50	0.60	0.65	0.73	0.94	1.10
17	Interaction Impact on Self.	0.03	0.07	0.13	0.16	0.19	0.26	0.35	0.44	0.53	0.57	0.62	0.82	0.96
18.1	Agent's Emotional Int. Pr.	0.05	0.12	0.22	0.26	0.32	0.45	0.57	0.69	0.86	0.94	1.04	1.41	1.68
18.3	User's Emotion Presence	0.06	0.11	0.23	0.27	0.33	0.44	0.57	0.68	0.85	0.93	1.07	1.40	1.64
19	User Agent Interplay	0.04	0.10	0.18	0.25	0.30	0.41	0.51	0.65	0.79	0.85	0.97	1.20	1.36
	Mean:	0.04	0.09	0.18	0.22	0.27	0.37	0.48	0.60	0.74	0.81	0.91	1.17	1.37
	SD:	0.01	0.02	0.05	0.06	0.07	0.10	0.15	0.21	0.28	0.30	0.34	0.43	0.48
	Median:	0.04	0.09	0.17	0.22	0.26	0.36	0.45	0.55	0.68	0.74	0.82	1.06	1.27
	Min:	0.03	0.06	0.12	0.16	0.19	0.26	0.32	0.40	0.50	0.55	0.60	0.77	0.96
	Max:	0.07	0.14	0.32	0.40	0.50	0.69	1.06	1.46	1.89	2.06	2.28	2.88	3.12

Table D.17
Sample sizes for each ASAQ construct/dimension of the long and short ASAQ versions based on Cohen's effect sizes (d) from Table D.15 and D.16 with .80 power and .05 alpha level (two-tailed t-test).

No.	Construct/Dimension	Long Version Effect size			Short Version Effect size		
		Small (25%)	Medium (50%)	Large (75%)	Small (25%)	Medium (50%)	Large (75%)
1.1	Human-Like Appearance	86	15	5	99	15	5
1.2	Human-Like Behaviour	273	55	17	252	57	19
1.3	Natural Appearance	164	31	11	298	59	21
1.4	Natural Behaviour	252	55	14	298	59	20
1.5	Agent's Appearance Suit.	273	72	28	393	99	33
2	Agent's Usability	252	75	25	325	75	29
3	Performance	436	90	35	393	90	34
4	Agent's Likeability	164	43	15	188	46	19
5	Agent's Sociability	233	51	21	325	82	30
6.1	Agent's Personality Pr.	201	49	16	436	110	35
7	User Acceptance of the A.	233	53	21	252	61	22
8	Agent's Enjoyability	164	42	15	273	69	28
9	User's Engagement	393	86	35	485	122	43
10	User's Trust	325	69	27	544	137	46
11	User Agent Alliance	485	116	41	544	129	48
12	Agent's Attentiveness	436	110	38	614	122	45
13	Agent's Coherence	436	104	39	614	154	53
14	Agent's Intentionality	233	48	20	357	99	31
15	Attitude	122	25	10	154	38	15
16	Social Presence	357	90	32	544	99	38
17	Interaction Impact on Self.	252	59	23	614	129	49
18.1	Agent's Emotional Int. Pr.	175	41	14	233	49	19
18.3	User's Emotion Presence	164	39	14	216	49	19
19	User Agent Interplay	273	72	26	252	61	23
	Max:	485	116	41	614	154	53

Table D.18
Sample sizes for each ASAQ construct/dimension for the 5th percentile error margin.

No.	Construct/Dimension	Long Version				Short Version			
		90% CI	95% CI	97.5% CI	99% CI	90% CI	95% CI	97.5% CI	99% CI
1.1	Human-Like Appearance	92	130	171	225	102	145	190	251
1.2	Human-Like Behaviour	235	334	436	576	261	370	484	640
1.3	Natural Appearance	186	264	346	456	293	416	544	719
1.4	Natural Behaviour	168	239	312	412	303	430	563	743
1.5	Agent's Appearance Suit.	454	645	843	1113	704	999	1307	1726
2	Agent's Usability	506	718	939	1240	476	676	884	1167
3	Performance	511	725	948	1252	591	839	1097	1449
4	Agent's Likeability	209	296	387	511	284	403	528	697
5	Agent's Sociability	305	433	567	748	452	642	840	1110
6.1	Agent's Personality Pr.	337	478	625	826	696	988	1292	1706
7	User Acceptance of the A.	297	422	551	728	432	614	803	1060
8	Agent's Enjoyability	211	299	391	517	518	736	963	1271
9	User's Engagement	495	703	919	1213	677	961	1257	1660
10	User's Trust	466	661	865	1142	701	996	1302	1720
11	User Agent Alliance	689	979	1280	1691	767	1089	1424	1881
12	Agent's Attentiveness	482	684	894	1181	575	817	1068	1411
13	Agent's Coherence	543	771	1009	1332	871	1237	1618	2137
14	Agent's Intentionality	345	490	641	846	468	665	869	1148
15	Attitude	173	245	320	423	243	344	450	595
16	Social Presence	551	783	1023	1352	634	901	1178	1556
17	Interaction Impact on Self.	333	472	618	816	804	1142	1494	1972
18.1	Agent's Emotional Int. Pr.	212	300	393	519	309	439	574	758
18.3	User's Emotion Presence	250	355	464	613	311	441	577	762
19	User Agent Interplay	414	588	769	1015	416	591	772	1020
	Max:	689	979	1280	1691	871	1237	1618	2137

Table D.19
Sample sizes for each ASAQ construct/dimension for the 10th percentile error margin.

No.	Construct/Dimension	Long Version				Short Version			
		90% CI	95% CI	97.5% CI	99% CI	90% CI	95% CI	97.5% CI	99% CI
1.1	Human-Like Appearance	21	30	39	51	23	33	43	57
1.2	Human-Like Behaviour	90	127	167	220	81	115	150	199
1.3	Natural Appearance	65	93	121	160	150	213	279	369
1.4	Natural Behaviour	75	107	139	184	84	119	155	205
1.5	Agent’s Appearance Suit.	142	202	264	349	187	265	347	458
2	Agent’s Usability	153	217	283	374	138	196	257	339
3	Performance	148	210	275	363	156	222	290	384
4	Agent’s Likeability	68	97	127	167	82	116	152	200
5	Agent’s Sociability	94	133	174	230	168	239	313	413
6.1	Agent’s Personality Pr.	77	109	143	189	154	218	286	377
7	User Acceptance of the A.	95	135	177	234	124	176	230	303
8	Agent’s Enjoyability	77	110	144	190	133	189	247	326
9	User’s Engagement	188	267	350	462	196	279	364	481
10	User’s Trust	132	188	246	325	209	297	389	513
11	User Agent Alliance	161	228	298	394	177	252	329	435
12	Agent’s Attentiveness	192	273	357	471	267	379	496	655
13	Agent’s Coherence	201	286	374	493	350	497	650	859
14	Agent’s Intentionality	102	144	189	249	135	191	250	330
15	Attitude	47	67	88	116	70	99	129	171
16	Social Presence	152	216	282	373	172	244	319	422
17	Interaction Impact on Self.	137	195	255	337	220	312	408	539
18.1	Agent’s Emotional Int. Pr.	54	77	101	133	102	145	189	250
18.3	User’s Emotion Presence	61	87	114	150	88	125	163	216
19	User Agent Interplay	131	186	243	321	98	139	182	241
	Max:	201	286	374	493	350	497	650	859

Table D.20
Sample sizes for each ASAQ construct/dimension for the 20th percentile error margin.

No.	Construct/Dimension	Long Version				Short Version			
		90% CI	95% CI	97.5% CI	99% CI	90% CI	95% CI	97.5% CI	99% CI
1.1	Human-Like Appearance	5	6	8	11	5	7	10	13
1.2	Human-Like Behaviour	43	61	80	105	39	55	72	95
1.3	Natural Appearance	15	22	28	37	30	43	56	74
1.4	Natural Behaviour	34	48	62	83	40	57	74	98
1.5	Agent’s Appearance Suit.	33	47	61	81	43	60	79	104
2	Agent’s Usability	36	52	68	89	48	68	90	118
3	Performance	69	98	128	169	40	57	75	98
4	Agent’s Likeability	24	34	45	59	25	36	47	62
5	Agent’s Sociability	35	50	66	87	34	48	63	84
6.1	Agent’s Personality Pr.	19	27	35	46	46	65	84	112
7	User Acceptance of the A.	34	48	63	83	30	43	56	74
8	Agent’s Enjoyability	22	31	41	54	36	52	68	89
9	User’s Engagement	54	77	101	133	64	91	118	156
10	User’s Trust	31	44	58	77	74	105	137	181
11	User Agent Alliance	60	85	111	146	71	101	133	175
12	Agent’s Attentiveness	70	100	130	172	68	97	127	167
13	Agent’s Coherence	52	73	96	126	74	105	137	181
14	Agent’s Intentionality	23	33	43	56	55	79	103	136
15	Attitude	14	20	26	34	18	25	33	44
16	Social Presence	44	62	82	108	42	60	78	103
17	Interaction Impact on Self.	37	53	69	91	76	108	141	186
18.1	Agent’s Emotional Int. Pr.	19	27	35	46	26	37	49	64
18.3	User’s Emotion Presence	21	29	38	50	22	31	40	53
19	User Agent Interplay	38	55	71	94	21	29	39	51
	Max:	70	100	130	172	76	108	141	186

Table D.21
Sample sizes for each ASAQ construct/dimension for the 25th percentile error margin.

No.	Construct/Dimension	Long Version				Short Version			
		90% CI	95% CI	97.5% CI	99% CI	90% CI	95% CI	97.5% CI	99% CI
1.1	Human-Like Appearance	2	3	4	6	2	4	5	6
1.2	Human-Like Behaviour	31	44	57	75	21	30	40	52
1.3	Natural Appearance	10	14	19	25	17	24	31	41
1.4	Natural Behaviour	44	63	82	108	34	48	63	83
1.5	Agent's Appearance Suit.	23	33	43	57	24	34	44	59
2	Agent's Usability	19	27	35	47	25	36	47	62
3	Performance	37	52	68	90	21	30	39	52
4	Agent's Likeability	19	27	35	46	20	29	37	49
5	Agent's Sociability	19	27	35	46	31	45	58	77
6.1	Agent's Personality Pr.	11	16	21	27	38	54	71	94
7	User Acceptance of the A.	20	29	37	49	17	24	31	42
8	Agent's Enjoyability	17	24	31	41	27	38	49	65
9	User's Engagement	33	47	61	81	41	59	77	101
10	User's Trust	17	23	31	41	68	97	126	167
11	User Agent Alliance	40	57	75	99	46	66	86	114
12	Agent's Attentiveness	41	59	77	101	46	65	85	113
13	Agent's Coherence	45	64	84	111	52	74	97	128
14	Agent's Intentionality	12	17	22	29	34	48	63	83
15	Attitude	8	12	15	20	16	23	30	39
16	Social Presence	46	66	86	114	38	54	71	93
17	Interaction Impact on Self.	25	35	46	61	48	68	89	117
18.1	Agent's Emotional Int. Pr.	12	17	22	29	13	18	24	32
18.3	User's Emotion Presence	13	19	24	32	15	21	27	36
19	User Agent Interplay	33	47	61	81	16	23	30	40
	Max:	46	66	86	114	68	97	126	167

Data availability

Data and analysis code of the results presented in this paper are available online (Fitrianie et al., 2025).

References

Albers, N., Bönsch, A., Ehret, J., Khodakov, B.A., Brinkman, W.-P., 2024. German and dutch translations of the artificial-social-agent questionnaire instrument for evaluating human-agent interactions. In: ACM International Conference on Intelligent Virtual Agents (IVA '24). ACM, p. 4. http://dx.doi.org/10.1145/3652988.3673928.

Arora, S., Liang, Y., Ma, T., 2017. A simple but tough-to-beat baseline for sentence embeddings. In: Proc. of International Conference on Learning Representations. pp. 1–16, URL https://openreview.net/forum?id=SyK00v5xx.

Baker, M., 2016. 1,500 scientists lift the lid on reproducibility. Nature 533, 452–454. http://dx.doi.org/10.1038/533452a.

Bartneck, C., Kulić, D., Croft, E., Zoghbi, S., 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int. J. Soc. Robot. 1, 71–81.

Bevacqua, E., Prepin, K., Niewiadomski, R., de Sevin, E., Pelachaud, C., 2010. Greta: Towards an interactive conversational virtual companion. Artif. Companions Society: Perspect. Present. Futur. 1–17.

Bickmore, T.W., Vardoulakis, L.P., Jack, B.W., Paasche-Orlow, M.K., 2013. Automated promotion of technology acceptance by clinicians using relational agents. In: Proc. of Intelligent Virtual Agents. In: Lecture Notes in Computer Science, 8108, Springer, pp. 68–78. http://dx.doi.org/10.1007/978-3-642-40415-3_6.

Blunch, N.J., 2013. Introduction to Structural Equation Modeling using IBM SPSS Statistics and AMOS, 2nd Eds. SAGE, City Road, London, http://dx.doi.org/10.4135/9781526402257.

Borsci, S., Malizia, A., Schmettow, M., Van Der Velde, F., Tariverdiyeva, G., Balaji, D., Chamberlain, A., 2022. The chatbot usability scale: the design and pilot of a usability scale for interaction with AI-based conversational agents. Pers. Ubiquitous Comput. 26, 95–119.

Bradley, M.M., Lang, P.J., 1994. Measuring emotion: The self-assessment manikin and the semantic differential. J. Behav. Ther. Exp. Psychiatry 25 (1), 49–59. http://dx.doi.org/10.1016/0005-7916(94)90063-9.

Brinkman, W.-P., Haakma, R., Bouwhuis, D.G., 2009. The theoretical foundation and validity of a component-based usability questionnaire. Behav. Inf. Technol. 28 (2), 121–137. http://dx.doi.org/10.1080/01449290701306510.

Cafaro, A., Bruijnes, M., van Waterschoot, J., Pelachaud, C., Theune, M., Heylen, D., 2017. Selecting and expressing communicative functions in a SAIBA-compliant agent framework. In: Intelligent Virtual Agents: 17th International Conference, IVA 2017, Stockholm, Sweden, August 27-30, 2017, Proceedings 17. Springer, pp. 73–82.

Cafaro, A., Vilhjálmsson, H.H., Bickmore, T., Heylen, D., Pelachaud, C., 2014. Representing communicative functions in saiba with a unified function markup language. In: Intelligent Virtual Agents: 14th International Conference, IVA 2014, Boston, MA, USA, August 27-29, 2014. Proceedings 14. Springer, pp. 81–94.

Camerer, C.F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., Wu, H., 2016. Evaluating replicability of laboratory experiments in economics. Science 351 (6280), 1433–1436. http://dx.doi.org/10.1126/science.aaf0918.

Cohen, J., 2013. Statistical Power Analysis for the Behavioral Sciences. Taylor and Francis.

de Vet, H.C.W., Adér, H.J., Terwee, C.B., Pouwer, F., 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. Qual. Life Res. 14, 1203–1218. http://dx.doi.org/10.1007/s11136-004-5742-3.

de Vet, H.C.W., Terwee, C.B., Knol, D.L., Mokkink, L.B., 2011. Measurement in Medicine - A Practical Guide. U.S.A., Cambridge University Press, New York.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., Morency, L.-P., 2014. SimSensei kiosk: A virtual human interviewer for healthcare decision support. In: Proc. of AAMAS '14. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, pp. 1061–1068.

DeVellis, R.F., Thorpe, C.T., 2021. Scale Development: Theory and Applications, 5th eds Sage, Thousand Oaks, CA.

Diehl, C., Schiffhauer, B., Eyssel, F., Achenbach, J., Klett, S., Botsch, M., Kopp, S., 2017. Get one or create one: the impact of graded involvement in a selection procedure for a virtual agent on satisfaction and suitability ratings. In: Proc. of Intelligent Virtual Agents. In: Lecture Notes in Computer Science, 10498, Springer, pp. 109–118. http://dx.doi.org/10.1007/978-3-319-67401-8_13.

Ekman, P., 1999. Basic emotions. In: Dalgleish, T., Powers, M.J. (Eds.), Handbook of Cognition and Emotion. Wiley, pp. 4–5. http://dx.doi.org/10.1002/0470013494.

Errington, T.M., Mathur, M., Soderberg, C.K., Denis, A., Perfito, N., Iorns, E., Nosek, B.A., 2021. Investigating the replicability of preclinical cancer biology. In: Pasqualini, R., Franco, E. (Eds.), ELife 10, e71601. http://dx.doi.org/10.7554/eLife.71601.

European Union, 2024. EU artificial intelligence act. https://artificialintelligenceact.eu/ai-act-explorer/. (Accessed January 2025).

Feldman-Barrett, L., Russell, J.A., 1998. Independence and bipolarity in the structure of current affect. J. Pers. Soc. Psychol. 74 (4), 967—-984. http://dx.doi.org/10.1037/0022-3514.74.4.967.

Fitrianie, S., Bruijnes, M., Abdulrahman, A., Brinkman, W.-P., 2025. Data and Analysis Underlying the Research into The Artificial Social Agent Questionnaire (ASAQ) - Development and Evaluation of a Validated Instrument for Capturing Human Interaction Experiences with Artificial Social Agents. http://dx.doi.org/10.4121/4fe035a8-45ff-4ffc-a269-380d09361029, Fitrianie, S., Bruijnes, M., Abdulrahman, A., Li, F., Brinkman, W.-P., 2023. Study 9: Concurrent validation and a normative dataset development. OSF Registries https://doi.org/10.17605/OSF.IO/6GZ29. Open-Ended Registration.

Fitrianie, S., Bruijnes, M., Brinkman, W.-P., 2020b. Study 4: Define questionnaire items. OSF Registries https://osf.io/qxeu5/registrations. Open-Ended Registration.

Fitrianie, S., Bruijnes, M., Brinkman, W.-P., 2021b. Study 6: Analysing reliability of questionnaire items. OSF Registries https://osf.io/hyxwb/registrations Open-Ended Registration.

Fitrianie, S., Bruijnes, M., Brinkman, W.-P., 2021c. Study 7: Confirmatory factor analysis. OSF Registries https://osf.io/nfgqx/registrations Open-Ended Registration.

Fitrianie, S., Bruijnes, M., Brinkman, W.-P., 2022c. Study 8: Cross validation final questionnaire set. OSF Registries https://osf.io/579hb/registrations Open-Ended Registration.

Fitrianie, S., Bruijnes, M., Li, F., Abdulrahman, A., Brinkman, W.-P., 2022a. The artificial-social-agent questionnaire: Establishing the long and short questionnaire versions. In: Proc of IVA'22. Association for Computing Machinery, New York, NY, USA, pp. 1–8. http://dx.doi.org/10.1145/3514197.3549612.

Fitrianie, S., Bruijnes, M., Li, F., Abdulrahman, A., Brinkman, W.-P., 2022b. Data and analysis underlying the research into the Artificial-Social-Agent Questionnaire: Establishing the long and short questionnaire versions. 4TU Data Repos. http://dx.doi.org/10.4121/19758436.

Fitrianie, S., Bruijnes, M., Li, F., Brinkman, W.-P., 2021a. Questionnaire items for evaluating artificial social agents - expert generated, content validated and reliability analysed. In: Proc. of IVA'21. ACM, NY, USA, pp. 84–86. http://dx.doi.org/10.1145/3472306.3478341.

Fitrianie, S., Bruijnes, M., Richards, D., Abdulrahman, A., Brinkman, W.-P., 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, pp. 159–161. http://dx.doi.org/10.1145/3308532.3329421.

Fitrianie, S., Bruijnes, M., Richards, D., Bönsch, A., Brinkman, W.-P., 2020a. The 19 unifying questionnaire constructs of artificial social agents: An IVA community analysis. In: Proc. of IVA'20. ACM, NY, USA, pp. 1–8. http://dx.doi.org/10.1145/3383652.3423873.

Goldberg, L.R., 1990. An alternative "description of personality": the big-five factor structure. J. Pers. Soc. Psychol. 59 (6), 1216—-1229. http://dx.doi.org/10.1037//0022-3514.59.6.1216.

Grigore, E.C., Pereira, A., Zhou, I., Wang, D., Scassellati, B., 2016. Talk to me: Verbal communication improves perceptions of friendship and social presence in human-robot interaction. In: Proc. of Intelligent Virtual Agents. In: Lecture Notes in Computer Science, 10011, pp. 51–63. http://dx.doi.org/10.1007/978-3-319-47665-0_5.

Gunes, H., Broz, F., Crawford, C.S., der Pütten, A.R.-v., Strait, M., Riek, L., 2022. Reproducibility in human-robot interaction: Furthering the science of HRI. Curr. Robot. Rep. 1–12. http://dx.doi.org/10.1007/s43154-022-00094-5.

Hair, Jr., J.F., Hult, G.T.M., Ringle, C.M., Sarstedt, M., Danks, N.P., Ray, S., Hair, J.F., Hult, G.T.M., Ringle, C.M., Sarstedt, M., et al., 2021. Evaluation of reflective measurement models. Partial. Least Squares Struct. Equ. Model. (PLS- SEM) using R: A Work. 75–90.

Hartholt, A., Traum, D., Marsella, S.C., Shapiro, A., Stratou, G., Leuski, A., Morency, L.-P., Gratch, J., 2013. All together now: Introducing the virtual human toolkit. In: Intelligent Virtual Agents: 13th International Conference, IVA 2013, Edinburgh, UK, August 29-31, 2013. Proceedings 13. Springer, pp. 368–381.

Heylen, D., Kopp, S., Marsella, S.C., Pelachaud, C., Vilhjálmsson, H., 2008. The next step towards a function markup language. In: Intelligent Virtual Agents: 8th International Conference, IVA 2008, Tokyo, Japan, September 1-3, 2008. Proceedings 8. Springer, pp. 270–280.

Hinkle, D.E., Wiersma, W., Jurs, S.G., 2003. Applied Statistics for the Behavioral Sciences, 5th Ed. Houghton Mifflin., Boston.

Holroyd, A., Rich, C., Sidner, C.L., Ponsler, B., 2011. Generating connection events for human-robot collaboration. In: 2011 RO-MAN. IEEE, pp. 241–246.

International Organisation for Standardisation, 2019. Ergonomics of human–system interaction – part 210: human-centred design for interactive systems. https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-2:v1:en.

Ioannidis, J.P.A., 2005a. Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. JAMA 294 (2), 218–228. http://dx.doi.org/10.1001/jama.294.2.218.

Ioannidis, J.P., 2005b. Why most published research findings are false. PLoS Med. 2 (8), e124. http://dx.doi.org/10.1371/journal.pmed.0020124.

Jaques, N., McDuff, D., Kim, Y.L., Picard, R., 2016. Understanding and predicting bonding in conversations using thin slices of facial expressions and body language. In: Proc. of Intelligent Virtual Agents. Springer International Publishing, pp. 64–74. http://dx.doi.org/10.1007/978-3-319-47665-0_6.

Kang, N., Ding, D., Riemsdijk, M.B.V., Morina, N., Neerincx, M.A., Brinkman, W.-P., 2021. Self-identification with a virtual experience and its moderating effect on self-efficacy and presence. Int. J. Human– Comput. Interact. 37 (2), 181–196. http://dx.doi.org/10.1080/10447318.2020.1812909.

Kipp, M., Heloir, A., Schröder, M., Gebhard, P., 2010. Realizing multimodal behavior: Closing the gap between behavior planning and embodied agent presentation. In: Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10. Springer, pp. 57–63.

Kolkmeier, J., Bruijnes, M., Reidsma, D., 2017. A demonstration of the ASAP realizer-Unity3D bridge for virtual and mixed reality applications. In: Intelligent Virtual Agents: 17th International Conference, IVA 2017, Stockholm, Sweden, August 27-30, 2017, Proceedings 17. Springer, pp. 223–226.

Kolkmeier, J., Vroon, J., Heylen, D., 2016. Interacting with virtual agents in shared space: Single and joint effects of gaze and proxemics. In: Proc. of Intelligent Virtual Agents. In: Lecture Notes in Computer Science, 10011, pp. 1–14. http://dx.doi.org/10.1007/978-3-319-47665-0_1.

Kopp, S., Krenn, B., Marsella, S., Marshall, A.N., Pelachaud, C., Pirker, H., Thórisson, K.R., Vilhjálmsson, H., 2006. Towards a common framework for multimodal generation: The behavior markup language. In: Intelligent Virtual Agents: 6th International Conference, IVA 2006, Marina Del Rey, CA, USA, August 21-23, 2006. Proceedings 6. Springer, pp. 205–217.

Kruger, J., Dunning, D., 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. J. Pers. Soc. Psychol. 77 (6), 1121.

Lawrence, N.W., 2014. Social Research Methods: Qualitative and Quantitative Approaches, Seventh ed. Pearson Education Limited, Harlow.

Lawshe, C.H., 1975. A quantitative approach to content validity. Pers. Psychol. 28, 563–575. http://dx.doi.org/10.1111/j.1744-6570.1975.tb01393.x.

Leichtmann, B., Nitsch, V., Mara, M., 2022. Crisis ahead? why human-robot interaction user studies may have replicability problems and directions for improvement. Front. Robot. AI 9, http://dx.doi.org/10.3389/frobt.2022.838116.

Li, F., Fitrianie, S., Bruijnes, M., Abdulrahman, A., Guo, F., Brinkman, W.-P., 2023. Mandarin Chinese translation of the artificial-social-agent questionnaire instrument for evaluating human-agent interaction. Front. Comput. Sci. 5, http://dx.doi.org/10.3389/fcomp.2023.1149305, URL https://www.frontiersin.org/articles/10.3389/fcomp.2023.1149305.

Li, F., Hu, J., Xie, K., He, T.-C., 2015. Authentication of experimental materials: A remedy for the reproducibility crisis? Genes & Dis. (ISSN: 2352-3042) 2 (4), 283. http://dx.doi.org/10.1016/j.gendis.2015.07.001.

Lisetti, C., Amini, R., Yasavur, U., Rishe, N., 2013. I can help you change! an empathic virtual agent delivers behavior change health interventions. ACM Trans. Manag. Inf. Syst. 4 (4), 1–28. http://dx.doi.org/10.1145/2544103.

Lord, F.M., Novick, M.R., 2008. Statistical Theories of Mental Test Scores. IAP.

Lugrin, B., Pelachaud, C., Traum, D. (Eds.), 2021. The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition, first ed. vol. 37, Association for Computing Machinery, New York, NY, USA.

Matthews, G., Deary, I.J., Whiteman, M.C., 2003. Personality traits, second ed. Cambridge University Press, http://dx.doi.org/10.1017/CBO9780511812736.

Mckeown, G., Valstar, M., Cowie, R., Pantic, M., Schroder, M., 2013. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. Affect. Comput. IEEE Trans. on 3, 5–17. http://dx.doi.org/10.1109/T-AFFC.2011.20.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In: Proc. of Neural Information Processing Systems 2013 - Volume 2. Curran Associates Inc., Red Hook, NY, USA, pp. 3111—-3119.

Mobley, A., Linder, S.K., Braeuer, R., Ellis, L.M., Zwelling, L., 2013. A survey on data reproducibility in cancer research provides insights into our limited ability to translate findings from the laboratory to the clinic. PloS One 8 (5), e63221. http://dx.doi.org/10.1371/journal.pone.0063221.

Mokkink, L.B., Terwee, C.B., Patrick, D.L., Alonso, J., Knol, P.W.S.D.L., Bouter, L.M., de Vet, H.C.W., 2010. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J. Clin. Epidemiol. 63 (7), 737–745. http://dx.doi.org/10.1016/j.jclinepi.2010.02.006.

Montgomery, D.C., Runger, G.C., 2003. Applied Statistics and Probability for Engineers, third ed. John Wiley and Son, Inc., Hoboken.

Moraila, G., Shankaran, A., Shi, Z., Warren, A.M., 2014. Measuring Reproducibility in Computer Systems Research. Technical report, University of Arizona.

Myers, I.B., McCaulley, M.H., 1985. Manual: A Guide to the Development and Use of the Myers-Briggs Type Indicator. Consulting Psychologists Press, Palo Alto, CA.

National Academies of Sciences, Engineering, and Medicine, 2019. Reproducibility and Replicability in Science. National Academies of Sciences, Engineering, and Medicine, pp. 39–54.

Nazari, Z., Lucas, G.M., Gratch, J., 2015. Opponent modeling for virtual human negotiators. In: Proc. of Intelligent Virtual Agents 2015. 9238, Springer International Publishing, Delft, Netherlands, pp. 39–49. http://dx.doi.org/10.1007/978-3-319-21996-7.

Neuman, W.L., 2013. Social research methods: qualitative and quantitative approaches. Always learning, Pearson Education, URL https://books.google.nl/books?id=Ybn3ngEACAAJ.

Nielsen, J., 2018. Card sorting: Uncover users' mental models for better information architecture. https://www.nngroup.com/articles/card-sorting-definition/. (Accessed January 2024).

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349 (6251), aac4716. http://dx.doi.org/10.1126/science.aac4716.

Pejsa, T., Gleicher, M., Mutlu, B., 2017. Who, me? How virtual agents can shape conversational footing in virtual reality. In: Proc. of Intelligent Virtual Agents. In: Lecture Notes in Computer Science, 10498, Springer, pp. 347–359. http://dx.doi.org/10.1007/978-3-319-67401-8_45.

Polkosky, M.D., Lewis, J.R., 2003. Expanding the MOS: Development and psychometric evaluation of the MOS-R and MOS-X. Int. J. Speech Technol. 6 (2), 161–182. http://dx.doi.org/10.1023/A:1022390615396.

Protzko, J., Krosnick, J., Nelson, L., Nosek, B.A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C.R., Lundmark, S., et al., 2023. High replicability of newly discovered social-behavioural findings is achievable. Nat. Hum. Behav. 1–9.

Ranjbartabar, H., Richards, D., 2016. A virtual emotional freedom therapy practitioner: (demonstration). In: Proc. of the 2016 International Conference on Autonomous Agents & Multiagent Systems. AAMAS '16, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, pp. 1471–1473.

Rönkkö, M., Cho, E., 2022. An updated guideline for assessing discriminant validity. Organ. Res. Methods 25 (1), 6–14. http://dx.doi.org/10.1177/1094428120968614.

Rosenkoetter, U., Tate, R.L., 2018. Assessing features of psychometric assessment instruments: A comparison of the COSMIN checklist with other critical appraisal tools. Brain Impair. 19 (1), 103–118. http://dx.doi.org/10.1017/BrImp.2017.29.

Sakamoto, D., Ishiguro, H., 2009. Geminoid: Remote-controlled android system for studying human presence. Kansei Eng. Int. 8 (1), 3–9. http://dx.doi.org/10.5057/er081218-1.

Scherer, K.R., 2005. What are emotions? And how can they be measured? Soc. Sci. Inf. 44 (4), 695–729. http://dx.doi.org/10.1177/0539018405058216.

Shneiderman, B., 2020. Human-centered artificial intelligence: Three fresh ideas. AIS Trans. Human- Comput. Interact. 12 (3), 109–124.

Sieg, A., 2018. Text similarities : Estimate the degree of similarity between two texts. https://medium.com/@adriensieg/text-similarities-da019229c894. (Accessed April 2020).

Stevens, J.P., 2009. Applied Multivariate Statistics for the Social Sciences, 5th Ed. Routledge, http://dx.doi.org/10.4324/9780203843130.

Tran, T.Q., Langlotz, T., Young, J., Schubert, T.W., Regenbrecht, H., 2024. Classifying presence scores: Insights and analysis from two decades of the igroup presence questionnaire (IPQ). ACM Trans. Comput.- Hum. Interact. http://dx.doi.org/10.1145/3689046.

Valstar, M., Baur, T., Cafaro, A., Ghitulescu, A., Potard, B., Wagner, J., André, E., Durieu, L., Aylett, M., Dermouche, S., Pelachaud, C., Coutinho, E., Schuller, B., Zhang, Y., Heylen, D., Theune, M., Waterschoot, J.v., 2016. Ask alice: An artificial retrieval of information agent. In: Proc. of the 18th ACM International Conference on Multimodal Interaction. ICMI '16, Association for Computing Machinery, New York, NY, USA, pp. 419—-420. http://dx.doi.org/10.1145/2993148.2998535.

Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D., 2003. User acceptance of information technology: Toward a unified view. MIS Q. 425–478.

van Welbergen, H., Reidsma, D., Ruttkay, Z.M., Zwiers, J., 2009. Elckerlyc: A BML realizer for continuous, multimodal interaction with a virtual human. J. Multimodal User Interfaces 3 (4), 271–284.

Zhang, Z., Bickmore, T., 2018. Medical shared decision making with a virtual agent. In: Proc. of Intelligent Virtual Agents. ACM, New York, NY, USA, pp. 113–118. http://dx.doi.org/10.1145/3267851.3267883.

Zhao, R., Romero, O.J., Rudnicky, A., 2018. SOGO: a social intelligent negotiation dialogue system. In: Proc. of Intelligent Virtual Agents. ACM, pp. 239–246. http://dx.doi.org/10.1145/3267851.3267880.