



OPEN ACCESS

EDITED BY

Athanasios Drigas,
National Centre of Scientific Research
Demokritos, Greece

REVIEWED BY

Aikaterini Doulou,
National Centre of Scientific Research
Demokritos, Greece
Victoria Bamicha,
National Centre of Scientific Research
Demokritos, Greece

*CORRESPONDENCE

Siska Fitrianie
✉ s.fitrianie@tudelft.nl

RECEIVED 15 October 2025

REVISED 03 December 2025

ACCEPTED 08 December 2025

PUBLISHED 04 March 2026

CITATION

Fitrianie S, Abdulrahman A, Buijnes M and
Brinkman W-P (2026) Establishing reference
points for artificial social agent evaluation: the
ASAQ representative set 2025.
Front. Comput. Sci. 7:1726078.
doi: 10.3389/fcomp.2025.1726078

COPYRIGHT

© 2026 Fitrianie, Abdulrahman, Buijnes and
Brinkman. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Establishing reference points for artificial social agent evaluation: the ASAQ representative set 2025

Siska Fitrianie^{1*}, Amal Abdulrahman¹, Merijn Buijnes² and
Willem-Paul Brinkman¹

¹Faculty of Electrical Engineering, Mathematics and Computer Science, Department of Intelligent Systems, Delft University of Technology, Delft, Netherlands, ²Faculty of Law, Economics and Governance, Utrecht University, Utrecht, Netherlands

KEYWORDS

artificial social agent, evaluation instrument, normative dataset, questionnaire, user study

1 Introduction

Artificial Social Agents (ASAs) such as conversational agents, robots, and virtual agents are approached as social actors because they can interact with people using verbal and non-verbal communication, including language, gestures, gaze, and emotional expression. Research shows that people form perceptions, beliefs, and expectations about ASAs through the experience of their interaction, similar to how they form impressions of other humans (e.g., [Lugrin et al., 2022](#); [Norouzi et al., 2018](#)). Studying the interaction experience reveals how people understand an ASA's behavior in terms of how it affects them, which is a prerequisite for how an interaction might affect task-related goals. For example, being motivated by an ASA to become physically active, engage with educational material, or immerse in entertainment.

Benchmarking this interaction experience helps our community to understand the progress we are making in our research and development of ASAs. Studying these interaction experiences, i.e., mental constructs, helps us to ground new measurements with existing ones. In other words, when we compare new results against a predefined set of normative data, we create a shared community understanding and strengthen the cohesion of our research. Looking at other research areas, we see representative data sets being offered alongside psychometric instruments, such as for personality inventories [e.g., IPIP-NEO-120 ([Johnson, 2014](#)) and a related data set ([Kajonius and Johnson, 2019](#))], assessment tools for depression [e.g., Beck Depression Inventory ([Beck et al., 1961](#)) and a normative set ([Roelofs et al., 2013](#))], health [e.g., SF-36 Health Survey ([Ware, 1999](#)) and its normative sets ([Pappa et al., 2005](#); [Roser et al., 2019](#); [Swift et al., 2022](#))], and intelligence [e.g., the Stanford-Binet Intelligence Scales ([Roid and Pomplun, 2012](#)) and a normative dataset ([Stevens and Bernier, 2021](#))]. But, also closer at home, when it comes to evaluation of software, System Usability Scale (SUS) ([Brooke, 1996](#)) also comes with a representative data set ([Lewis and Sauro, 2018](#)).

Creating benchmark set to go along with ASA Questionnaire (ASAQ) ([Fitrianie et al., 2025a,b](#)) allows us to benchmark peoples experience with an ASA. This is measured on 24 constructs and dimensions covering an extensive part of our community shared interests, such as believability, likeability, and sociability of ASA. ASAQ has been published alongside the norm set "ASAQ representative set 2024," which includes the experience of 1,066 individuals with 29 agents. That set is based on a third person perspective, i.e., filling out a questionnaire after seeing a video of someone else interacting with an agent. Although pragmatic for validating the questionnaire, the ASAQ authors also acknowledge possible

limitations of this set on generalization toward experiences based on actual interaction (Fitriani et al., 2025a).

A key question when developing a benchmark is what should constitute as a benchmark? Which people should be included in the sample, and which agents? For ASAQ representative set 2024, the research platform Prolific was used, which allows data collection across the world. When using this platform to develop our benchmarking set, we need to know which agents are publicly available that have a global reach and have a sizeable user group. Therefore, our first step in building the benchmark set was to survey contemporary ASA usage.

2 Method

We recruited participants for this study through the crowdsourcing platform, Prolific, between November 30 and December 19, 2023. For this, we applied the following inclusion criteria, where eligible participants were those who: (1) had not taken part in prior ASAQ validation studies, (2) had a Prolific approval rate above 95%, and (3) were proficient in English. Recruitment spanned multiple time zones, with a staggered approach in 6-h intervals to elicit global participant distribution. The study consisted of two sequential phases: (1) screening the population for familiarity with contemporary ASAs, and (2) establishing the ASAQ Representative Set 2025. For this study, we received approval from the university human research ethics Committee (no. 2685, dated 13 January 2023), preregistered the study (Fitriani et al., 2023), and made the analysis script and data publicly available (Fitriani et al., 2026). We compensated participants according to Prolific's payment guidelines.

To develop a benchmark set based on individuals' interaction experiences with widely known agents, we started by creating an initial agent list using input from the OSF working group on Artificial Social Agent Evaluation Instrument.¹ Twelve workgroup members from all over the world brainstormed on popular and widely used ASAs, selecting agents that various people, e.g., age groups and locations, might have interacted with at home. This resulted in a pre-selection of 11 agents, namely: Amazon's Alexa, Google's Bard chatbot, Microsoft's Bing chatbot, OpenAI's ChatGPT, Microsoft's CoPilot, Android's Google Assistant, IKEA's customer service chatbot, Replika chatbot, Apple's Siri, iRobot's Roomba vacuum cleaner, and Microsoft's Xiaoice. To further diversify the agent group, we included a dog, asking some participants to complete the questionnaire based on their interactions with a dog. Furthermore, with an eye on the future, we also incorporated an online version of the classic Eliza chatbot (Weizenbaum, 1966), making it possible to expose people in the future to the same agent. Finally, we included a non-existent agent, "Xonderfloip," as a distractor check, resulting in a list of 14 agents. Participants were asked to indicate the timing of their last interaction with the agents, with options ranging from "today" to "never." Of the 1,296 individuals initially recruited, 1,253 participants responded "never" to interactions with the distractor

agent, meeting the criteria for inclusion in the subsequent phase of the study.

Allowing people to compare their agent with agents in the benchmark set, we aimed for a statistical power of 0.80 to detect at least a medium-sized effect in future independent *t*-tests with an alpha level of 0.05 (Cohen, 1992). Consequently, the benchmark set required a minimum of 64 samples per agent. To ensure participants had interacted with the agents recently, we only used agents used within the last 6 months, narrowing the agent group from 14 to 10. Including the Eliza chatbot and the dog, we selected the agents: Alexa, Bard, Bing, ChatGPT, CoPilot, Google Assistant, Roomba, and Siri. Participants were assigned to evaluate a single agent they were familiar with, or to interact with the Eliza chatbot for 5 min before assessment to establish their own interaction experience with this agent. Exclusion criteria in this phase were: (1) failing more than 20% of attention checks; (2) providing incoherent responses to open-ended questions (e.g., unintelligible or non-sensical answers, or indicating no interaction with the assigned ASA); and (3) completing fewer than 10 dialogue turns for those assigned to the Eliza chatbot. Each participant was allowed to participate only once, with only their first completion included in the analysis.

Out of 1,253 available participants, we invited 777 individuals until we ended up with 666 participants who met the inclusion criteria [per agent: $M = 66$, $SD = 1$, range = [64–68]]. Among the exclusions, 47 participants did not complete the study with their assigned ASA, five failed attention checks (providing from 3 to 7 incorrect answers out of 10), and one was removed due to an open-ended response indicating no interaction with the assigned agent. Additionally, 58 participants assigned to the Eliza chatbot were excluded for completing fewer than 10 dialogue turns. Additionally, we requested participants to describe their experiences with the ASA to which they were assigned, in their own words, aiming for future research.

The resulting dataset included participants from the two phases: Phase 1 ($n = 1,253$) and a subset of these participants in Phase 2 ($n = 666$). The majority of participants identified as male (Phase 1: 54.5%; Phase 2: 57.8%), followed by female (Phase 1: 44.9%; Phase 2: 41.9%), with a small proportion identifying as other (Phase 1: 0.6%; Phase 2: 0.3%). The mean age was similar across both Phases (Phase 1: $M = 30$, $SD = 9.2$; Phase 2: $M = 29.8$, $SD = 9.2$), with the largest age groups being 18–25 (Phase 1: 38.9%; Phase 2: 39.8%) and 26–35 (Phase 1 and Phase 2: 39.6%). Education levels were comparable between groups, with the highest proportions holding an undergraduate degree (Phase 1 and Phase 2: 41.4%) or a graduate degree (Phase 1: 25%; Phase 2: 23.9%). Socioeconomic status, assessed via the MacArthur Scale (Adler et al., 2000) (1 = lowest, 10 = highest), was distributed across the scale, with the largest proportions in the middle ranges (e.g., at level 6, Phase 1: 25.9% at level 6; Phase 2: 28.4%). Geographically [based on the United Nations Regional Groups (United Nations, 2024)], most participants resided in Western Europe (Phase 1: 46.8%; Phase 2: 42.9%), followed by Africa (Phase 1: 21.1%; Phase 2: 22.8%) and Eastern Europe (Phase 1: 18.2%; Phase 2: 20.1%). Smaller proportions were from Latin America and the Caribbean (Phase 1: 11.7%; Phase 2: 12.2%), with limited presentation from the United States (Phase 1: 1.2%; Phase 2: 0.6%), and other regions.

¹ <https://osf.io/6duf7/>

TABLE 1 Summary of participants' usage of the 13 ASAs participated between November 30 and December 13, 2023 ($n = 1,253$). The reported % of total any-use reported for each ASA, and when this use last occurred.

Agent	Phase: Screening agent population set								Phase: Establishing Dataset		
	Total users (%)	Today	This week	This month	Last half year	This year	More than a year ago	Never	ASAQ Score		
									<i>n</i>	Long	Short
Alexa	655 (52.27)	118	94	81	94	98	170	598	67	4	9
Bard	448 (35.75)	37	90	143	53	94	31	805	66	5	10
Bing	644 (51.40)	64	140	166	77	115	82	609	68	6	9
ChatGPT	1,121 (89.47)	295	362	240	106	102	16	132	67	0	8
CoPilot	203 (16.20)	17	53	48	24	33	28	1,050	67	7	13
A dog (animal)	854 (68.16)	546	91	53	25	28	111	399	66	30	32
Eliza	28 (2.23)	1	2	3	2	4	16	1,225	68	-30	-30
Google Assistant	1,066 (85.08)	180	273	240	108	150	115	187	67	0	6
IKEA	335 (26.74)	3	13	38	57	86	138	918			
Replika	131 (10.45)	3	5	9	24	37	53	1,122			
Roomba	340 (27.13)	24	73	48	40	56	99	913	64	-6	0
Siri	896 (71.51)	163	191	142	85	123	192	357	66	3	8
Xiaoice	100 (7.98)	5	8	16	9	35	27	1,153			

We present the ASAQ score only for the ASAs we measured ($n = 666$).

Users of this dataset might select sub-datasets based on these characteristics to study specific groups.

3 Data description

Table 1 provides an overview of participant interactions with 12 ASAs and a dog. ChatGPT emerged as the most widely used agent, with 89.47% of 1,253 participants reporting interactions. Google Assistant (85.08%) and Siri (71.51%) also demonstrated high usage rates. In contrast, less commonly used agents included Replika (10.45%), Xiaoice (7.98%), and Eliza (2.23%).

Among the ASAs, ChatGPT and Google Assistant exhibited the highest proportions of recent interactions (today and this week), reflecting their integration into daily life. For instance, 295 participants interacted with ChatGPT *today*, and 362 *this week*. As anticipated, agents such as Eliza showed minimal recent interactions, with the majority of participants reporting *never* having engaged with them (1,225).

The study generated a representative set of nine ASAs and a dog, collecting 666 unique participant ratings on the 90 first-person perspective items of the ASAQ. Sample sizes per agent ranged from 64 to 68. Analysis of the ASAQ long version revealed variability in the ASAQ scores across agents, ranging from -30 (Eliza) to +30 (the dog). The data set, showing a detailed presentation of the scores of the ASAs on each of the 24 constructs and dimensions of the ASAQ, can be accessed publicly online (Fitrianie et al., 2026). The ASAQ constructs and overall item content remained consistent with the ASAQ representative set 2024; the

only difference is the participants' point of view, with the 2024 set collected from a third-person perspective (watching a video of a human-ASA interaction) and the 2025-set from a first-person perspective (interacting directly with an ASA). Items reflect the relevant perspective [e.g., "The user can rely on [the agent]" vs. "I can rely on [the agent]"]. The ASAQ construct and dimension scores, derived from both the long and short versions of the ASAQ, for all agents in the Representative Set 2025 are provided in Supplementary Tables S1–S4 accompanying this article.²

4 Recommendation for future use

The ASAQ Representative set 2025 extends the previously established ASAQ representative set 2024, offering an enhanced resource for researchers. The dataset highlights the varying interaction experiences people have in direct interaction with well-known agents. The reported use of contemporary ASAs (e.g., ChatGPT, Google Assistant, and Siri) demonstrates how rapidly conversational agents have become embedded in daily life. The inclusion of a non-artificial social agent (a dog) adds depth to the dataset, allowing for comparisons to other social experiences. Additionally, the variability in ASAQ scores, ranging from -30 for Eliza to +30 for dogs, provides anchor points for researchers to compare their own ASA against when using the ASAQ. Furthermore, the dataset allows for the ranking

² For all resources and updates, visit the ASAQ project website <https://asaq.ewi.tudelft.nl>.

of results across each ASAQ construct or dimension relative to the agents included in the ASAQ Representative Set. To facilitate analysis, researchers can utilize ASAQ charts, which offer a clear, at-a-glance visualization of their ASAs scores across all 24 constructs/dimensions, enabling direct comparisons with the representative ASAs. This resource promotes robust and standardized reporting in studies focused on human-agent interactions, which advances methodological consistency in the field.

Fitriani et al. (2025a) provide guidelines on leveraging the ASAQ, in third person perspective. It also includes recommendations for selecting between the long and short versions of the ASAQ (based on experimental paradigms and the scale of the survey) and determining appropriate sample sizes for studies. With the here presented dataset, it is possible to create similar guidelines for the first person perspective use of the ASAQ.

Two limitations about this dataset should be noted. First, apart from Eliza, participants evaluated ASAs based on their most recent interaction, which relies on recall and may introduce bias due to differences in time since use, ASA version, and interaction context. Second, participants were recruited through Prolific Academic, which is available only in selected countries and as such is not representative of the global population. These factors limit the generalizability of the ASAQ results to other demographic groups or controlled settings.

For future research, as ASAs evolve quickly, ASAQ norm values may require periodic updating to remain accurate and relevant. Therefore, we encourage researchers to include the dataset of ASAQ measurements of their ASAs. This enables systematic comparison of emerging technologies and designs against established systems. Continued ASAQ dataset-releases facilitate recalibration of interpretations, as the next popular or successful ASA might become the next anchor point against which other researchers compare their work. Thus, releasing the dataset of your ASA experiments using ASAQ, increases the relevance of your work and allows for reflection of evolving trends in ASA development and usage.

Data availability statement

The datasets presented in this study can be found in <http://doi.org/10.4121/638738c4-3905-4e71-8364-523c74af7ee8>.

Ethics statement

The studies involving humans were approved by the Human Research Ethics Committee, Delft University of Technology (Approval No. 2685, dated January 13, 2023). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

SF: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. AA: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Validation, Writing – review & editing. MB: Conceptualization, Methodology, Validation, Visualization, Writing – review & editing. W-PB: Conceptualization, Formal analysis, Funding acquisition, Methodology, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. Parts of this work were funded by the 4TU Pride and Prejudice project.

Acknowledgments

We acknowledge the efforts of all OSF workgroup members, and in particular mention the large recurring efforts of Nele Albers, Andrea Bönsch, Jonathan Ehret, Fengxiang Li, and Deborah Richards. The authors also thank the reviewers for their valuable suggestions.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. 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.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of

their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fcomp.2025.1726078/full#supplementary-material>

References

- Adler, N. E., Epel, E. S., Castellazzo, G., and Ickovics, J. R. (2000). Relationship of subjective and objective social status with psychological and physiological functioning: preliminary data in healthy, white women. *Health Psychol.* 19, 586–592. doi: 10.1037/0278-6133.19.6.586
- Beck, A., Ward, C. H., Mendelson, M., Mock, J., and Erbaugh, J. (1961). An inventory for measuring depression. *Arch. Gen. Psychiatry* 4, 561–571. doi: 10.1001/archpsyc.1961.01710120031004
- Brooke, J. (1996). “SUS: a ‘quick and dirty’ usability scale,” in *Usability Evaluation In Industry, 1st Edn.*, eds. P. W. Jordan, B. Thomas, B. A. Weerdmeester, and I. L. McClelland (London: CRC Press; Taylor and Francis), 189–194. doi: 10.1201/9781498710411
- Cohen, J. (1992). Quantitative methods in psychology: a power primer. *Psychol. Bull.* 112, 155–159. doi: 10.1037/0033-2909.112.1.155
- Fitrianie, S., Abdulrahman, A., Bruijnes, M., and Brinkman, W.-P. (2026). Data and analysis underlying the research into establishing reference points for artificial social agent evaluation: the ASAQ representative set 2025. *4TU.ResearchData*. doi: 10.4121/638738c4-3905-4e71-8364-523c74af7ee8
- Fitrianie, S., Bruijnes, M., Abdulrahman, A., and Brinkman, W.-P. (2025a). The Artificial Social Agent Questionnaire (ASAQ) - development and evaluation of a validated instrument for capturing human interaction experiences with artificial social agents. *Int. J. Hum. Comput. Stud.* 199:103482. doi: 10.1016/j.ijhcs.2025.103482
- Fitrianie, S., Bruijnes, M., Abdulrahman, A., Li, F., and Brinkman, W.-P. (2023). *Study 9: Concurrent Validation and a Normative Dataset Development*. OSF Registries. Available online at: <https://archive.org/details/osf-registrations-6gz29-v1> (Accessed January 10, 2026).
- Fitrianie, S., Bruijnes, M., Abdulrahman, A., Li, F., and Brinkman, W.-P. (2025b). Artificial Social Agent Questionnaire (ASAQ). *APA PsycTests*. doi: 10.1037/t95286-000
- Johnson, J. A. (2014). Measuring thirty facets of the five factor model with a 120-item public domain inventory: development of the IPIP-NEO-120. *J. Res. Pers.* 51, 78–89. doi: 10.1016/j.jrp.2014.05.003
- Kajonius, P. J., and Johnson, J. A. (2019). Assessing the structure of the five factor model of personality (IPIP-NEO-120) in the public domain. *Eur. J. Psychol.* 15:260. doi: 10.5964/ejop.v15i2.1671
- Lewis, J. R., and Sauro, J. (2018). Item benchmarks for the system usability scale. *J. Usabil. Stud.* 13, 158–167. doi: 10.5555/3294033.3294037
- Lugrin, B., Pelachaud, C., and Traum, D. (eds.) (2022). *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application, 1st Edn.*, vol. 48. New York, NY: Association for Computing Machinery. doi: 10.1145/3563659
- Norouzi, N., Kim, K., Hochreiter, J., Lee, M., Daher, S., Bruder, G., et al. (2018). “A systematic survey of 15 years of user studies published in the Intelligent Virtual Agents Conference,” in *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (New York, NY: Association for Computing Machinery), IVA’18, 17–22. doi: 10.1145/3267851.3267901
- Pappa, E., Kontodimopoulos, N., and Niakas, D. (2005). Validating and norming of the Greek SF-36 health survey. *Qual. Life Res.* 14, 1433–1438. doi: 10.1007/s11136-004-6014-y
- Roelofs, J., van Breukelen, G., de Graaf, L. E., Beck, A. T., Arntz, A., and Huibers, M. J. H. (2013). Norms for the Beck Depression Inventory (BDI-II) in a large dutch community sample. *J. Psychopathol. Behav. Assess.* 35, 93–98. doi: 10.1007/s10862-012-9309-2
- Roid, G. H., and Pomplun, M. (2012). “The Stanford-Binet intelligence scales,” in *Contemporary Intellectual Assessment: Theories, Tests, and Issues, 5th Edn.*, vol. 656, eds. D. P. Flanagan and P. L. Harrison (New York, NY: The Guilford Press), 249–268.
- Roser, K., Mader, L., Baenziger, J., Sommer, G., Kuehni, C. E., and Michel, G. (2019). Health-related quality of life in Switzerland: normative data for the SF-36v2 questionnaire. *Qual. Life Res.* 28, 1963–1977. doi: 10.1007/s11136-019-02161-5
- Stevens, A., and Bernier, R. (2021). “Stanford-Binet intelligence scales and revised versions,” in *Encyclopedia of Autism Spectrum Disorders*, ed. F. R. Volkmar (Cham: Springer International Publishing), 4604–4607. doi: 10.1007/978-3-319-91280-6_754
- Swift, B., Naci, H., Taneri, B., Becker, C. M., Zondervan, K. T., and Rahmioglu, N. (2022). The Cyprus women’s health research (COHERE) initiative: normative data from the SF-36v2 questionnaire for reproductive aged women from the Eastern Mediterranean. *Qual. Life Res.* 31, 2011–2022. doi: 10.1007/s11136-022-03100-7
- United Nations (2024). *Regional Groups of Member States* (Accessed November 06, 2024).
- Ware, J. E. J. (1999). “SF-36 health survey,” in *The Use of Psychological Testing for Treatment Planning and Outcomes Assessment, 2nd Edn.*, ed. M. E. Maruish (Lawrence Erlbaum Associates Publishers), 1227–1246.
- Weizenbaum, J. (1966). Eliza a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 36–45. doi: 10.1145/365153.365168

NOMENCLATURE

Resource Identification Initiative
Artificial Social Agent Questionnaire (ASAQ), [RRID:SCR_027534](https://n2t.org/RRID:SCR_027534).