Questionnaire Items for Evaluating Artificial Social Agents - Expert Generated, Content Validated and Reliability Analysed

Fitrianie, Siska; Bruijnes, Merijn; Li, Fengxiang; Brinkman, Willem Paul

DOI
10.1145/3472306.3478341

Publication date
2021

Document Version
Final published version

Published in
Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, IVA 2021

Questionnaire Items for Evaluating Artificial Social Agents - Expert Generated, Content Validated and Reliability Analysed

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

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

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

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

ABSTRACT

In this paper, we report on the multi-year Intelligent Virtual Agents (IVA) community effort, involving more than 90 researchers worldwide, researching the IVA community interests and practice in evaluating human interaction with an artificial social agent (ASA). The joint efforts have previously generated a unified set of 19 constructs that capture more than 80% of constructs used in empirical studies published in the IVA conference between 2013 to 2018. In this paper, we present expert-content-validated 131 questionnaire items for the constructs and their dimensions, and investigate the level of reliability. We establish this in three phases. Firstly, eight experts generated 431 potential construct items. Secondly, 20 experts rated whether items measure (only) their intended construct, resulting in 207 content-validated items. Next, a reliability analysis was conducted, involving 192 crowd-workers who were asked to rate a human interaction with an ASA, which resulted in 131 items (about 5 items per measurement, with Cronbach's alpha ranged [.60 − .87]). These are the starting points for the questionnaire instrument of human-ASA interaction.

CCS CONCEPTS

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

KEYWORDS

Artificial social agent; user study; evaluation instrument; questionnaire; reliability analysis

ACM Reference Format:
Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, and Willem-Paul Brinkman. 2021. Questionnaire Items for Evaluating Artificial Social Agents - Expert Generated, Content Validated and Reliability Analysed. In *21th ACM International Conference on Intelligent Virtual Agents (IVA '21), September 14–17, 2021, Virtual Event, Japan.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3472306.3478341

IVA '21, September 14–17, 2021, Virtual Event, Japan

ACM ISBN 978-1-4503-8619-7/21/09.
https://doi.org/10.1145/3472306.3478341

INTRODUCTION

In this paper, we show our progress towards a validated and standardised measurement instrument (i.e. a questionnaire) for evaluating human interaction with an artificial social agent (ASA). The work presented in this paper is part of a larger effort that includes all sub-fields of the ASA community. Currently, over 90 people (self-selected to) have participated in the Open Science Framework work group "Artificial Social Agent Evaluation Instrument".[1] In previous work [3, 4], we investigated in which constructs the IVA community is interested, resulting in 19 unified constructs (and 15 dimensions), which are related to the the interaction between the user and ASA (see Figure 1). Constructs are, for example, Agent's Believability, Agent's Sociability, User-Agent Interplay, User-Agent Alliance, User's Trust and User's Engagement (for more details see [4]).

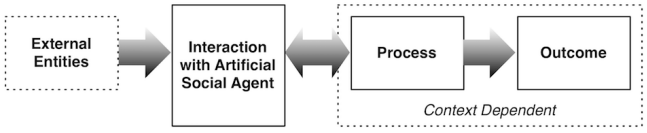

Figure 1: A world model of human-ASA interactions: the instrument will measure only the 'Interaction with ASA'.

In the current paper we describe the initial set of questionnaire items for each of these constructs (and their dimensions). The developed items are applicable for a variety of ASAs, ranging from chatbots and computer-controlled virtual humanoid agents to virtual and physical social robots. Our approach presented in this paper consists of three steps: 1) generate questionnaire items for each of the constructs, 2) determine the content-validity of the items, and 3) determine the reliability of the items. The pre-registration, data, and analyses of this work are publicly available at our Open Science Framework-repository.

[1] Join the work group's efforts at: https://osf.io/6duf7/

STEP 1 - GENERATE ITEMS

Members of the work group were invited, as independent experts, to propose as many items as possible for the constructs and dimensions identified in [4]. Their instructions were that items should: 1) be answerable on an interval scale labelled 'agree-disagree'; 2) be answerable for the different types of ASA that researchers from the community currently investigate; 3) be formulated as a singular statement; 4) be formulated in such a way that they can easily be changed so that they can be answered by a person who interacted with an agent (i.e., first person point of view) and by someone who observed an interaction with an agent (i.e., third point of view); 5) not refer to a particular physical part or modality or function of an ASA; and 6) not limit to a particular task of ASA.

The experts (n = 8) generated 431 new items (on average 17 items per construct (or dimension)). Next, three judges continued to check whether the generated items adhered to the instructions and, if necessary, reformulated the items (for example, into statements that are easy to understand and grammatically correct). Edits were discussed until unanimous agreement was reached, and the discussions are documented in the Open Science Framework repository.[2] The resulting 431 items were included in the next step.

STEP 2 - CONTENT-VALIDITY

In this step, work group members (n = 20) evaluated, as independent experts, the content-validity of the questionnaire items. The panel of experts assessed whether items could effectively measure the construct for which they were intended (based on Lawshe [5]). The rationale was to keep items that were found appropriate to measure a construct, and remove items that were found appropriate for multiple constructs or not appropriate for the construct for which they were intended.

The evaluation was broken into tasks (each task was performed on average by 10 experts; ranging from 8 to 15 experts per task). Each task showed the name and definition of a construct and four items: two items written for that construct (target items) and two items written for a different construct (distractor items). The experts selected all items that, in their view, would effectively measure the construct with an yes/no answer per item. Based on the design of the tasks, for each item, we counted: 1) True Positive (TP), times an item is intended and identified as a target; 2) True Negative (TN), times the item is intended and identified as a distractor; 3) False Positive (FP), times the item is intended as a distractor, but identified as a target; and 4) False Negative (FN), times the item is intended a target, but identified as a distractor. We corrected for (50%) chance on answering as intended on target items, TP_c, and on distractor items, FP_c:

$$TP_c = \frac{TP - 0.5(TP + FN)}{0.5(TP + FN)} \quad (1)$$

$$FP_c = \frac{FP - 0.5(TN + FP)}{0.5(TN + FP)} \quad (2)$$

When an item was intended as a target for a construct, the TP_c value above .40 was regarded as at least a moderate level of measuring the construct. For example, when 12 experts rated an item intended as a target, and if 10 recognised it as such (TP = 10) and two not (FN = 2), TP_c = .67, this item was not rejected on this ground. In contrast, when the item was included as a distractor in a task for another construct, the FP_c value above −.40 indicated a cause for concern as it was associated with an unintended construct. For example, when 11 experts were confronted with an item intended as a distractor, and if 2 rated the item as appropriate for the non-intended construct (FP = 2) and 9 did not (TN = 9), FP_c = −.64, this item would not be rejected on this ground.

Next, from the remaining items three judges selected the 'best' items for each construct based on the TP_c and FP_c values (aiming for about eight items, thus balancing coverage of the construct and reducing the number of items per construct). When more than eight items remained in a construct, a similarity test was conducted to measure how similar or dissimilar the item-texts were using a combination of the Word2Vec embedding [6], smooth inverse frequency [1], and cosine similarity methods [7]. The items that were the most dissimilar from the other items for that construct were selected. Additionally, the three judges compared the items' semantic, lexical, and pragmatic sides to select the items with the most distinctive meaning, the most distinctive choice of words, and that are easiest to understand. All decisions were discussed until unanimous agreement was reached. The discussions are documented in the Open Science Framework repository.[3]

The resulting 207 items for 26 constructs and dimensions (7-8 items each) were regarded to have an acceptable expert-content-validity (TP_c M = .89, SD = .14, range [.46..1]; FP_c M = −.75, SD = .18, range [−1.. − .43]).

STEP 3 - RELIABILITY ANALYSIS

In this step, the goal was to select the items that show an adequate reliability within their construct. For this, we recruited participants (n = 192) from the online crowd-worker platform Prolific to rate an interaction between a human user and an ASA on the 207 construct items. Participants were paid according to the platform's standards. This study was approved by the data management officer and the Human Research Ethics Committee TUDelft (no. 1402 (18-12-2020)).

Participants viewed one 30-second video of a human-ASA interaction (i.e., robot ASIMO (Advanced Step in Innovative Mobility) by Honda). They were randomly assigned to rate half of the questionnaire items, which were adapted to either the first-person point of view (e.g. *The agent and I look alike*) or to the third-person point of view (e.g. *The agent and the user look alike*). Items were rated on a 7-point scale from 'disagree' (value -3) to 'agree' (value 3) with a middle point 'neither agree nor disagree' (value 0). Finally, participants were included for analyses if they passed 12 out of 15 attention-check questions. The order of items and check-questions was random. Related data and files to this study are available at the Open Science Foundation-repository.[4]

Four judges removed the items that showed an unacceptable level of correlation within their construct. The judges kept at least 5 items per construct/dimension (aiming in balancing the coverage and reducing the number of items within the constructs), as this study involved only one ASA and the results might not generalise to other agents. The key factor for their decision was the reliability

[2] https://osf.io/32hfb/wiki/home/

[3] https://osf.io/qxeu5/wiki/home/

[4] https://osf.io/hyxwb/wiki/home/

coefficient of Cronbach's alpha, whereby a value below .60 was regarded as an unacceptable level of reliability [2]. The exclusion criteria for items were, after scores of reverse worded items were reversed:

(1) The item is negatively correlated with the total score of all items in the construct;
(2) The item's internal correlation with the total score of all items in the construct ($std.r$) is low (e.g. $< .50$), however, the removal of such an item should not reduce the alpha value;
(3) The item's absolute standardised mean difference (SMD) between the first and the third point of view is higher than $Q_3 + 1.5 * IQR$, with Q_3 as the third quartile and IQR as the interquartile range of the SMD between point of views across all items measured. A higher SMD shows that participants rate an item differently on the point of views; and
(4) The item correlated substantially with other constructs (e.g. $> .50$).

These four criteria were applied if: 1) the total number of items in the construct was > 5 (balancing coverage and reducing number of items per construct); 2) the alpha value of the construct was $>= .60$; and 3) the remaining set of items would still cover the theoretical domain that the construct intends to measure. Finally, the calculations were updated after removal of an item.

The resulting 131 items (5 to 6 items per construct/dimension) showed an average reliability of Cronbach's $\alpha = .76$, ($SD = .07$), ranging from .60 to .87. Using the classification of DeVellis [2], we observed: 3 out of 26 (11.5%) constructs are between .60 and .65 (undesirable); 2 (7.7%) are between .65 and .70 (minimally acceptable); 12 (46.2%) are between .70 and .80 (respectable); and 9 (34.6%) are between .80 and .90 (very good). The SMD score between point of views was small ($M = .23$, $SD = .22$, range [0 .. .97]), indicating that point of view differences might be limited. The items set includes 22 reverse-scored items (16.8%), and 54 items (40%) are point of view specific.

DISCUSSION

The work in this paper, toward the unified community supported questionnaire, has resulted in 131 expert-generated questionnaire items that are content-validated and demonstrated on average an respectable level of reliability. The next steps are:

(1) A confirmatory factor analysis to examine items' associations with the latent constructs, i.e., construct validity, involving a diverse set of ASAs;
(2) Establish the final item-set with the provision to create a long and short questionnaire version;
(3) Determine criteria validity and concurrent validity;
(4) Translate the questionnaire; and
(5) Develop a normative data set.

REFERENCES

[1] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *Proc. of International Conference on Learning Representations 2017.*
[2] Robert F. DeVellis. 2003. *Scale development: theory and applications.* Thousand Oaks, CA: Sage.
[3] Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Amal Abdulrahman, and Willem-Paul Brinkman. 2019. What Are We Measuring Anyway? - A Literature Survey of Questionnaires Used in Studies Reported in the Intelligent Virtual Agent Conferences. In *Proc. of IVA'19.* ACM NY USA, 159–161. https://doi.org/10.1145/3308532.3329421
[4] Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Andrea Bönsch, and Willem-Paul Brinkman. 2020. The 19 Unifying Questionnaire Constructs of Artificial Social Agents: An IVA Community Analysis. In *Proc. of IVA'20.* ACM NY USA, Article 21, 8 pages. https://doi.org/10.1145/3383652.3423873
[5] Charles H. Lawshe. 1975. A quantitative approach to content validity. *Personnel psychology* 28 (1975), 563–575.
[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proc. of Neural Information Processing Systems 2013 - Volume 2* (Lake Tahoe, Nevada). Curran Associates Inc., Red Hook, NY, USA, 3111—-3119.
[7] Andrien Sieg. 2018. Text Similarities : Estimate the degree of similarity between two texts. https://medium.com/@adriensieg/text-similarities-da019229c894. Accessed: April 2020.