What are we measuring anyway? A literature survey of questionnaires used in studies reported in the intelligent virtual agent conferences

Bruijnes, Merijn; Fitrianie, Siska; Richards, Deborah; Abdulrahman, A.; Brinkman, Willem-Paul

Publication date
2019

Document Version
Final published version

Published in
BNAIC/BENELEARN 2019

What are We Measuring Anyway? A Literature Survey of Questionnaires Used in Studies Reported in the Intelligent Virtual Agent Conferences

Merijn Bruijnes[1], Siska Fitrianie[1], Deborah Richards[2], Amal Abdulrahman[2], and Willem-Paul Brinkman[1]

[1] Delft University of Technology, Delft, The Netherlands
[2] Macquarie University, Sydney, NSW, Australia
{m.bruijnes,s.fitrianie,w.p.brinkman}@tudelft.nl
{deborah.richards@,amal.abdulrahman@hdr.}mq.edu.au

Abstract. Research into artificial social agents aims at constructing these agents and at establishing an empirically grounded understanding of them, their interaction with humans, and how they can ultimately deliver certain outcomes in areas such as health, entertainment, and education. Key for establishing such understanding is the community's ability to describe and replicate their observations on how users perceive and interact with their agents. In this paper, we address this ability by examining questionnaires and their constructs used in empirical studies reported in the intelligent virtual agent conference proceedings from 2013 to 2018. The literature survey shows the identification of 189 constructs used in 89 questionnaires that were reported across 81 papers. We found unexpectedly little repeated use of questionnaires as the vast majority of questionnaires (more than 76%) were only reported in a single paper. We expect that this finding will motivate joint effort by the IVA community towards creating a unified measurement instrument and in the broader AI community a renewed interest in replicability of our (user) studies.

In our paper [1], we open a discussion about methodological issues that exist in human-computer interaction (HCI) and specifically in the evaluation of Artificial Social Agents (ASA). ASAs, such as *intelligent virtual agents* (IVA), social robots, and chatbots, are computer controlled entities that can autonomously interact with humans following the social rules of human-human interactions. The motivation of the work is driven by the replication crisis in the social and life sciences (e.g. [2]). Many of the methods employed by HCI researchers come from the fields that are currently in crisis. Hence, we ask the question "do our studies have similar issues?" A variety of ideas to improve research practices have been proposed and these ideas likely can be beneficial to the methods used in the field of HCI. Some actionable points leading to open and reproducible science are pre-registration of experiments, replication of findings, collaboration and education of researchers. While discussing each of these (and potentially more) issues is beyond the scope of this paper, it is clear that the replication crisis needs our attention. As we reflect on our methods it makes sense to discuss

in general our scientific methods and practices, we therefore welcome critical and constructive input, on this work, and in the discussion on methodology in HCI.[3]

In this paper we ask what is the IVA community currently measuring of the user-interaction experience. Although several measuring techniques exists, e.g. behavioural measures, physiological measures, and observational measures, we limit the scope to questionnaires because of their popularity. We conducted a literature survey and examined the reported questionnaires and their constructs. This we argue gives an insight into the ability to replicate results which requires agreement in what to measure and with what measuring instruments.

We examined questionnaires and their constructs used in user-studies reported in the intelligent virtual agent conference proceedings from 2013 to 2018. The use of evaluation questionnaires in IVA papers has increased over the past six years: from 16% (2013) to 63% (2018) of accepted papers. We identified 189 constructs used in 89 questionnaires that were reported across 81 papers. We found unexpectedly little repeated use of questionnaires as the vast majority of questionnaires (more than 76%) were only reported in a single paper. Only 7 questionnaires were used by more than two studies. Such diverse measurement instruments make comparisons over studies impossible.

Our work is part of a larger effort that includes all sub-fields of the ASA community and aims at developing a validated standardised questionnaire instrument for evaluating human interaction with ASAs. To achieve this, we plan multiple steps, including: (1) Determine the conceptual model (i.e. examine existing questionnaires and foster discussions among experts); (2) Determine the constructs and dimensions (i.e. check face validity among experts and grouping of existing constructs); (3) Determine an initial set of constructs items (i.e. content validity analysis: reformulate items into easy to understand and 'ASA-appropriate' questionnaire items); (4) Confirmatory factor analysis to examine construct validity; (5) Establish the final item set with the provision to create a long and short questionnaire version; (6) Determine criteria validity (i.e. predictive validity: agreement with predicted future observations) and concurrent validity (e.g. agreement with other 'valid' measures); (7) Translate the questionnaire; and (8) Develop a normative data set. We have set up an open work-group to share ideas and to help implement the necessary steps. Currently, over 80 people participate in the work-group's open science framework platform[3] and we encourage more people to join. Ultimately, this will help us to address the methodological issues that we, as a relatively young field, face.

References

1. Fitrianie, S., Bruijnes, M., Richards, D., Abdulrahman, A., Brinkman, W.P.: What are we measuring anyway? - a literature survey of questionnaires used in studies reported in the intelligent virtual agent conferences. In: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents. pp. 159–161 (2019)
2. Pashler, H., Wagenmakers, E.J.: Editors introduction to the special section on replicability in psychological science: A crisis of confidence? Perspectives on Psychological Science **7**(6), 528–530 (2012)

[3] Join our efforts at: https://osf.io/6duf7/