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

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

Merijn Bruijnes
University of Twente
Enschede, the Netherlands
m.bruijnes@utwente.nl

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

Amal Abdulrahman
Macquarie University
Sydney, NSW, Australia
amal.abdulrahman@hdr.mq.edu.au

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

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.

ACM Reference Format:
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 *ACM International Conference on Intelligent Virtual Agents (IVA '19), July 2–5, 2019, PARIS, France.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3308532.3329421

1 INTRODUCTION

In this paper 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) and social robots, are computer controlled entities that can autonomously interact with humans following the social rules of human-human interactions. In this paper, we show our progress in addressing one particular methodological issue by creating a meta model of the constructs that researchers are interested in when studying ASAs where we limit our scope to user evaluations in IVA research.

The motivation of this work is driven by the crisis of methodology that the social and life sciences are facing as the results of many scientific studies are difficult or impossible to replicate in subsequent investigation (e.g. [11]). The Open Science Collaboration [3] observed, for example, that the effect size of replications was about half of the reported original effect size and that whereas 97% of the original studies had significant results, only 39% of the replication studies had significant results. In fact it has been suggested that more than 50% of psychological research results might be false (i.e. theories hold no or very low verisimilitude) [7]. Many of the methods employed by Human-Computer Interaction (HCI) researchers come from the fields that are currently in a replication crisis. Hence, we ask the question "do our studies (in this paper we focus on user evaluations of intelligent virtual agents) have similar issues?"

A variety of ideas to improve research practices have been proposed and it is likely these ideas 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.

The main question of this paper is what is the IVA community currently measuring of the interaction experience? Although several measuring techniques exists, e.g. behavioral measures, physiological measures, and observational measures, in this paper 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 finish the paper with the future plans to establish a standardized ASA evaluation questionnaire. As we reflect on our methods it makes sense to discuss in general our scientific methods and practices, we therefore

IVA '19, July 2–5, 2019, PARIS, France

ACM ISBN 978-1-4503-6672-4/19/07.
https://doi.org/10.1145/3308532.3329421

welcome critical and constructive input, on this work, and in the discussion on methodology in HCI.[1]

2 METHOD

We reviewed and collected constructs from empirical user studies reported in IVA conferences from 2013 to 2018 (Figure 1). For this, the first author screened 215 full and short papers presenting empirical user studies at the abstract and full-text stages. After reading the abstracts, 152 papers conducting empirical analysis in their user studies were included. Then, 84 papers remained for inclusion after a full-text review, as these papers used questionnaire instruments for their empirical analysis. A second independent coder double coded 60 randomly selected papers (i.e. 10 papers for each year). Both coders have a computer science background. High inter-coder reliability was achieved in the abstract and full-text reviews, respectively, 95% (Cohen's kappa $\kappa = .87$) and 82% ($\kappa = .64$).

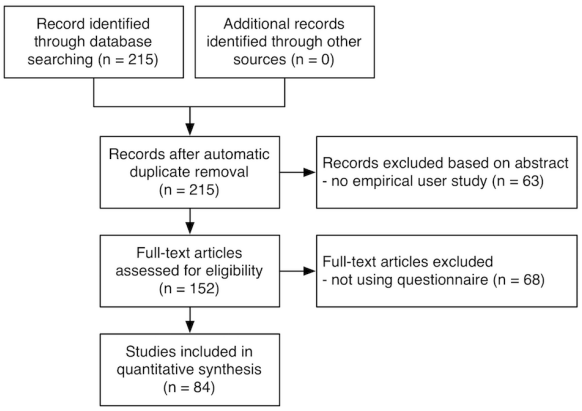

Figure 1: Prisma [9] diagram diagram of the screening process as completed by the first author

The next step was to extract the measurement constructs from the selected questionnaires. A construct expresses the specific phenomenon or aspect, e.g. naturalness, bonding, or trust, a questionnaire tries to capture. In a questionnaire, a construct is measured by a single or set of items. Some studies clearly provided their questionnaires with the measurement constructs. However, other studies only listed the goals they aimed to measure, only reported a set of questions, or simply gave references to existing questionnaires, often without clarifying which items had been included or whether any had been modified. We interpreted the measurement goal of these constructs based on the full-text assessment. Furthermore, we did not include those constructs that: (1) measured constructs that were predefined prior the interaction, e.g., user's demographics and context of interaction; and (2) measured context-dependent actions or behaviors (e.g. food intake or study time) or context-dependent expected results (e.g. weight loss, exam score). The first exclusion criterion aimed to exclude exogenous constructs that are determinants of the experience of interaction with ASA, while the second criterion aimed to exclude exogenous constructs where the interaction with ASAs could be seen as determinants for these factors. Based on these criteria, 3 papers were removed as all their constructs met exclusion criteria. We included constructs that at least measured interaction between a human user and artificial social agents. This covered different agent types, contexts, domains and development stages of IVA research. The list of collected constructs completed with their references and questionnaire items (if available) can be found in [4].

From the 81 papers that we extracted, 189 constructs measured human interaction with ASAs [4]. For each construct the name, its measurement goal, and the questionnaire items were retrieved. The questionnaire items of 70 constructs could not be found. The first author extracted all measurement constructs. To measure the reliability of these constructs, another independent coder with a computer science background double coded a sample of 25 papers. Independently the coders identified a total of 71 constructs, whereby 61 were identified by both coders, resulting in 86% agreement.

3 RESULTS

The use of evaluation questionnaires in IVA papers has increased over the past six years: from 16% (2013) to 63% (2018) of accepted full and short papers (Figure 2). From the 81 studies investigated, 25 developed new questionnaires while 56 studies used or adapted existing questionnaires. Existing questionnaires could be either previously published as an IVA paper or retrieved from other literature. In addition, 27 studies combined more than one of these existing questionnaires.

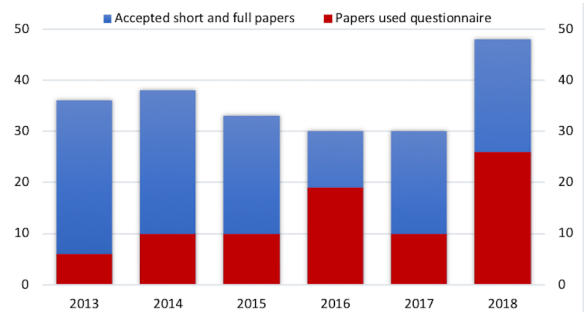

Figure 2: Number accepted IVA papers versus the number of studies using questionnaires from 2013 to 2018.

From 64 existing questionnaires applied in studies, only 7 of them were used by more than two studies, while 43 were used only once. Combining with studies that developed their own questionnaires, 68 questionnaires were used only once. Table 1 shows the most frequently used questionnaires and the number of times it was used. In particular, 14 studies reported by researchers from one research group consistently adapted one set of questions measuring the users' attitude towards the interaction with an agent and their satisfaction.

It appears that the number of new measurement constructs has not reached its saturation (Figure 3). Each year new constructs were introduced without observing any decline in this trend. In 2018 alone, 62 new constructs were introduced (which we extracted from 26 studies). Yet, some studies (15 out of 56 studies) used a combination of existing questionnaires with newly developed constructs.

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

Table 1: Most used questionnaires and usage frequency.

Questionnaire	Freq.
Relational Agent Group, e.g. [12][15]	14
The Working Alliance Inventory [6]	6
Social Presence [5]	5
Presence Scale [10]	4
Presence [13]	4
The Godspeed Questionnaire [1]	3
Warmth, Competence and Human-Like [2]	3

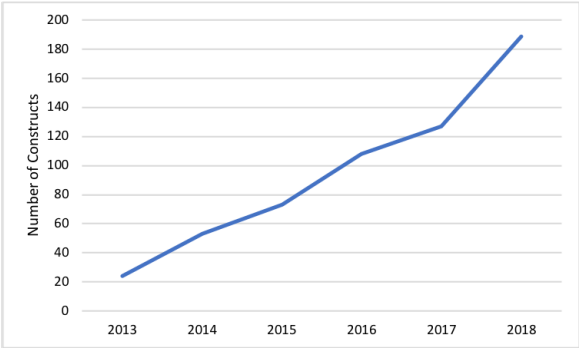

Figure 3: Cumulative frequency of new constructs introduced between 2013 and 2018.

There were variations in the scope of use of measurement constructs and in the level of abstraction of constructs. Some constructs aimed at measuring a specific aspect, while others assessed multiple aspects. An example of a specific construct measured the degree of the user's trust toward an agent (e.g. in [14]). On the other hand, the construct 'quality of interaction' assessed the overall quality of interaction [12]. This included, among other things, a questionnaire item about trust. Thus, one construct might measure different aspects within that single construct. In addition, the application of the same questionnaire might vary. For example, the Working Alliance Inventory questionnaire [6] was applied in [15] to measure the user's attitude toward an agent, while in [8] it was used to measure the collaboration and trust between the user and an agent. These observations have implications for the development of a theoretical model of measurement constructs that takes into account the relation between the identified constructs.

4 DISCUSSION AND CONCLUSION

We began this paper by noting that a methodological crisis is looming over the research in the field of HCI and specifically over the evaluation of IVAs. In this paper, we focus on the community's ability to make claims about how humans experience the interaction with an ASA. After examining the literature, we observed a relatively low level of reuse of questionnaires, and a continuous trend of studying new constructs. We therefore argue that we should move towards creating a unified measurement instrument, which researchers could use as a common base to measure a set of shared constructs to describe the interaction experience with an ASA.

The work presented in this paper is part of a larger effort that includes all sub-fields of the ASA community and aims at developing a validated standardized questionnaire instrument for evaluating human interaction with ASAs. To achieve this, we have put forward a plan consisting of multiple steps, including: (1) Determine the conceptual model (i.e. examine existing questionnaires and foster discussions among experts); (2) Determine the constructs and dimensions (i.e. check face validity among experts and grouping of existing constructs); (3) Determine an initial set of 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 70 people participate in the work-group's open science framework platform and we hope more people will join. Ultimately, this will help us to address the methodological issues that we, as a relatively young field, face.

REFERENCES

[1] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *Int. J. of Social Robotics* 1, 1 (2009), 71–81.

[2] Kirsten Bergmann, Friederike Eyssel, and Stefan Kopp. 2012. A Second Chance to Make a First Impression? How Appearance and Nonverbal Behavior Affect Perceived Warmth and Competence of Virtual Agents over Time. In *Proc. of IVA*. Springer, 126–138.

[3] Open Science Collaboration et al. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), aac4716.

[4] Siska Fitrianie. 2019. *List of Constructs and their Items: IVA Proceedings 2013 - 2018.* http://osf.io/yv6f4/.

[5] Chad Harms and Frank Biocca. 2004. Internal consistency and reliability of the networked minds measure of social presence. In *7th Annual International Workshop: Presence.*

[6] Adam O. Horvath and Leslie S. Greenberg. 1989. "Development and validation of the working alliance inventory". *J. of Counseling Psychology* 36, 2 (1989), 223–233.

[7] John PA Ioannidis. 2005. Why most published research findings are false. *PLoS medicine* 2, 8 (2005), e124.

[8] Natasha Jaques, Daniel McDuff, Yoo Lim Kim, and Rosalind Picard. 2016. Understanding and Predicting Bonding in Conversations Using Thin Slices of Facial Expressions and Body Language. In *Proc. of IVA*. Springer, 64–74.

[9] David Moher, Alessandro Liberati, and Douglas G. Altman Jennifer Tetzlaff. 2009. *PRISMA 2009 Flow Diagram.* Prisma statement. PMID: 19621072.

[10] Kristine L. Nowak and Frank Biocca. 2003. The Effect of the Agency and Anthropomorphism on Users' Sense of Telepresence, Copresence, and Social Presence in Virtual Environments. *Presence: Teleoperators and Virtual Environments* 12, 5 (2003), 481–494.

[11] Harold Pashler and Eric-Jan Wagenmakers. 2012. EditorsâĂŹ introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science* 7, 6 (2012), 528–530.

[12] Ameneh Shamekhi, Mary Czerwinski, Gloria Mark, Margeigh Novotny, and Gregory A. Bennett. 2016. An Exploratory Study Toward the Preferred Conversational Style for Compatible Virtual Agents. In *Proc. of IVA*. Springer, 40–50.

[13] Bob G. Witmer and Michael J. Singer. 1998. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoper. Virtual Environ.* 7, 3 (1998), 225–240.

[14] Jason Wu, Sayan Ghosh, Mathieu Chollet, Steven Ly, Sharon Mozgai, and Stefan Scherer. 2018. NADiA: Neural Network Driven Virtual Human Conversation Agents. In *Proc. of IVA*. 173–178.

[15] Zhe Zhang, Timothy Bickmore, Krissy Mainello, Meghan Mueller, Mary Foley, Lucia Jenkins, and Roger A. Edwards. 2014. Maintaining Continuity in Longitudinal, Multi-method Health Interventions Using Virtual Agents: The Case of Breastfeeding Promotion. In *Proc. of IVA*. Springer, 504–513.