Evaluation of Chatbots Usability Experimentation

Ranci Ren
Dep. Ing. Informática
Univ. Autónoma de Madrid
Madrid, Spain
ranci.ren@estudiante.uam.es

*John W. Castro**
Dep. Ing. Informática y Ciencias de la Computación
Universidad de Atacama
Copiapó, Chile
john.castro@uda.cl

Silvia T. Acuña
Dep. Ing. Informática
Univ. Autónoma de Madrid
Madrid, Spain
silvia.acunna@uam.es

***Abstract*—Context: The interest in developing chatbots is on the rise as the usability evaluation is an essential step in the chatbot development process; the number of experimental studies of chatbot usability has grown as well. Objective: Aggregating and concluding the features and metrics used to evaluate the usability of chatbots in experiments, to identify the state of the art of chatbots usability experimentation. Method: A systematic mapping study has been conducted, searching in five scientific databases. Results: Of 363 papers, 14 papers with experiments were selected as the primary studies. The published works in this area were initiated in 2018. Control tools are applied commonly in experiments. Various advantages and shortages of chatbot usability experiments were revealed, for example, most of the experiments do not provide raw data, and only one of the identified works replicated the experiment. Conclusions: An increased interest in usability experimentation of chatbots is observed in recent years. The chatbot usability experiment should be more replicable to improve the reliability of experimental results.**

Keywords—Usability, Chatbots, Family of Experiments

I. INTRODUCTION

A chatbot, also known as a chatterbot, is a domain-specific text-based software that supports human users with specific services [1]. The remarkable advancement of natural language processing and machine learning is causing a seismic shift, in that sense, this created unlimited possibilities, productive and useful experiences through chatbots who can access and interact with digital services in many different applications [2][3]. Compared with other communication channels (e.g., e-mail), not all users are willing to fully trust the chatbot due to understanding ability and response quality, chatbot design is still far from reading users' minds, in this context, it is necessary for better integration between usability evaluation and the chatbot [4]. Usability evaluation refers to how well users can learn and use software to meet their requirements and refers to how satisfied users are during the process [5]. In software engineering (SE), usability has been recognized as a crucial quality characteristic for success in the competitive commercial world [7]. The choice of evaluation methodology must be applied appropriately for the current research question or issue [5]. Apparently, usability evaluation of chatbots is not a mature field so far [4]. In general, usability evaluation of chatbots learns and borrows experience from experimentation in software engineering (ESE). We noticed that the families of experiments are being run in increasing numbers in ESE [8]. It is the unanimous opinion of the scientific community that the veracity of the base experiment results can only be established by replication and contrast of results [9]. A family of experiments is a set of experimental replications with the same goal [8]. The families of experiments allow to obtain a greater statistical power due to the greater number of involved subjects [10], increase the internal validity of joint conclusions and the reliability of the findings. Due to the strengths of families of experiments, we pay special attention to the adoption of families of experiments in chatbot usability evaluation. To explore the chatbot usability experimentation, we did a preliminary investigation, and we did not find any previous study or literature review that could bring us a consolidated view. As described by Ren et al. [4], we noticed that chatbots and their relevant usability evaluation had become prevalent themes and the number of publications started to grow from the year of 2015; however, they did not pay attention to the usability experiment of chatbots. For this purpose, we conducted a systematic mapping study (SMS) on top of a baseline study [4] with the aim of (i) explore the state-of-the-art on chatbots usability experimentation, (ii) identify the metrics used in experiments to measure chatbot usability in SE. The nature of our contribution is providing a map of what has been published since we have included all reported references in the literature of our SMS on chatbot usability experimentation. With this information, researchers interested in conducting experiments and/or replications related to the usability of chatbots will obtain a baseline of aspects that they should consider.

Paper organization. In Sec. 2, we outline the research method of the SMS. In Sec. 3, we provide the answer to each of the research questions. In Sec. 4, we discuss the results and threats to validity. Finally, we outline the conclusions of our study in Sec. 5.

II. RESEARCH METHOD

The secondary study reported in this paper has been developed following the guidelines established by Kitchenham and Charters [11].

Objectives and Research Questions. The main objective of this study was to map the usability experiments of chatbot in aspects of publication status, and measured metrics in experiments. This gave rise to our research questions: (**RQ1**) What is the state of the art of chatbots usability experimentation? (**RQ2**) How to evaluate the usability of chatbots in experiments?

Search String Selection. We first piloted various synonymic search strings. The rationale behind the selection of our final search string is that it returns the most records, and the results are more balanced between the different databases. Our final search string was: *(usability OR "usability techniques" OR "usability practice" OR "user interaction" OR "user experience") AND (chatbots OR "chatbots development" OR "conversational agents" OR chatterbot OR "artificial conversational entity" OR "mobile chatbots")*.

Databases and Search Protocol. The IEEE Xplore, ACM Digital Library, SpringerLink, Scopus and ScienceDirect academic databases (DBs) were used in the SMS process. The selection criteria used to retrieve the fundamental studies are

* Corresponding Author.

DOI reference number: 10.18293/SEKE2021-053

summarized below. We dismissed an article whenever at least one of the exclusion criteria was met. *Inclusion criteria*: ((the abstract or title mentions an issue regarding the chatbots and usability) OR (the abstract mentions an issue related to usability engineering or HCI techniques) OR (the abstract mentions an issue related to the user experience)) AND (the paper describes the experiment of chatbot usability). *Exclusion criteria*: (the paper does not present any evaluation or experiment related to chatbot usability) OR (the paper does not present any issue related to the chatbots and usability) OR (the paper does not present any issue related to the chatbots and user interaction) OR (the paper does not present any issue related to the chatbots and user experience) OR (the paper is written in a language other than English).

Search Process. We reviewed works about the experiments of chatbot usability, which were published from 2014 to June 2020. Once we identified the search strings and defined search fields, we started our search process. A total of 363 *Retrieved Papers* were found in the different DBs. Then the duplicate papers were removed from the retrieved papers, 323 papers were filtered to the group of *Non-Duplicate Retrieved Papers*. A peer review was carried out on these 323 papers applying the inclusion and exclusion criteria to the title and abstract. Discrepancies were resolved through a discussion. As a result, we obtained 86 *Candidate Papers*. To determine if candidate papers have relevance regarding the usability of chatbots and the execution of the chatbot usability experiment, we reviewed each candidate paper again, applying the inclusion and exclusion criteria. However, this time we especially reviewed the full text. The results were cross-checked by two HCI experts. Finally, 14 papers formed the *Experiment Paper*s used in this study. The results of selection were assessed by two HCI experts, each disagreement has been discussed and resolved during the meeting. The remained 14 experiment papers for the analysis and extraction of the results are shown in Appendix A.

III. RESULTS

RQ1: What is the state of the art of chatbots usability experimentation? The raw data were poorly reported among 14 experiments as only one experiment provided access to their raw data in the paper. As shown in Fig. 1, a synthetic view of the identified primary studies, the results have been segmented into two areas. The left-side consists of two scatter (XY) plots (top and bottom) with bubbles at the junctions of the year-type of publication categories (left side - top) and usability feature-type of publication categories (left side - bottom). With regard to the types of publications, 50% of publications are conference papers, 28.6% are journal article, 21.4% are chapters book. The size of each bubble was determined by the number of experiment papers that had been classified into each category. The right-side of Fig. 1 presents the number of primary studies published per year. As can be seen from the upper right part of Fig. 1, the interest in chatbots usability experimentation is increasing and is very recent; initial works are from 2018. Considering that the search was carried out until June 2020, the number of identified works in our SMS for 2020 is high. Satisfaction is the most widely measured usability feature. Note that the number of papers in the lower part of Fig. 1 does not match the number of papers in the upper part. The reason is that the same paper can discuss several usability features. In aspect of the types of chatbots, most chatbots are deployed as the personal assistant [PS2][PS4][PS6][PS10][PS11][PS13], especially in the health care domain [PS5][PS7][PS14], some act as e-commerce tools [PS9][PS12], collaborative tool [PS8] and recommender [PS3].

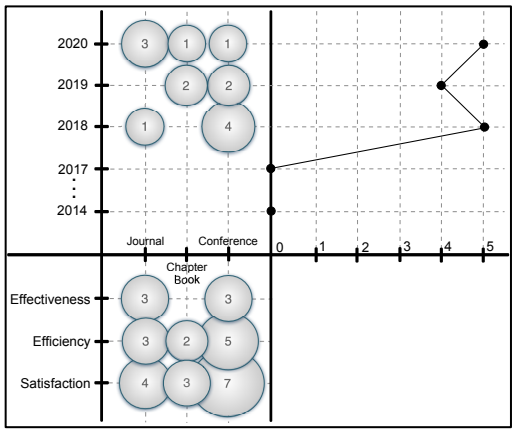

Figure 1. Mapping showing the primary study distribution.

RQ2: How to evaluate the usability of chatbots in experiments? Compared with the work of [4], we notice that more varieties of questionnaires were opted for to investigate the usability of the chatbots in recent years (see Table I), except SUS and ad-hoc. In [PS2], the AttrakDiff2 questionnaire was used to measure how attractive a product is based on its hedonic and pragmatic qualities. The Likert scale was the most used metric among those questionnaires from the past until now [PS3][PS6][PS9][PS12]. Over the usability evaluation process, pre-test and post-test questionnaires were combined for use in [PS5][PS10] to deepen the result of evaluation.

TABLE I. USABILITY TECHNIQUES

Usability Techniques	Experiments
Questionnaire	[PS1][PS2][PS3][PS4][PS6][PS7][PS8][PS9] [PS11][PS12][PS13][PS14]
Interview	[PS1][PS7][PS10][PS13]
Think-aloud	[PS5][PS13]
Direct observation	[PS5]

The Replication of Experiments. Upon the usability experiments we reviewed, there is only one study presented by Huff-Jr et al. [PS6] conducted a replication of an experiment with consistent experimental design but different participant background. They used a within-subjects mixed-method design, and they analyzed data by analyzing qualitative contents and a multilevel linear model. The total sample size of replications is 35, although the authors do not report the corresponding sample size of each replication. To the best of our knowledge, a family of experiments should include at least three experiments [8], while a single experiment had been replicated in their work—that is, it forms a set of two experiments since two experiments are able to aggregate the data to evaluate the effect of chatbots—we classified them as a family of experiments.

Sample Sizes. Although the sample size varies in different usage and development phases, as the recently published experiments have, the sample sizes of usability experiments for chatbots are relatively small. Of the 14 experiments, 50% of experiments contained less than 30 subjects, 28.6% contained between 30 and 50 subjects, and 14.3% contained between 100 and 500 subjects. One experiment did not detail the sample size [PS11].

Types of Subjects. 35.7% of the experiments include students, while most of the researchers did not limit academic background and grade. There were 29% that included experienced users or experts and company employees. Three experiments included farmers, children, and residents, respectively. Two experiments did not detail the types of subjects. Only one experiment used compared group, graduates and undergraduates [PS8].

Experimental Design and Procedure. 71.4% of experiments were defined as a within-subject design. Since the sample sizes of identified experiments are relatively small, the within-subject design has better statistical power by doubling data points. In SE, experimental design plays a role in controlling for extraneous variables: mature experiments are run with pre-established protocols defining the experimental settings and the set of procedures that must be strictly adhered to during the execution and analysis of the experiments. By contrast, many usability experiments of chatbots are formed without any a priori plan or experimental design definition. Furthermore, the prior experience and technical knowledge have an impact on the global usability of Conversational Agents [PS13], while the pre-user experience or knowledge related to chatbot seems didn't be measured during some experiments [PS1][PS11].

Statistical Techniques. Statistical techniques are categorized from two perspectives: statistical descriptions and statistical inference. The statistical descriptions (Table II) are representation methods that integrate multiple datasets in a visual way to give context to the data and to improve reader understanding. There is an experiment that has not yet been executed [PS11]. Among 13 experimental results of chatbots usability, box plot and descriptive statistics were the most used presentation formats. Statistical inference was used to analyze 11 experiment results. 7 experiments used parametric tests [PS2][PS5][PS7][PS9][PS12][PS13][PS14], and 4 experiments used non-parametric tests [PS1][PS3][PS5][PS9]. The majority of the authors did not explain the motivation behind adopting the technique or indicate the challenges or advantages of adopting the technique.

TABLE II. STATISTICAL DESCRIPTIVE REPRESENTATION

Statistical Descriptive Representation	N	Experiments
Box plot	6	[PS1][PS2][PS6][PS8][PS12][PS14]
Descript. statistics table	4	[PS2][PS3][PS4][PS14]
Histogram	3	[PS3][PS5][PS10]
Line chart	2	[PS2][PS13]
Scatter plot	2	[PS2][PS7]
Textual description	2	[PS9][PS13]

IV. DISCUSSION AND VALIDITY THREATS

Although our goal is to present an analysis for chatbot usability experimentation, we noticed that the interfaces of most current chatbots take a form of an NL dialog: the development of chatbots has become standardized because there are many build platforms for different goals and usages that have been widely used [PS1][PS6][PS10]. Of the initial 363 papers selected in well-known electronic research databases, 14 studies were selected following a rigorous process, from selecting studies to solve disagreements found during the selection process. The comparison of two or more treatments and randomization of subjects are our key points to identify if the study described an experiment [12] when we reviewed each paper. The usability experiment of chatbot correlates to chatbot development; however, there is only one experiment related to a usability experiment of chatbot in an advanced or modified version [PS12]. To obtain reliable experimental results, all aspects of treatment (except for the manipulation of factors) should remain similar across all groups, as irrelevant variables pose a threat to validity. We noticed that many studies did not clearly state extraneous variables control in their experiment designs. For example, they did not discuss the possible learning effects between different sessions [PS6][PS10]. We observed that most chatbots were experimented based on some specificities—including the relatively small sample size, the subjects with a specific background, the tasks being preset, and whether it was the users' first encounter with a chatbot—as the expansion of experimental results to the industrial field is fairly limited. Besides, there is a work that did not published the experimental results as of our search date [PS11].

The first threat to the validity of this work is the bias in the paper selection process. Although the selection criteria and results have been double-checked and accepted by other authors, the publications were evaluated and classified based on the judgment and experience of the authors, and other researchers may have evaluated the publications differently. The second point is related to the type of studies included in this work. We expanded the search scope by using search strings that identify a wider range of publications. On the one hand, this SMS was developed using five databases as they were considered the most complete and most used database in SE. On the other hand, this search includes only papers written in English. Nonetheless, relevant papers produced by additional databases or resources or written in other languages could have overlooked.

V. CONCLUSION AND FUTURE WORK

RQ1: What is the state of the art of chatbots usability experimentation? From our SMS perspective, chatbots usability experiments are being run in increasing numbers (see Fig. 1). With regard to publication venue, half of the reviewed papers in our SMS are published through conferences. We notice that control tools are applied commonly in experiments, most studies used real-life products as control tools [PS1][PS2] [PS5][PS8]. To determine whether the chatbot was able to provide a similar experience to the user, some developed diverse version of chatbots with different functions or expression [PS3][PS9][PS10].

We also observed that many experiments did not define the research question or hypothesis follow ESE methods [12], or the proposed research questions are related to usability but are not limited to usability. In general, most studies investigate not only usability factors but also the quality of the interaction or chatbot performance [PS3][PS7][PS8][PS10], in order to understand the chatbot usability comprehensively and also some studies investigate the relationships between the usability and other factors (e.g., acceptability) [PS5][PS10][PS14]. The majority of the experiments did not provide access to raw data. This situation prevents rigorous peer-review and does not allow third-party researchers to reanalysis using aggregation methods that may be more appropriate than the original one [8].

RQ2: How to evaluate the usability of chatbots in experiments? We notice: (i) the questionnaire is the most used usability technique; (ii) the family of experiments was barely

used in this field so far since only one experiment contained replications was found; (iii) the within-subject design is the most popular design on chatbots usability experimentation; (iv) 50% of the experiments included a small sample size (less than 30 subjects) and the most subjects are students; (v) the number of tasks is relatively small, as most of the experiments applied less than six tasks; and (vi) parametric tests were the most used inference to analyze the experimental result in experiments.

We suggest that the researchers: (i) provide access to full raw data to guarantee the replicability of the experiment and transparency of results; (ii) consider the family of experiments or conduct replications of the baseline experiment to consolidate the experimental result and to increase the statistical power; (iii) more third-party evaluations should be considered in chatbot usability evaluation, as they do not suffer from the bias introduced in the previous development process. Considering that the work is limited by search date, databases and search strings, this study could be replicated in a future study. Based on this research results, we plan to conduct a family of experiments to evaluate a chatbot's usability with an advanced version to fill the gap and explore the topic further.

ACKNOWLEDGMENT

This research was funded by the Spanish Ministry of Science, Innovation and Universities research grant PGC2018-097265-B-I00 and MASSIVE project (RTI2018-095255-B-I00). Also, it received support from the Madrid Region R&D programme (FORTE project-P2018/TCS-4314).

APPENDIX A: PRIMARY STUDIES

[PS1] S. Katayama, A. Mathur, M. Van den Broeck, T. Okoshi, J. Nakazawa and F. Kawsar, "Situation-aware emotion regulation of conversational agents with kinetic earables", in Proc. 8th Intern. Conf. on Affective Computing and Intelligent Interaction (ACII'19), Cambridge, UK, 2019, pp. 725-731.

[PS2] S. Lee, H. Ryu, B. Park, and M. H. Yun, "Using physiological recordings for studying user experience: Case of conversational agent-equipped TV", Intern. Journal of Human Computer Interaction, vol. 36, no. 9, pp. 815-827, Feb. 2020.

[PS3] F. Narducci, P. Basile, M. de Gemmis, P. Lops, and G. Semeraro, "An investigation on the user interaction modes of conversational recommender systems for the music domain", User Modeling and User-Adapted Interaction, vol. 30, pp. 251-284, Mar. 2020.

[PS4] J. Guo, D. Tao, and C. Yang, "The effects of continuous conversation and task complexity on usability of an AI-based conversational agent in smart home environments", in: S. Long, B. Dhillon (Eds.), Man–Machine–Environment System Engineering, MMESE'19, (pp. 695-703). Lecture Notes in Electrical Engineering, vol 576. Springer, Singapore, 2020.

[PS5] A. Ponathil, F. Ozkan, B. Welch, J. Bertrand, and K. C. Madathil, "Family health history collected by virtual conversational agents: An empirical study to investigate the efficacy of this approach", Journal of Genetic Counseling, pp. 1-12, Mar. 2020.

[PS6] E. W. Huff-Jr, N. A. Mack, R. Cummings, K. Womack, K. Gosha, and J. E. Gilbert, "Evaluating the usability of pervasive conversational user interfaces for virtual mentoring", in: P. Zaphiris, A. Ioannou (Eds.), Learning and Collaboration Technologies. Ubiquitous and Virtual Environments for Learning and Collab., HCII'19 (pp. 80-98). Lecture Notes in Computer Science, vol 11591, Springer, Cham, 2019.

[PS7] R. Håvik, J. D. Wake, E. Flobak, A. Lundervold, and F. Guribye, "A conversational interface for self-screening for ADHD in adults", in: S. Bodrunova et al. (Eds.), Internet Science, INSCI'19 (pp. 133-144). Lecture Notes in Computer Science, vol 11551. Springer, Cham, 2019.

[PS8] R. Ren, J. W. Castro, A. Santos, S. Pérez-Soler, S. T. Acuña, and J. de Lara, "Collaborative modelling: Chatbots or on-line tools? An experimental study", in Proc. Evaluation and Assessment in Software Engineering (EASE'20), Trondheim, Norway, 2020, pp. 260-269.

[PS9] E. Elsholz, J. Chamberlain, and U. Kruschwitz, "Exploring language style in chatbots to increase perceived product value and user engagement", in Proc. 2019 Conf. on Human Information Interaction and Retrieval (CHIIR'19), Glasgow, Scotland, UK, 2019, pp. 301-305.

[PS10] M. Jain, P. Kumar, I. Bhansali, Q. V. Liao, K. N. Truong, and S. N. Patel, "FarmChat: A conversational agent to answer farmer queries", in Proc. ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT'18), vol. 2, no. 4, 2018, pp. 170:1-170:22.

[PS11] Q. N. Nguyen, and A. Sidorova, "Understanding user interactions with a chatbot: A self-determination theory approach", in Proc. 24th Americas Conference on Information Systems: Digital Disruption (AMCIS'18), New Orleans, LA, USA, 2018.

[PS12] M. Jain, R. Kota, P. Kumar, and S. N. Patel, "Convey: Exploring the use of a context view for chatbots", in Proc. Conf. on Human Factors in Comp. Systems (CHI'18), Montreal, QC, Canada, 2018, pp. 468:1-468:6.

[PS13] M.-L. Chen, and H.-C. Wang, "How personal experience and technical knowledge affect using conversational agents", in Proc. 23rd Intern. Conf. on Intelligent User Interfaces Companion (IUI'18), Tokyo, Japan, 2018, pp. 53:1-53:2.

[PS14] C. Sinoo, S. van der Pal, O. A. B. Henkemans, A. Keizer, B. P. B. Bierman, R. Looije, and M. A. Neerincx, "Friendship with a robot: Children's perception of similarity between a robot's physical and virtual embodiment that supports diabetes self-management", Patient Education and Counseling, vol. 101, pp. 1248-1255, Jul. 2018.

REFERENCES

[1] J. Guichard, E. Ruane, R. Smith, D. Bean, and A. Ventresque, "Assessing the robustness of conversational agents using paraphrases," in Proc. IEEE Intern. Conf. on Artificial Intelligence Testing (AITest'19), Newark, CA, USA, 2019, pp. 55-62.

[2] M. Jain, R. Kota, P. Kumar, and S. N. Patel, "Convey: Exploring the use of a context view for chatbots," in Proc. Conf. on Human Factors in Comp. Systems (CHI'18), Montreal, Canada, 2018, pp. 468:1-468:6.

[3] Q. N. Nguyen, and A. Sidorova, "Understanding user interactions with a chatbot: A self-determination theory approach," in Proc. 24th Americas Conference on Information Systems: Digital Disruption (AMCIS'18), New Orleans, LA, USA, 2018.

[4] R. Ren, J. W. Castro, S. T. Acuña, and J. de Lara, "Usability of chatbots: A systematic mapping study," in Proc. 31st Intern. Conf. on Software Engineering & Knowledge Engineering (SEKE'19), Lisbon (Portugal), 2019, pp. 479-484.

[5] S. Greenberg, and B. Buxton, "Usability evaluation considered harmful (some of the time)," in Proc. SIGCHI Conf. on Human Fact. in Comp. Systems (CHI'08), Florence, Italy, 2008, pp. 111-120.

[6] A. Seffah, M. C. Desmarais, and E. Metzker, "HCI, Usability and software engineering integration: Present and future," in: A. Seffah, J. Gulliksen, M. C. Desmarais (Eds.), Human-Centered Soft. Eng. — Integration Usability in the Soft. Devel. Lifecycle (pp. 37-57). Human-Computer Interac. Series, vol. 8, Springer, Dordrecht, 2005.

[7] K. Curcio, R. Santana, S. Reinehr, and A. Malucelli, "Usability in agile software development: A tertiary study," Computer Standards & Interfaces, vol. 64, pp. 61-77, May. 2019.

[8] A. Santos, O. Gómez, and N. Juristo, "Analyzing families of experiments in SE: A systematic mapping study," IEEE Transactions on Soft. Eng., vol. 46, no. 5, pp. 566–583, may. 2020.

[9] N. Juristo, "Once is not enough: Why we need replication," in: T. Menzies, L. Williams, and T. Zimmermann (Eds.), Perspectives on Data Science for Soft. Engin. Morgan Kaufmann, 2016, pp. 299-302.

[10] E. Fernández, O. Dieste, P. Pesado, and R. García, "The importance of using empirical evidence in software engineering," in: G. Simani, and H. Padovani (Eds.), Computer Science & Technology Series. XVI Argentine Congress of Computer Science Selected Papers, Ed. Universidad de la Plata (EDULP), 2011, pp. 181-189.

[11] B. A. Kitchenham, and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele University and Durham University, Keele, UK, EBSE Technical Report version 2.3 (EBSE-2007-012007), 2007.

[12] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslen, Experimentation in Software Engineering, Berlin, Germany: Springer, 2012.