A family of experiments for evaluating the usability of a collaborative modelling chatbot

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*—Recent natural language processing developments have facilitated the adoption of chatbots in typically collaborative software engineering tasks. Families of experiments can overcome limitations in terms of the sample size of individual experiments. To experimentally evaluate the usability of a chatbot for collaborative modelling (i.e., SOCIO) and tackle some of the typical shortcomings of individual experiments, we conducted a family of three experiments to evaluate the usability of SOCIO against the Creately online collaborative tool. Results show that the participants were more satisfied with the chatbot than with the online collaborative tool and that they also created class diagrams faster using the chatbot. We conclude that chatbots may be helpful for creating class diagrams.**

Keywords—Chatbots, Family of Experiments, Usability, Modelling

I. INTRODUCTION

Modelling is a fundamental part of the software development process, and it is often a collaborative activity [1]. A plethora of cloud-based platforms have recently emerged for synchronous mechanisms (e.g., Lucidchart, Gliffy and Creately). The SOCIO chatbot, a collaborative modelling tool, was developed to provide an alternative method for building models or meta-models using Twitter or Telegram (nick @ModellingBot) [2]. Along with the SOCIO chatbot, users benefit from social network collaboration and ubiquity to perform the lightweight modelling task [2].

Experiments can assess the effectiveness of software engineering (SE) treatments (e.g., tools) and check whether or not the hypotheses about the effectiveness of such treatments hold. Unfortunately, isolated experimental results may be unreliable due to small sample sizes [3], while families of experiments increase the reliability of joint conclusions and internal validity and rule out the detrimental effects of publication bias on conclusions [4][5]. It is critical to assess chatbot usability because they are increasingly being used across many domains [6][7], and poor interactions would have an impact on user willingness to use the service [8]. To increase the reliability and generalizability of individual experimental results, we used a family of experiments to assess the usability of the SOCIO chatbot.

In our family of three experiments, we compared the usability of the chatbot SOCIO with Creately (https://creately.com/app). Creately is a real-time collaborative tool built on Adobe's Flex/Flash technologies. We chose Creately as the control tool since no previous studies had assessed Creately usability, even though it is the most used online collaborative modelling tool [9], and it has similar functionality to SOCIO. Along the way, we made several findings with respect to efficiency, effectiveness and satisfaction issues in response to our research question:

RQ: Compared to Creately, does the use of SOCIO positively affect user efficiency, effectiveness, and satisfaction with respect to class diagram construction in a family of experiments?

Our findings contribute: (1) empirical evidence that the SOCIO chatbot improves usability and (2) direct suggestions from users, as a starting point for understanding the impact of three human-computer interaction (HCI) usability characteristics (effectiveness, efficiency and satisfaction) that affect collaborative modelling tool and chatbot design.

Paper organization. In Sect. 2, we present the related work in usability experiments for chatbots. In Sect 3, we describe the design of the family of experiments, show the data analysis and results of this family of experiments. The paper finishes with the threats to validity section (Sect. 5) and discussion and conclusions (Sect. 6).

II. RELATED WORK

In [10], we reported a wider systematic mapping study (SMS) to identify the state of the art with respect to chatbot usability and applied HCI techniques in order to analyse how to evaluate chatbot usability. We concluded that chatbot usability is an incipient field of research, where the published studies are mainly surveys, usability tests, and rather informal experimental studies. Hence, it is necessary to perform more formal experiments to measure user experience and exploit these results to provide usability-aware design guidelines. We then updated the SMS, focusing on papers published from November 2018 to June 2020 applying the same procedure and search string as in [10]. In particular, we reviewed chatbot usability evaluation experiments to discover the recent trends and methodologies in the experimental software engineering field. Based on Ren et al.'s selection criteria [10], we also included papers describing controlled chatbot usability experiments and we excluded papers reporting only an evaluation or a quasi-experiment related to chatbot usability.

Finally, we retrieved ten primary studies ([11]-[20]) reporting experiments on the usability of chatbots which we used in this study. Only one study, designed as a within-subjects mixed-method experiment with different participant backgrounds, carried out replications of experiments [16]. Satisfaction continues to be the most popular usability characteristic, since it was evaluated more often. Task completion time and task completion are the efficiency and effectiveness characteristics attracting most interest, respectively. Within the primary studies, most chatbots are used as personal assistants [12][16][17][18][20].

* Corresponding Author.

DOI reference number: 10.18293/SEKE2021-043

Nevertheless, none of the chatbots were applied as modelling tools like SOCIO.

So far there have been three studies of the usability of SOCIO: the baseline experiment of this paper [11], and two separate evaluations [2][21]. All of these studies used questionnaires, and all participants had a SE background. Two small-scale evaluation experiments for SOCIO (with 19 and 8 participants) were reported in [2][21]. They measured SOCIO chatbot applicability for building an e-commerce class diagram in 15 minutes [2] and a consensus mechanism for choosing different modelling alternatives, where subjects had to choose the best of three options for two projects [21]. However, these studies focused on evaluating SOCIO separately using simple tasks. An evaluation experiment with a larger number of subjects (54 participants) comparing the SOCIO chatbot with the web-based application Creately was reported in [11]. Even though the subjects had identical backgrounds, session and task were confounded, highlighting the potentially detrimental effects of combining experimental results.

With the aim of moving beyond the limitations of the above studies, we build a family of experiments - which is defined as a group of at least three experiments with the same goal - by means of replication. Families of experiments allow surpassing the limitations in terms of sample size of individual experiments, and also, evaluating the effects of the treatments under different settings [22]. Families provide certain advantages for evaluating the effectiveness of SE treatments [4][5]: (i) because access to the raw data is granted in families, researchers can apply consistent pre-processing and analysis techniques to analyse the experiments, and, in turn, increase the reliability of joint conclusions; (ii) researchers conducting families may opt to reduce the amount of changes made across the experiments with the aim of increasing the internal validity of joint conclusions; and (iii) because families do not rely on already published results, joint conclusions are not affected by the detrimental effects of publication bias. Due to the advantages of families of experiments, we followed this approach to conduct our research.

III. Family Design

Since our family contains a SE baseline experiment and two replications, we designed the experiment according to the guidelines proposed by Santos et al. [22].

A. Objectives, Hypotheses and Variables

The objective of our family of experiments was to evaluate the usability, in terms of efficiency, effectiveness, and satisfaction, of the SOCIO chatbot through comparison in controlled experiments with the Creately web tool. The null hypotheses governing this research question is: *H.x.0 There is no significant difference in EFFICIENCY | EFFECTIVENESS | SATISFACTION with respect to class diagram construction using SOCIO or Creately.* This hypothesis is broken down into three specific null hypotheses, one for each usability characteristic (where x represents 1. Efficiency, 2. Effectiveness and 3. Satisfaction).

The main independent variable across all experiments is the modelling tool. The treatments are the SOCIO chatbot and the Creately web application. The response variable within the family is usability. Based on definitions of usability in ISO/IEC 25010:2011 [23], ISO 9241-11:2018 [24] and ISO/IEC/IEEE 29148:2018 [25], and Hornbæk's guide [26], efficiency, effectiveness and satisfaction are commonly measured attributes for evaluating product usability. In view of this, we measure usability as efficiency, effectiveness and satisfaction.

We measured efficiency in terms of *speed* and *fluency*. Speed corresponds to the time taken to complete the tasks. Fluency corresponds to the number of discussion messages exchanged between the teammates during task development via the Telegram group. We measured effectiveness as *completeness*, based on the perceived success of each class diagram compared with the ideal class diagram that we built to measure the solutions produced by all participants [11][26]. In particular, the speed, fluency, and completeness metrics refer to social complexity and sociability and are typically evaluated when measuring macro-level usability (tasks requiring hours of collaboration) [23][24][26]. To assess and quantify satisfaction, we modified the System Usability Scale (SUS) questionnaire [21][27] to suit our experiments. Ease-of-use and learnability are two measured sub characteristics included in SUS questions [26]. There are ten SUS questions –each question is scored on a five-point Likert scale– and four open-ended questions. Finally, we adopted Brooke's equation [27] to derive the numerical value of each participant's satisfaction score. The median of the scores given by all three members of each team –to each question– is selected as the team score.

B. Design of the Experiments

All three experiments in our family have an identical experimental design. The study employed a two-sequence and two-period within-subject crossover design (see Table I). We chose a crossover design to avoid the influence of the period on the treatment and assure that there was no learning effect between the two periods [28].

TABLE I. Experimental Design

Group	***Period 1 (Task 1)***	***Period 2 (Task 2)***
Group 1 (SC-CR)	SOCIO	Creately
Group 2 (CR-SC)	Creately	SOCIO

The participants were grouped into three-member teams, where each team was considered as a subject. We put the full participant name list in a random team generator (www.randomlists.com/team-generator) to generate teams. All teams were assigned to either of two groups (Group 1 or Group 2), where each group applied the treatments in a different order. Participants did not receive any training and signed an informed consent before the experiment. After a 10-minute tutorial on the tool that they were to use in each period, they were required to perform the task in 30 minutes. Group 1 implemented Task 1 using SOCIO in the first period followed by Task 2 with Creately in the second period (i.e., SC-CR sequence).

On the other hand, Group 2 implemented Task 1 with Creately first, followed by Task 2 with SOCIO (i.e., CR-SC sequence). Task 1 was to develop a class diagram representing a store, including the management of products and customers. Task 2 consisted of designing the class diagram of a school to support courses and students. At the end of each period, all participants filled in a modified and validated SUS questionnaire. We did not ask participants which tool they preferred until the end of the second period.

C. Subjects

Participants were recruited from two universities in two countries: (1) the *Universidad de las Fuerzas Armadas ESPE Extensión Latacunga* (ESPE-Latacunga) in Ecuador (UNIV-1), and (2) the *Escuela Politécnica Superior of the Universidad Autónoma de Madrid* (EPS-UAM) in Spain (UNIV-2), all participants are undergraduate students who were completing a degree in Computer Engineering. Each participant only participates once. A total of three experiments were run. 18 subjects (54 participants) of the baseline experiment (EXP1) are from UNIV-1. 10 subjects (30 participants) of the second experiment (EXP2) are from UNIV-2. The third experiment (EXP3) contains 11 subjects from UNIV-2 and 5 subjects from UNIV-1 (48 participants in total). The subjects were selected using convenience sampling: participants were students of academic staff teaching SE-related courses, and all participants volunteered to participate. All participants were required to complete a pre-test questionnaire that assessed demographic data and related experience and knowledge.

As Fig. 1 shows, average subject experience appears to be slightly heterogeneous, but the gaps between each experiment appear to be small (i.e., never greater than 1). Although 37% of participants have no experience in using Telegram, they are regular social media users. This ensures that they can complete the task since no complicated operations are required. However, the inclusion of subjects with no previous experience with chatbots does pose a threat to the validity of the results. Despite the fact that none of the participants were native English speakers, they all claimed to have at least an intermediate level of English. As there are no significant differences between the three experiments in terms of age, gender, knowledge background, social media usage habits, smartphone or tablet ownership, we consider that the participants across the countries are comparable.

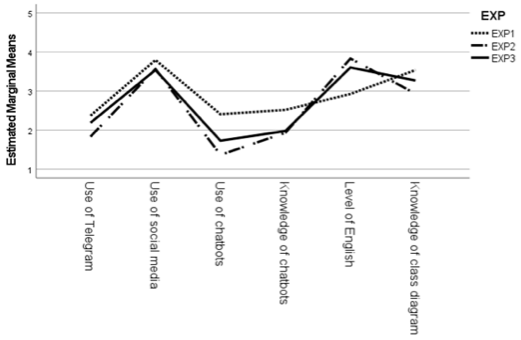

Fig. 1. Profile plot for subject experience

IV. Results and Data Aggregation

A. Analysis Approach

In response to the research question, we follow Santos et al.'s guidelines [22] to analyse the family of experiments. For each metric, we provide: (i) a profile plot showing the mean effect of the treatments across the experiments (ii) a violin-plot and the descriptive statistics divided by treatment and by experiment; and (iii) the joint results of all the experiments together applying a one-stage individual participant data (IPD) meta-analysis, reporting the contrast between treatments as an extra parameter in the linear mixed model (LMM) model to account for the difference between results across experiments [22][29]. The profile-plots give a bird's eye view of the data at family level and check for the existence of patterns across the results [22]. The descriptive statistics and violin-plots ease the understanding of the data in each experiment. We followed an IPD meta-analysis approach rather than a meta-analysis of effect sizes, because we had access to the raw data of the experiments [29].

As all the experiments have an identical (i.e., a cross-over) design, we analyse them following Vegas et al.'s advice [28]. In particular, we analyse the experiments using linear mixed models (LMMs) [28]. We used LMMs rather than their non-parametric counterparts because: (i) commonly used non-parametric models are not useful for studying the effect of multiple factors at the same time (e.g., period, treatment, and sequence on the outcomes); (ii) the overall sample size (i.e., 44 teams, each with two data-points —one per session, a total of 88 data points) may suffice to make the central limit theorem hold [30], and thus, interpret the results despite data non-normality.

In particular, we fit a three-factor LMM [31] for each metric: period (i.e., 1 or 2), treatment (i.e., SOCIO, or Creately), and sequence (i.e., SOCIO-Creately or Creately-SOCIO). We add an extra parameter to the LMM to account for the difference between results across the experiments (i.e., Experiment), which is a common feature of stratified individual participant data (IPD) models [22]. We interpret the statistical significance of the results with the corresponding ANOVA table of LMMs.

B. Response Variables

1) Efficiency

As Fig. 2 and 3 and the descriptive statistics (Table II) show, the aggregate time appears to be less for SOCIO than for Creately in two out of three of the experiments. The difference in performance between the treatments is statistically significant in the ANOVA table (Table IV). According to the pairwise contrast between the treatments in Table V, **the participants took an average of 1.14 minutes longer with Creately than with SOCIO.**

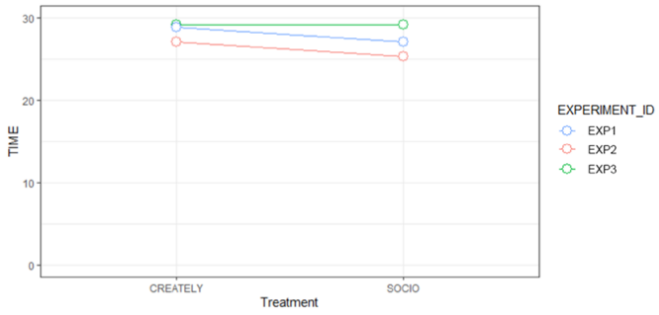

Fig. 2. Profile plot for time spent on tasks

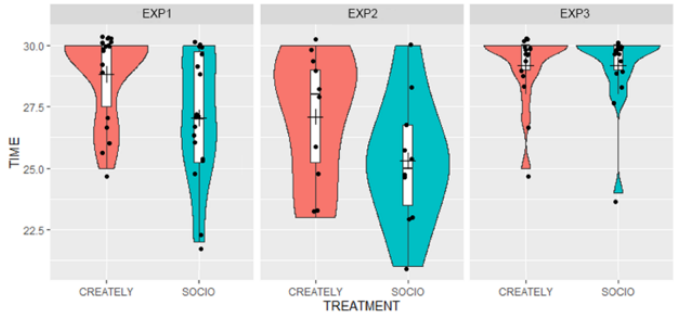

Fig. 3. Violin-plot for time spent on tasks

TABLE II. DESCRIPTIVE STATISTICS FOR EFFICIENCY. LEGEND: TR=TREATMENT; CR=CREATELY; SC=SOCIO;FLUEN=FLUENCY

Metric	Exp	TR	Team	Mean	Std. Dev.	Median
SPEED	EXP1	CR	18.00	28.83	1.76	30
	EXP1	SC	18.00	27.06	2.62	27
	EXP2	CR	10.00	27.10	2.69	28
	EXP2	SC	10.00	25.30	2.63	25
	EXP3	CR	16.00	29.19	1.42	30
	EXP3	SC	16.00	29.19	1.56	30
FLUEN	EXP1	CR	18.00	19.56	16.30	13.50
	EXP1	SC	18.00	9.61	11.51	5.00
	EXP2	CR	10.00	75.40	40.84	68.00
	EXP2	SC	10.00	57.00	19.98	51.00
	EXP3	CR	16.00	63.00	45.85	70.00
	EXP3	SC	16.00	65.81	46.00	66.00

As we can see in the plots and the descriptive statistics (Figs. 4 and 5 and Table II), the participants tend to send more messages with Creately than with SOCIO. Besides, as Table IV highlights, the difference in the number of messages is statistically significant. In particular, **the participants send up to 7.23 more messages with Creately than with SOCIO**, as shown in Table V.

Considering different textual communication styles, we treat a complete single sentence or an emoji as a message. Although message exchange was encouraged, we considered a low number of messages is an indicator that fewer communication efforts were required, since users were immediately able to observe the changes in the class diagram.

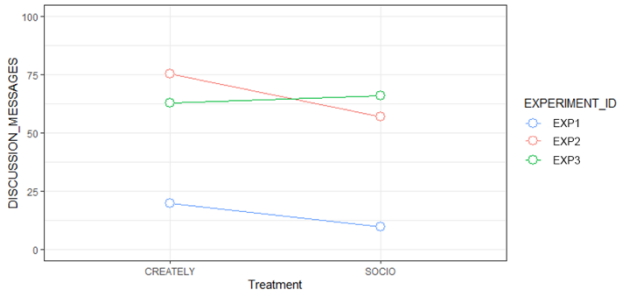

Fig. 4. Profile plot for discussion messages

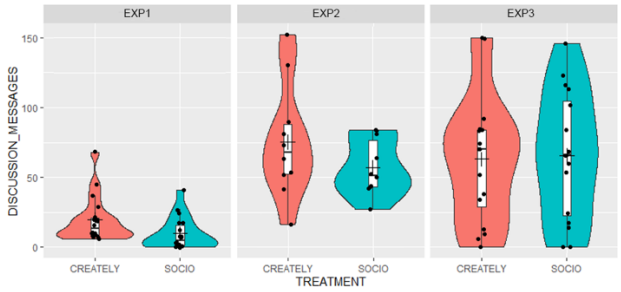

Fig. 5. Violin-plot for discussion messages

2) *Effectiveness*

As we can see in Figs. 6 and 7 and Table III, completeness appears to be similar for both tools. Besides, as shown in Table IV and Table V, the observed difference in completeness (-0.003) was negligible and not statistically significant. In sum, **Creately and SOCIO appear to perform similarly in terms of completeness.**

3) *Satisfaction*

As Figs. 8 and 9 and Table III show, the participants appear to be more satisfied with SOCIO than with Creately in EXP1 and EXP2. The opposite applies to EXP3, albeit to a lesser extent. As Table IV and Table V show, the difference in satisfaction scores appears to be significant at the 0.1 level. In other words, **participants appear to have higher satisfaction scores with SOCIO**.

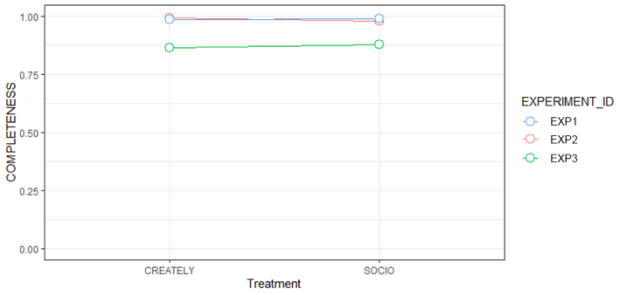

Fig. 6. Profile-plot for completeness

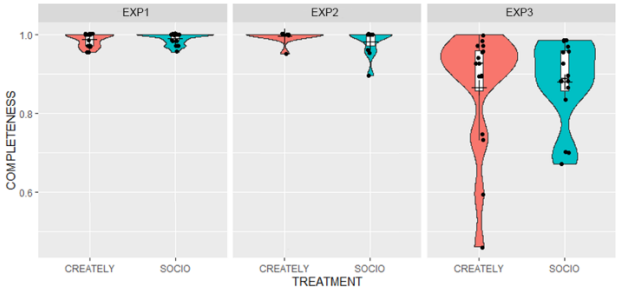

Fig. 7. Violin-plot for completeness

TABLE III. DESCRIPTIVE STATISTICS FOR COMPLETENESS AND SATISFACTION. LEGEND: COMP=COMPLETENESS; SATIS= SATISFACTION

Metric	Exp	TR	Team	Mean	Std. Dev.	Median
COMP	EXP1	CR	18.00	0.99	0.02	1.00
	EXP1	SC	18.00	0.99	0.01	1.00
	EXP2	CR	10.00	0.99	0.02	1.00
	EXP2	SC	10.00	0.98	0.04	1.00
	EXP3	CR	16.00	0.86	0.15	0.92
	EXP3	SC	16.00	0.88	0.11	0.89
SATIS	EXP1	CR	18.00	64.72	11.50	66.25
	EXP1	SC	18.00	71.32	11.18	70.00
	EXP2	CR	10.00	43.50	21.86	43.75
	EXP2	SC	10.00	66.00	16.12	72.50
	EXP3	CR	16.00	60.16	17.78	61.25
	EXP3	SC	16.00	55.62	15.51	55.00

TABLE IV. ANOVA TABLE OF TREATMENT

Metric	numDF	denDF	F-value	p-value
SPEED	1	42	6.187	**0.0169**
FLUEN	1	42	4.1183	**0.0488**
COMP	1	42	0.068	0.7955
SATIS	1	42	3.4203	0.0714

TABLE V. CONTRAST BETWEEN TREATMENTS

Metric	Estimate	SE	df	t-ratio	p-value
SPEED	1.14	0.457	42	2.487	**0.0169**
FLUEN	7.23	3.56	42	2.029	**0.0488**
COMP	**-0.0033**	0.0126	42	-0.261	0.7955
SATIS	-6.16	3.33	42	-1.849	0.0714

V. THREATS TO VALIDITY

Replications are subject to *conclusion validity*. To mitigate any possible influence, we resorted to parametric statistical tests (i.e., LMM [31]) to analyse the data and ensured result robustness by meta-analysing the data with the one-stage IPD model and an extra factor to account for the difference in

results [22][32]. We also evaluated the quality of the constructed class diagrams with respect to different aspects so as to give a better understanding for the time metric. In order to ensure the transparency of the results, we provide the original data, statistical analysis carried out and collaboration examples with chatbot SOCIO and Creately in the supplementary materials at https://bit.ly/34v7OTs.

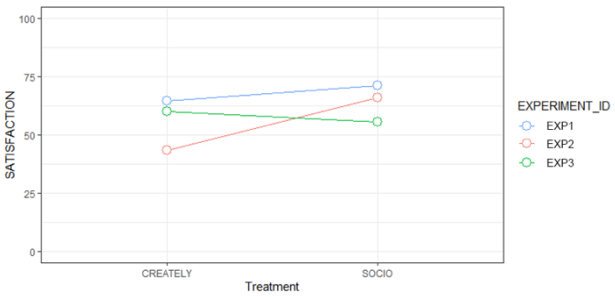

Fig. 8. Profile-plot for satisfaction

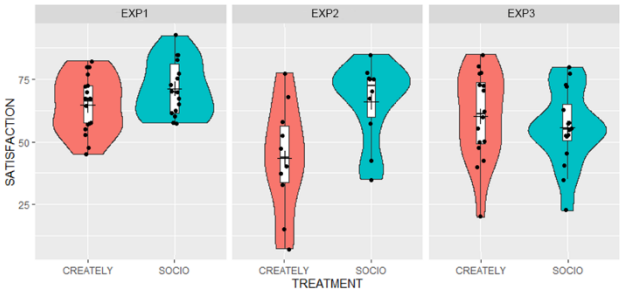

Fig. 9. Violin-plot for satisfaction

Unacknowledged variables confounded with the investigated variable may pose threats to *internal validity*. Since subject background (e.g., different universities) is another potential independent variable, this threat may compromise the validity of the results. In EXP3, we tried to mitigate this threat by conducting the experiment in the same universities as EXP1 and EXP2.

In terms of *construct validity*, we acknowledge that self-assessment questions may not properly reflect the knowledge background of the participants because they may not be able to honestly assess either their knowledge level or their characteristics—even if they used chatbot/social network as frequently as each other. This fact add bias on the response variable of satisfaction.

A probable *external threat* is the generalization of results. As usual in SE experiments [33], we had to rely on toy tasks to evaluate and compare the performance of two tools. Having said this, the subjects that make up this family of experiments are computer science students with sufficient knowledge of the field. In view of this, our findings are limited to academia and are not generalizable to industry.

VI. Discussion and Conclusions

To the best of our knowledge, there is no other chatbot offering a similar service to SOCIO. Although SOCIO chatbot usability was evaluated in two other small-scale evaluation results [2][21] previously, the number of subjects was smaller than provided by this family of experiments and it was not compared with other tools. Our family of experiments is the first and only research to evaluate the usability of the SOCIO chatbot comprehensively with regard to effectiveness, efficiency and satisfaction. Particularly, this family consolidates the previous result of the baseline experiment [11] thanks to a bigger sample size and more powerful statistical results. The information aggregated at family level is much more accurate than for individual experiments, which, in many cases, are unable to observe the existing differences. For instance, the treatment is not statistically significant for all variables in EXP3, which is not the case in the family of experiments (see supplementary material).

We followed a mixed method to provide joint results and identify variables impacting results. With our family of experiments, we observed that subjects take longer and send a larger number of messages to build class diagrams with Creately than with SOCIO. In other words, **SOCIO outperforms Creately in terms of efficiency**. Regarding effectiveness, **results are similar for both tools**. For satisfaction, we can conclude that **participants were more satisfied with SOCIO than with Creately**.

In addition, with the aim of identifying concrete opinions related to the satisfaction from subjects, we extended SUS questionnaire (see supplementary material) with four open-ended questions (concerning positive and negative aspects of the two tools, suggestions and user preferences) in order to gather definite satisfaction-related opinions from subjects. By analysing responses to open-ended questions, we find some insight as follows. Many participants remarked that they found both tools to be satisfactory in terms of responsiveness, ease of use, and collaboration capabilities. Creately was praised for its friendly interface. SOCIO was more fun to use. Quite a few participants complained about the SOCIO chatbot help web page, whereas the biggest problems with Creately were related to real-time collaboration, which produced some errors when loading on some of the user's computers.

This research contributes to the empirical analyses of the evaluation of chatbot usability, in particular, the chatbot SOCIO. There is the existence of statistically significant differences with medium effect size. Additionally, our experiments provide further information for developers regarding the usability evaluation of SOCIO chatbot and Creately. We conclude that chatbots may aid in the creation of class diagrams. In particular, their speed may be valuable, especially in view of the satisfaction shown by the participants with their use.

Future studies will focus on investigating this updated versions of the SOCIO chatbot. Accordingly, it is possible to clarify the required evidence-based SOCIO chatbot improvements. Currently, work is underway to develop four different updated versions of the SOCIO chatbot: (1) Provide different help when the SOCIO chatbot does not understand the user well according to a different situation. (2) Add functionalities requested by users: users will be able to delete any elements that they like by clicking the buttons underneath, and users will be able to choose how many steps to cancel or redo at a time instead of deleting or redoing one by one. (3) Provide option to select the appearance of the class diagrams. (4) Update and supplement the help page for all three versions.

ACKOWLEDGEMENTS

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) and also received support from the Madrid Region R&D programme (FORTE project, P2018/TCS-4314).

REFERENCES

[1] M. Franzago, D. D. Ruscio, I. Malavolta, and H. Muccini, "Collaborative model-driven software engineering: A classification framework and a research map", IEEE Trans. Softw. Eng., vol. 44, no. 12, pp. 1146-1175, 2018.

[2] S. Pérez-Soler, E. Guerra, J. De Lara, and F. Jurado, "The rise of the (modelling) bots: Towards assisted modelling via social networks", Proc. 32nd IEEE/ACM Int. Conf. Autom. Softw. Eng. (ASE'17). Urbana, IL, USA, pp. 723-728, 2017.

[3] T. Dybå, V. B. Kampenes, and D. I. K. Sjøberg, "A systematic review of statistical power in software engineering experiments", Infor. and Softw. Techn., vol. 48, no. 8, pp. 745-755, 2006.

[4] T. P. A. Debray, K. G. M. Moons, G. van Valkenhoef, O. Efthimiou, N. Hummel, R. H. H. Groenwold, J. B. Reitsma, and GetReal Methods Review Group, "Get real in individual participant data (IPD) meta-analysis: A review of the methodology", Res. Synth. Methods, vo. 6, no. 4, pp. 293-309, 2015.

[5] G. H. Lyman, and N. M. Kuderer, "The strengths and limitations of meta-analyses based on aggregate data", BMC Medical Res. Method., vol. 5, no. 1, pp. 1-7, 2005.

[6] G. Daniel, J. Cabot, L. Deruelle and M. Derras, "Xatkit: A Multimodal Low-Code Chatbot Development Framework, " IEEE Access, vol. 8, pp. 15332-15346, 2020.

[7] L. Erlenhov, F. Gomes de Oliveira Neto, R. Scandariato and P. Leitner, "Current and Future Bots in Software Development, " Proc. 2019 IEEE/ACM 1st Intern. Workshop on Bots in Software Engineering (BotSE'19). Montreal, QC, Canada, pp. 7-11, 2019.

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

[9] F. Lanubile, C. Ebert, R. Prikladnicki and A. Vizcaíno, "Collaboration tools for global software engineering", IEEE Software, vol. 27, no. 2, pp. 52-55, 2010.

[10] R. Ren, J. W. Castro, S. T. Acuña, and J. de Lara, "Evaluation techniques for chatbot usability: A systematic mapping study", Int. J. of Softw. Eng. and Knowl. Eng., vol. 29, no. 11n12, pp. 1673-1702, 2019.

[11] 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", Proc. Eval. Assessm. Softw. Eng. (EASE'20). Trondheim, Norway, pp. 1-9, 2020.

[12] S. Lee, H. Ryu, B. Park, and M. H. Yun, "Using physiological recordings for studying user experience: Case of conversational agent-equipped TV", Int. J. Hum. Comput. Interact., vol. 36, no. 9, pp. 815-827, 2020.

[13] 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, 2020.

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

[15] 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", Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII'19). Cambridge, UK, pp. 1-7, 2019.

[16] 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: Zaphiris P., Ioannou A. (eds). Learning and Collaboration Technologies. Ubiquitous and Virtual Environments for Learning and Collaboration (HCII'19), pp. 80-98. Lect. Notes Comp. Sci., vol 11591, Springer, Cham, 2019.

[17] 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: Long S., Dhillon B. (eds). Man–Machine–Environment System Engineering (MMESE'19), pp. 695-703. Lect. Notes Elect. Eng., vol 576. Springer, Singapore, 2020.

[18] R. Håvik, J. D. Wake, E. Flobak, A. Lundervold, and F. Guribye, "A conversational interface for self-screening for ADHD in adults", In: Bodrunova S. et al. (eds). Internet Science (INSCI'18), pp. 133-144. Lect. Notes Comp. Sci., vol 11551. Springer, Cham, 2019.

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

[20] M. Jain, P. Kumar, I. Bhansali, Q. V. Liao, K. N. Truong, and S. N. Patel, "FarmChat: A conversational agent to answer farmer queries", Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. (IMWUT'18), vol. 2, no. 4, pp. 1-24, 2018.

[21] S. Pérez-Soler, E. Guerra, and J. De Lara, "Collaborative modeling and group decision making using chatbots in social networks", IEEE Softw., vol. 35, no. 6, pp. 48-54, 2018.

[22] A. Santos, S. Vegas, M. Oivo, and N. Juristo, "A procedure and guidelines for analyzing groups of software engineering replications", IEEE Trans. Softw. Eng., pp. 1-22, 2019.

[23] ISO/IEC 25010, "ISO/IEC 25010:2011 - Systems and Software Engineering — Systems and Software Quality Requirements and Evaluation (SQuaRE) — System and Software Quality Models", ISO, 2011.

[24] ISO 9241-11. "Ergonomics of Human-System Interaction — Part 11: Usability: Definitions and Concepts." 2018.

[25] ISO/IEC/IEEE 29148:2018. "Systems and Software Engineering — Life Cycle Processes — Requirements Engineering." 2018.

[26] K. Hornbæk, "Current practice in measuring usability: Challenges to usability studies and research", Int. Journal of Human-Computer Studies, vol. 64, no. 2, pp.79-102. 2006.

[27] J. Brooke, "SUS-A quick and dirty usability scale". In: P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland (Eds.). Usability Evaluation in Industry, Chapter 21, pp. 189-194, 1996.

[28] S. Vegas, C. Apa, and N. Juristo, "Crossover Designs in Software Engineering Experiments: Benefits and Perils," IEEE Trans. Softw. Eng., vol. 42, no. 2, pp. 120–135, 2016.

[29] A. Whitehead, Meta-analysis of controlled clinical trials. John Wiley & Sons, 2002.

[30] J. de Winter, "Using the Student's t-test with extremely small sample sizes", Practical Assessment, Research, and Evaluation, vol. 18, pp. 1-13, 2013.

[31] B. T. West, K. B. Welch, and A. T. Galecki, Linear mixed models: A practical guide using statistical software. CRC Press, 2014.

[32] R. D. Riley, P. C. Lambert, and G. Abo-Zaid, "Meta-analysis of individual participant data: Rationale, conduct, and reporting", BMJ (Online first), pp. 1-7, 2010.

[33] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, Experimentation in software engineering. Springer Publishing Company, Inc., 2012.