IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 November 2022, accepted 5 December 2022, date of publication 12 December 2022,
date of current version 20 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3228772

RESEARCH ARTICLE

Using the SOCIO Chatbot for UML Modeling: A Second Family of Experiments on Usability in Academic Settings

RANCI REN[1], SARA PÉREZ-SOLER[1], JOHN W. CASTRO[2], OSCAR DIESTE[3], AND SILVIA T. ACUÑA[1]

[1]Departamento de Ingeniería Informática, Escuela Politécnica Superior, Universidad Autónoma de Madrid, 28049 Madrid, Spain
[2]Departamento de Ingeniería Informática y Ciencias de la Computación, Universidad de Atacama, Copiapó 1532297, Chile
[3]Escuela Técnica Superior de Ingenieros Informáticos, Universidad Politécnica de Madrid, 28660 Boadilla del Monte, Spain

Corresponding author: John W. Castro (john.castro@uda.cl)

This work was supported in part by the Spanish Ministry of Science, Innovation and Universities Research Grant under Grant PGC2018-097265-B-I00; in part by the Foundations for Augmented Trustworthy Low-Code Software Development (FINESSE) Project under Grant PID2021-122270OB-I00; in part by the Madrid Region Research and Development Program (FORmal Models and Technologies for Emerging Applications (FORTE) Project) under Grant P2018/TCS-4314; and in part by the Fundamentos Para la Ingeniería Automatizada de Chatbots (SATORI)-Universidad Autónoma de Madrid (UAM) Project under Grant TED2021-129381B-C21.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Review Committee of the Escuela de Doctorado de la Universidad Autónoma de Madrid.

ABSTRACT After improving the SOCIO chatbot prototype model, we wanted to know how/if its usability has changed. An evidence-based empirical evaluation of the usability of SOCIO V1 (updated version) requires an extensive verification of the experimental results. A family of experiments is a method of verification whereby we can check if the experimental results are reproducible. Through comparison with the updated control tool Creately, we aimed to gain a better understanding of the usability of the collaborative modeling chatbot and how it could be improved based on experimental evidence of changes in terms of efficiency, effectiveness, satisfaction, and quality. A total of 87 students from three countries were recruited. We conducted a family of three experiments to compare the usability of SOCIO V1 and updated Creately in academic settings. Students appeared to be more satisfied with SOCIO V1, and SOCIO V1 scored better on completeness. There were no significant differences between the two tools regarding efficiency and quality. This study provides evidence on how to employ a family of experiments to improve chatbot usability and enrich knowledge on chatbot usability experimentation.

INDEX TERMS Chatbot, usability, family of experiments.

I. INTRODUCTION

Collaborative modeling is an approach that deals with methods, processes, and tools for enhancing collaboration, communication, and coordination (3C) in teamwork [1]. Synchronization is used pervasively in software engineering (SE) collaborative modeling, providing for simple and efficient design changes in the collaboration environment. Many real-time collaboration modeling tools have been developed for target groups, e.g., Lucidchart, Creately, and Cacoo. Collaborative design is becoming increasingly relevant [1]. To this end, it is common for many, often geographically distributed, people to interact to build a product, such as a UML class diagram [1]. They interact verbally, either in writing, e.g., via chat, or audiovisually, e.g., via videoconference. When using videoconference, it may be preferable to use visual tools like Creately to build the class diagram to be constructed. When using chat, mechanisms such as chatbots can be integrated which participate in the conversation and simultaneously help create the class diagram [2].

Social networks like Telegram and Twitter have gained popularity and recognition [3]. With a view to integrating

The associate editor coordinating the review of this manuscript and approving it for publication was Orazio Gambino.

collaborative modeling tools into social networks, our colleague de Lara and his research group developed SOCIO chatbot (nick @ModellingBot), a collaborative modeling chatbot integrated into Telegram and Twitter [2]. SOCIO chatbot is an alternative collaborative modeling option to help stakeholders from different backgrounds perform lightweight tasks [2].

Usability deals with all sorts of activities related to software that is under development or has already been developed. Usability is defined in ISO/IEC50 25010:2011 [4] as a subset of quality in use, characterized explicitly by efficiency, effectiveness, and satisfaction. Experimentation is critical for evaluating usability in SE research [5]. Experimentation is a valuable tool for all software engineers involved in evaluation [6]. Back in 1998, Tichy reported his perspective on experimentation in software engineering (ESE) [7] as follows: ''Experimentation can help build a reliable base of knowledge and thus reduce uncertainty about which theories, methods, and tools are adequate.'' Nowadays, however, ESE is still a young and immature field where there is much debate on the appropriate research typology and evaluation criteria. Additionally, experiment replication types are not standardized at either the intra or interdisciplinary level [5].

A single experiment is unlikely to output reliable empirical results [5]. The outcomes of the experiment should be validated by replication. Lykken claimed in 1968 that ''the majority of theories should be evaluated through multiple corroborations and the majority of empirical generalizations through constructive replication'' [8]. Empirical evaluation has evolved considerably since its early beginnings, and the need for replication has been widely acknowledged in various scientific disciplines, including social science, business, and philosophy [5]. Replications of experiments have proven the need to be careful about accepting evidence that has not been subjected to strict corroborations [5]. To increase the robustness of the gathered experimental evidence, SE experiment replication is an indispensable part of ESE research [9]. The general purpose of replication is to check a previously observed finding. If the same results are reproduced in different replications, we can infer that these results are regularities existing in the portion of reality under study [10].

Quantitative analysis is widely used in experimental analysis and usability evaluation. Quantitative analysis interprets hard data collection [11]. However, qualitative analysis is a valuable paradigm for investigations where the data cannot be expressed numerically due to the complexity of the subjective characteristics and opinions involved. Thematic analysis is one of the most common forms of qualitative analysis. Thematic analysis is widely used across a range of epistemologies and research questions [12]. Thematic analysis has a number of advantages for evaluating the feedback from participants [12], [13], [14]: (i) researchers can apply a highly flexible approach that can be adapted to the needs of thematic analysis, (ii) it is an effective strategy for comparing and contrasting the perspectives of various research participants, revealing similarities as well as differences, and (iii) it is advantageous for summarizing significant characteristics of an extensive data set, as it helps to create a concise and ordered report.

This paper investigates a modified version of SOCIO with improved usability characteristics (SOCIO V1), based on the findings of a previous family of experiments. This second family of experiments tests whether or not SOCIO V1 consolidates the implemented usability characteristics. This article reports one of a number of families of experiments. In our case, the experimentation is performed in an academic setting, because the Unified Modeling Language (UML) modeling task is performed by senior computer engineering and mathematics students.

Our family of experiments aims to answer the following research question (RQ): How can chatbot usability be improved based on evidence from a family of experiments in academic settings?

In response to the research question, we designed an identical experiment for each experiment. We quantitatively analyzed the data using violin plots, descriptive statistics, and meta-analysis combined with linear mixed models (LMM) for each metric of each variable. Then we complemented the quantitative analysis by means of thematic analysis. The main contributions of the paper are: (1) the provision of evidence to enrich the body of knowledge to improve chatbot usability through the family of experiments, (2) demonstration of how chatbot usability can be improved by means of the family of experiments; and (3) provision of a summary and suggestions based on user feedback on how to improve software modeling.

The remainder of the paper is structured as follows. Section 2 describes the experiment background, indicating our improvement based on previous work. Section 3 reviews related work. Section 4 describes the design of our family of experiments. Section 5 reports the experimental result and quantitative and qualitative analysis. Section 6 describes the threats to validity. Section 7 discusses the experimental results. Section 8 outlines the conclusions and future work.

II. BACKGROUND

The first family of experiments comparing the usability of the basic version of SOCIO and Creately [15] was conducted in 2019. In this family of experiments, we adopted Creately (creately.com) as the control tool for comparison with the SOCIO chatbot, as Creately is one of the most commonly used modeling tools [16]. Creately is a web-based real-time collaboration tool for creating more than 50 types of diagrams, including UML diagrams.

SOCIO chatbot is a collaboration tool for creating class diagrams. By communicating with SOCIO in natural language (English), the team could create a class diagram in a group chat on Telegram or Twitter. From our first family of three experiments implementing a basic version of the SOCIO chatbot, we observed quantitative and qualitative feedback from this study, involving 132 participants. Quantitative data results provide: (1) proof of unsatisfactory

chatbot usability, and (2) insights on how to improve chatbots. The conclusion and future work of the previous publication [15] lists the usability improvements for SOCIO as follows:

- *Provide more help.* As some participants complained that they did not know where they went wrong when the chatbot did not understand their commands, they suggested the need for better help. Other participants said they were using the chatbot for the first time and therefore needed more help from the help page and during the interaction with the chatbot.
- *Delete any element that the user wants to delete, regardless of whether it was created by themselves or another team member.* In fact, some participants in the first experiments of the family suggested that the /undo command should be modified to enable a participant to undo his/her own action instead of the last action performed by the team.
- *Beautify the user interface.* Regarding the interface, some participants claimed that the look of the class diagram generated by the chatbot is old-fashioned and unchangeable.

After discussing with HCI experts and the entire SOCIO chatbot developer team, we prioritized the aspects on the list according to the evidence that we gathered from the results of the data aggregation. We decided to develop three updated versions with different advances. The changes that we made to versions 2 and 3 are outlined in Appendix A.

A. COMMON CHANGE

Change the guidance and help page. To provide more help to users, the following changes were made to all three updated versions:

1. Show the attribute types accepted by the chatbot as tips (int, double, float, date, string).
2. Update the guidance page in both English and Spanish (the native language of our subjects).
3. Provide examples on the help page to better explain the commands to help build the class diagram. For instance, we specify that point 3.5.6 would help relate entities and point 1.2 is helpful for directly making a command.

B. UPDATED VERSION 1 (SOCIO V1)

Alternative context-sensitive help. Apart from modifying the help page, we decided that, with a view to providing users with more help, the chatbot should have more than one optional response when it does not understand the user's command. Note that the improvement of context-sensitive help messages only affects SOCIO V1 for Task 1 and Task 2 and does not affect Creately.

1. When the user's command is properly formatted but is not understood by the chatbot, the SOCIO V1 chatbot sends an unchanged the project diagram. In the light of this, we modified the response to be an autoreply, alerting the user that the chatbot does not understand the command and providing some sample sentences that the chatbot can understand (see Figs. 1 and 2).
2. When a user's command is not in the correct format, we provide suggestions on how to organize the command correctly. For instance, we change the autoreply from "I don't understand this command" to "I don't understand this command. You can use all these commands: + command list" to remind users of the commands they can use (see Figs. 3 and 4).

In this article, we adopt the updated version 1 to conduct the second family of experiments with the aim of improving SOCIO chatbot's usability. Because Creately is a commercial product, it has undergone significant improvements. For example, the development team has upgraded the user interface, which no longer relies on Adobe Flash. We approached the Creately support team to request access to the version of Creately (used in the first family of experiments). However, they could not provide this version since Creately Classic was built on Adobe Flash, which is no longer supported by Adobe. Consequently, we used the updated Creately in the second family of experiments.

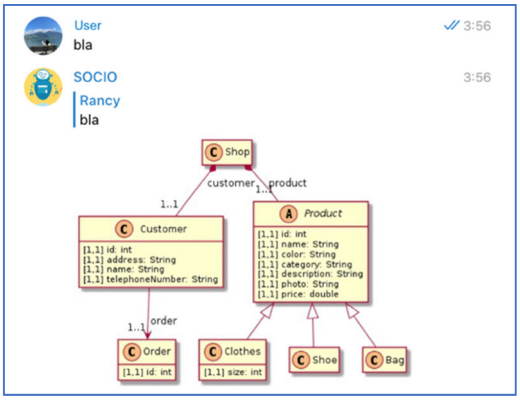


FIGURE 1. Before the first modification of Version 1.

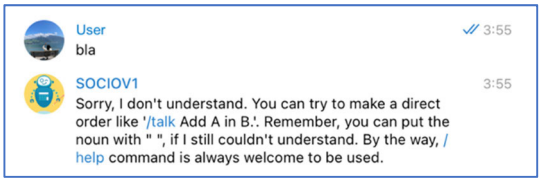


FIGURE 2. After the first modification of Version 1.

FIGURE 3. Before the second modification of Version 1.

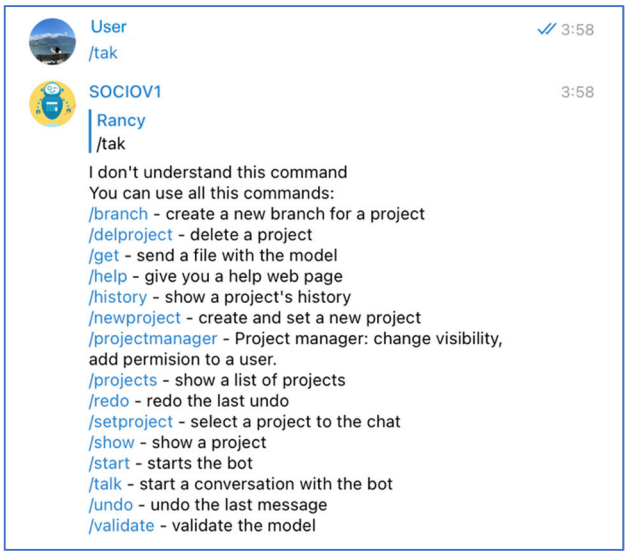


FIGURE 4. After the second modification of Version 1.

III. RELATED WORK

Ren et al. conducted a secondary study on chatbot usability experimentation [17]. They found that more and more chatbots had been evaluated with respect to various aspects, ranging from usability to practicability (or quality of outcome). We found that many chatbots had been evaluated through experimentation. However, most of the findings were based on observations from isolated experiments, and results have seldom been evaluated over again, irrespective of whether or not the chatbots were updated and improved. The second study reported that only one out of the 28 retrieved chatbot experiments [18] measured an improved version of the chatbot compared to the original version. The researchers designed a voice-activated chatbot that requires wake-up words. To get early feedback on the usability and the nature of any potential flaws, they conducted the first experiment with the first bot prototype that employed a simple heuristic to assess whether the user was addressing the bot. Following the enhancements to the early version of the bot, they conducted a second experiment with eight novices. These findings provide fair confidence that the second (improved) prototype bot is more useable. However, the researchers conducted experiments with different designs. In other words, to the best of our knowledge, most experiments on chatbot usability either have not been reproduced or have been reproduced according to the lesson they learned from the previous experiment.

It is pretty challenging to verify whether the results of independent experiments arise by chance, whether they are artificial, or whether the results conform to the regularities of the portion of reality under examination [5]. An effective validation method is to replicate the experiment to check that the results are reproducible [5]—this elucidated importance of replication in ESE.

A group of at least three replications could form a family of experiments to provide reliable validation [10]. Basili et al. [9] used the term family of experiments in 1999 to refer to a group of experiments pursuing the same goals whose results can be combined. Santos et al. further distinguishes the family of experiments through collections of experiments, either systematic literature reviews or replications of experiments [10]. Compared to individual experiments, Basili et al. [9] and Santos et al. [10] pointed out that a successful family of experiments has the advantage of increasing the validity and reliability of the outcomes of a single experiment.

However, we have not found any family or replication of experiments on chatbots following improvements. We regard this as being necessary in ESE in order to explore how to improve the usability of chatbots based on evidence. Therefore, we conducted a second family of experiments, reported in this article, with the improved version of the chatbot to explore how the usability of the chatbot was improved based on evidence.

IV. FAMILY DESIGN

This section describes the design of our family of experiments.

A. OBJECTIVES, HYPOTHESES AND VARIABLES

Based on findings from the previous study [15], we set out to investigate through replication within this family of experiments how to improve the usability of a chatbot by including usability characteristics in the application.

Note that our aim was to identify the application of usability characteristics in chatbot development rather than to help teams build a better UML diagram in academic settings. The null hypotheses that govern this research question are as follows:

H.1.0 There is no significant difference in efficiency using SOCIO V1 or improved Creately when building the class diagram.

H.2.0 There is no significant difference in effectiveness using SOCIO V1 or improved Creately when building the class diagram.

H.3.0 There is no significant difference in satisfaction using SOCIO V1 or improved Creately when building the class diagram.

H.4.0 There is no significant difference in the quality of the class diagram built using SOCIO V1 or improved Creately.

As mentioned above, we developed an updated version of the chatbot SOCIO called SOCIO V1 with context-sensitive help, and the control tool Creately was equipped with a better interface that did not rely on Adobe Flash. SOCIO V1 and the improved Creately were used to perform the family of experiments.

For each experiment run, the independent variable was the modeling tool, and the chatbot SOCIO V1 and the improved online application Creately were treatments. According to the above experimental setting [15], the response variables (dependent variables) within this family were three usability

characteristics (i.e., efficiency, effectiveness, and satisfaction) and the quality of the outcome.

Based on definitions from ISO/IEC 25010:2011 [4], ISO 9241-11 [19], ISO/IEC/IEEE 29148 [20] and Hornbæk's guidelines [21], efficiency, effectiveness, and satisfaction are commonly measured characteristics for evaluating software usability. Precisely, we measure usability as follows:

1) EFFICIENCY

Efficiency is measured in terms of time to complete a task and fluency.

2) TIME

Once we completed the tutorial for the tool, participants were sent the task, and time was counted as of when the task was received. We manually recorded how many minutes each team took to complete each task. We recorded the start and stop times for remote experiments via Telegram chat. For offline, face-to-face experiments, we recorded when we asked participants to start on-site and when each team finished. Each team was given a maximum of 30 minutes to complete a task. If a team finished the task early, the time at which they finally submitted the outcome was recorded as the task completion time.

3) FLUENCY

Fluency was measured by the number of discussion messages generated by teammates. We counted the number of discussion messages manually. Discussion messages are generally about task performance, tool use, and team management topics. Any irrelevant communication or discussion messages were not counted, e.g., emotional expressions and questions put to the experimenter. Of the discussion messages, SOCIO V1 and Creately both share a common type of discussion message: messages regarding how to use the tool. To gain a better understanding of user opinions, we also analyzed this type of message in the experimental results of the discussion messages afterwards.

4) EFFECTIVENESS

We measured *effectiveness* as *completeness*, based on the perceived success of each class diagram compared with the ideal class diagram (see lab package) that we (i.e., the experimenters) built to measure the solutions produced by teams [21], [22].

To calculate the completeness score, we counted how many elements were included compared to the ideal class diagram. We counted each class, relationship and attribute as one element. For instance, the ideal class diagram for Task 1 contains 32 elements. We counted the number of elements included by the teams and divided this number by the ideal number of elements (32) to calculate the completeness score for each team completing each task. Thus, the highest score for each team is 1. Note that when counting the included elements, the name and characteristic of the element does not necessarily have to be absolutely correct. At this point, we are measuring whether the participant managed to create the element, e.g., both "college" and "university" are counted as being correct.

5) SATISFACTION

We tailored the *System Usability Scale (SUS) questionnaire* to our experiments to assess *satisfaction* quantitatively and qualitatively. Each questionnaire included 10 five-point Likert scale SUS questions (1 for "Strongly Disagree" and 5 for "Strongly Agree") and three to four open-ended questions about positive comments, negative comments, and tool suggestions. At the end of the second experimental session, we asked about participants' preferences for either of the two tools.

To calculate the numerical value of each participant's satisfaction score, we used Brooke's equation [23] below to calculate the quantitative SUS result. The team score was calculated using the median of the scores of the three team members for each question:

$$SUSscore = [\sum_{n=1}^{5}(P_{2n-1} - 1) + (5 - P_{2n})] \times 2.5. \quad (1)$$

6) QUALITY OF THE OUTCOME

We also measured the *quality of the outcome* as the quality of the class diagrams generated by the teams used as a measure of effectiveness [21].

To gauge the quality of each team's class diagram, we used an ideal class diagram as a benchmark. However, a class diagram can have more than one solution, all of which are "correct." Software engineering experts designed the ideal class diagram before the experiment was carried out. To assess quality, we employed the following metrics [24]:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (2)$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \quad (3)$$

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (4)$$

$$\text{Error} = (\text{FP} + \text{FN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (5)$$

$$\text{Success} = \text{TP}/(\#\text{Number of ideal class diagram elements}) \quad (6)$$

By comparing the ideal class diagram with the true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) for each class diagram, the following formulas were computed:

TP (true positive): Number of elements found in the ideal and team class diagrams.

FN (false negative): Number of elements found in the ideal class diagram but not in the team class diagram.

FP (false positive): Number of elements found in the team class diagram but not in the ideal class diagram.

TN (true negative): There are no true negatives in the model comparison; hence, the value is always 0.

B. DESIGN OF THE EXPERIMENTS

A baseline experiment (EXP1) and two replications (EXP2 and EXP3) form the family of experiments in academic settings. Considering the relatively small sample size of the baseline experiment (i.e., 15 subjects) and the resulting potential for inaccurate and/or biased results [25], we followed

the theoretical guidelines set out by Juristo and Gómez [5], employing an identical experimental design for all three experiments. Note that the experimental process of this second family is identical to the first family [15] in order to compare SOCIO and SOCIO V1 vertically. Each of the three experiments was structured as a two-sequence and two-period within-subject crossover design (see Table 1).

TABLE 1. Experimental design.

Group	Task 1 Period 1	Task 2 Period 2
Group1	SOCIO V1	Creately
Group2	Creately	SOCIO V1

The two replications adhere to the baseline experiment with few variations. To assure that the replications are similar, and the results are comparable, researchers reuse the same experimental protocol and experimental material employed in the baseline experiment, and the replications are jointly run with the experimenter that conducted the baseline experiment.

Three experiments were run at three different sites. The baseline experiment (EXP1) took place at the *Universidad de las Fuerzas Armadas ESPE Extensión Latacunga* (ESPE-Latacunga) in Ecuador (UNIV-1), the first replication (EXP2) was conducted at the *Universidad Autónoma de Yucatán* (UADY) in Mexico (UNIV-2), and the second replication was run at the *Escuela Politécnica Superior of the Universidad Autónoma de Madrid* (EPS-UAM) in Spain (UNIV-3).

Due to COVID-19 lockdown in Mexico, Ecuador and Spain, test sessions for EXP1 and EXP2 were organized remotely via desktop sharing and video conferencing software. EXP3 was conducted in a face-to-face manner.

As both tools are collaborative, the experiments took place in a groupwork setting. The experiments were conducted using three-member teams, and each team was construed as an experimental subject. In each experiment, participants were randomly assigned to one of two groups (Group 1 or Group 2) and then grouped into three-member teams. Accordingly, each group applied the treatments differently (SOCIO V1-Creately/Creately-SOCIO V1). The experimental design is blocked by the period (i.e., the task).

At the beginning of the experiment, each participant was asked to complete a familiarity questionnaire and a consent form. After a 10-minute introduction to the tool that the participants would be using before each period, they were given a maximum of 30 minutes to complete the task using the tool. Group 1 carried out Task 1 with SOCIO V1 in the first period and Task 2 with Creately in the second (i.e., SOCIO V1-Creately sequence). On the other hand, Group 2 completed Task 1 in the first period using Creately and then completed Task 2 using SOCIO V1 in the second period (i.e., Creately-SOCIO V1 sequence). All participants were asked to complete a modified and validated SUS questionnaire connected with the tool they had just used following the completion of each experimental task (i.e., all participants had to fill in the modified SUS questionnaire twice with respect to two modeling tools). Additionally, participants were asked in the second SUS questionnaire whether they preferred SOCIO V1 or Creately.

Two distinct experimental tasks were designed (each assigned to a different experimental period). The first task was to create a class diagram for an online store that includes product and customer management. The second task was to create a class diagram for a college in order to facilitate the organization of courses and pupils. The complexity of the class diagrams was adapted to the duration of the experimental periods. Throughout the experiment, participants of the same team were only permitted to communicate via Telegram groups. This was done to ensure that all experimental data was captured. From the first family of experiments, we observed that most participants tended to run out of time. This may have affected their task completeness. Considering that (1) subject availability was limited and subject fatigue needed to be avoided and (2) we would not be able to measure the effectiveness variable if all the participants had had the option of completing each task, we did not extend the time limit in the following experimental series.

C. SAMPLE

The participants in our family of experiments were students recruited using the convenience sampling method at UNIV-1, UNIV-2, and UNIV-3. The sample in the family was composed of 96 participants: 45 students at UNIV-1, 27 students at UNIV-2, and 24 students at UNIV-3. All the participants were students completing a BSc in Computer Science degree or Joint BSc in Computer Science and Mathematics. Because SOCIO is a modeling chatbot, users had to be acquainted with UML in order to build the model (class diagram). In view of this, we recruited only students with a background in computer science or related fields to ensure that the participants would be able to complete the modeling tasks. To guarantee that all interactions between team members were conducted via Telegram, we made sure that the students' professors were present to oversee the process. In addition, each teammate was seated separately to make sure that there was no whispering.

For technical and methodological reasons (e.g., system failure, incomplete questionnaires, experiment withdrawals), teams 8 and 14 from EXP1 and 6 from EXP2 did not complete the experiment. The study was, therefore, limited to only 87 participants (see Table 2). The sample included 14 women and 73 men who ranged in age from 20-27 (mean 22.54, SD 1.29).

A postal survey was carried out with 87 participants. Table 3 summarizes the analysis performed on aggregated familiarity results.

Considering that 93% of participants had experience with Telegram and 66% used Telegram regularly, we believe that the use of social networking platforms does not affect chatbot usability. However, 34% of participants had no experience with chatbots, and 26% had little knowledge of chatbots,

TABLE 2. Overview of subjects.

EXP	Time	Affiliation	#Participants	Teams
EXP1	Jan 2021	UNIV-1	45	15
EXP2	Jan 2021	UNIV-2	27	9
EXP3	Nov 2021	UNIV-3	24	8

TABLE 3. Familiarity result.

Have you ever used Telegram?	
Yes	93%
No	7%
Have you ever used a chatbot?	
Yes	66%
No	34%
Which social networks do you use regularly? (Multiple choice)	
WhatsApp	98%
Telegram	66%
Twitter	52%
Facebook Messenger	59%
Instagram	78%
Rate your English level	
5	15%
4	24%
3	49%
2	11%
1	0%
Rate your knowledge of class diagrams	
5	5%
4	46%
3	38%
2	7%
1	0%
Rate your knowledge of chatbots	
5	5%
4	10%
3	36%
2	18%
1	26%

which could detract from the sensitivity and credibility of the experimental results. In addition, we also asked about the level of English: 88% of the participants believed that they had at least an intermediate level of English. Because the task did not require complex English communication, we believe that their English proficiency was good enough to get the job done.

V. RESULTS AND AGGREGATION OF DATA

To answer the research questions, we provide a quantitative and qualitative description of the nature of this study for data synthesis and analysis.

For quantitative analysis, we performed a global analysis of the whole family of experiments and illustrated the individual experiments. The descriptive statistics and violin plots were used to provide readers or other researchers with a better understanding of the normality of each experimental data item. We analyzed the quantitative family result following Santos et al.'s guidelines [26]. The individual participant data (IPD) meta-analysis approach combined with a three-factor LMM was used to study the effect on the outcomes of multiple factors (e.g., period, treatment, and sequence) [10], [27]. We added a parameter to the LMM to account for differences in outcomes across experiments (i.e., Experiment). We then used the corresponding ANOVA table of the LMM to illustrate the statistical significance of the results.

Finally, we adopted the thematic analysis method [28] to gain further insight into user responses and analyze the qualitative data of the three open-ended questions.

A. QUANTITATIVE ANALYSIS

The following section analyzes each response variable (i.e., efficiency, effectiveness, satisfaction, and quality). We concentrate on their respective metrics (i.e., time and discussion messages for efficiency; completeness for effectiveness; satisfaction for satisfaction; and precision, recall, accuracy, error, and perceived success for quality).

For each metric, we provide: (i) descriptive statistics and violin plots divided by treatment (i.e., SOCIO V1, Creately) and by experiment (i.e., EXP1, EXP2, and EXP3), and (ii) the results of all the experiments pooled using a one-stage IPD meta-analysis and the contrast between treatments across the experiments [29].

1) FIRST VALIDATION FOR H.1.0: EFFICIENCY

Efficiency was measured in terms of time and fluency. Time is the amount of time taken to accomplish the tasks. Fluency refers to the number of discussion messages exchanged between teammates during class diagram development. Figs. 5 and 6 illustrate the violin plot for time and fluency across the experiments. The respective summaries of descriptive statistics are shown in Tables 4 and 7, grouped by experiment and treatment.

a: TIME

As shown in Fig. 5, the aggregate task completion time with SOCIO V1 appears to be similar to Creately in EXP1 and slightly less than Creately in EXP2 and EXP3. As the descriptive statistics (Table 4) show, time spent on task performance appears to be similar for both Creately and SOCIO V1. Besides, as shown in the ANOVA table (Table 5) and the pairwise contrast between the treatments (Table 6), a negligible – and statistically non-significant– difference in the time was observed between Creately and SOCIO V1 (0.43 minutes). In sum, Creately and SOCIO V1 appear to perform similarly in terms of time.

Interestingly, we identified a trend where most teams in EXP1 and EXP2 spent as long as possible on completing and/or improving their class diagrams. Accordingly, we observed relatively lower task completeness than for EXP3. Based on these observations, a possible conclusion is that these participants needed more time to accomplish the task.

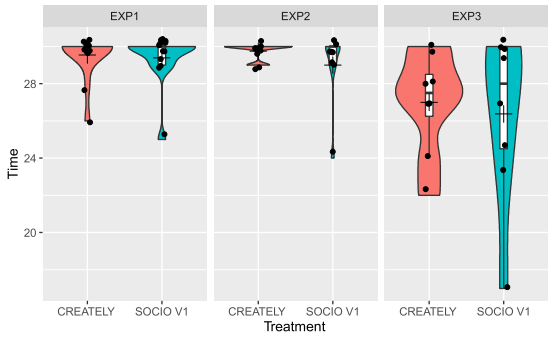


FIGURE 5. Violin plot for time spent on tasks (jitter added to the points).

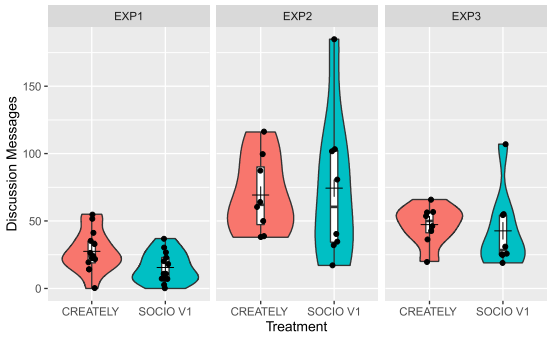


FIGURE 6. Violin plot for discussion messages (jitter added to the points).

TABLE 4. Descriptive statistics for time spent on tasks.

EXP	Treatment	Team	Mean	Std. Dev.	Median
EXP1	Creately	13	29.54	1.20	30.0
EXP1	SOCIO V1	13	29.39	1.39	30.0
EXP2	Creately	8	29.75	0.46	30.0
EXP2	SOCIO V1	8	29.00	2.07	30.0
EXP3	Creately	8	27.00	2.78	27.5
EXP3	SOCIO V1	8	26.38	4.60	28.0

TABLE 5. ANOVA table for time.

Measure	numDF	denDF	F-value	p-value
(Intercept)	1	27	7110.427	<.0001
Sequence	1	25	3.650	0.0676
Treatment	1	27	0.972	0.3329
Period	1	27	1.219	0.2792
Experiment	2	25	6.349	0.0059

TABLE 6. Contrast between treatments for time.

Contrast	Estimate	SE	df	t-radio	p-value
CR-SC	0.431	0.455	27	0.947	0.3519

TABLE 7. Descriptive statistics for discussion messages.

EXP	Treatment	Team	Mean	Std. Dev.	Median
EXP1	Creately	13	27.46	15.53	24.0
EXP1	SOCIO V1	13	15.46	11.33	11.0
EXP2	Creately	8	69.25	28.84	62.0
EXP2	SOCIO V1	8	74.38	55.66	60.5
EXP3	Creately	8	47.25	14.43	50.0
EXP3	SOCIO V1	8	42.75	29.32	28.5

TABLE 8. ANOVA table for discussion messages.

Measure	numDF	denDF	F-value	p-value
(Intercept)	1	27	86.05308	<.0001
Sequence	1	25	0.54658	0.4666
Treatment	1	27	1.43370	0.2416
Period	1	27	8.35061	0.0075
Experiment	2	25	10.85980	0.0004

TABLE 9. Contrast between treatments for discussion messages.

Contrast	Estimate	SE	df	t-radio	p-value
CR-SC	5.64	4.35	27	1.296	0.2058

b: DISCUSSION MESSAGES

Bear in mind that people have different messaging styles: some prefer to send a variety of short messages in succession, and others prefer to send long messages. As mentioned in [15], we counted each sentence containing the complete subject, predicate, and object as one discussion message.

As the violin plot (Fig. 6) and descriptive statistics (Table 7) show, the participants appear to send more messages with Creately than with SOCIO V1 in two out of our three experiments (EXP1 and EXP3). The opposite holds for EXP2. As ANOVA (Table 8) and the contrast table (Table 9) show, a negligible –and statistically non-significant– difference in the discussion message was observed between SOCIO V1 and Creately (5.64). This suggests that **SOCIO V1 and Creately appear to perform similarly in terms of discussion messages.**

2) SECOND VALIDATION FOR H.1.0: EFFICIENCY-TOOL USAGE MESSAGES

In the knowledge that there was a wide range of discussion messages, they were classified into the following types: task performance (e.g., how to divide labor), tool use, and discussions about UML knowledge. However, as we were researching chatbot usability, we were interested in discussions on tool usage, that is, how to use the tools properly. Therefore, we extracted discussion messages of this type and then performed an additional analysis.

As the plot (Fig. 7) and the descriptive statistics (Table 10) show, the participants are more likely to send more messages on proper tool use with Creately than with SOCIO V1. Besides, as shown in the ANOVA table (Table 11), the difference between the number of tool usage messages is statistically significant (p-value <0.05). According to the pairwise contrast between the treatments in Table 12, **the**

TABLE 10. Descriptive statistics for tool usage messages.

EXP	Treatment	Team	Mean	Std. Dev.	Median
EXP1	Creately	13	10.08	8.48	7.0
EXP1	SOCIO V1	13	2.62	1.94	3.0
EXP2	Creately	8	19.88	7.94	16.5
EXP2	SOCIO V1	8	4.75	1.83	4.0
EXP3	Creately	8	9.13	3.04	10.0
EXP3	SOCIO V1	8	13.25	9.77	8.5

participants using Creately sent up to 6.38 more tool usage messages than SOCIO V1 users.

In sum, we cannot reject the null hypothesis H.1.0. SOCIO V1 and Creately appear to perform similarly regarding time and discussion messages. However, SOCIO V1 has the advantage of reducing the communication effort on tool usage for the participants with respect to the first experiment [15].

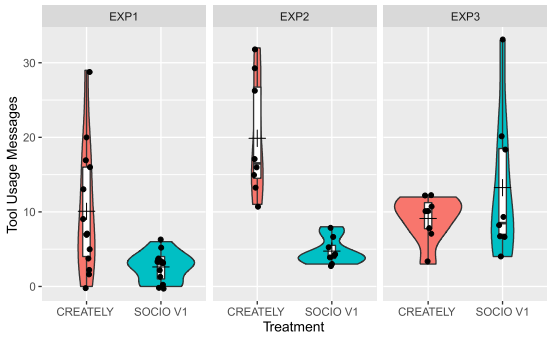


FIGURE 7. Violin plot for tool usage messages (jitter added to the points).

TABLE 11. ANOVA table for tool usage messages.

Measure	numDF	denDF	F-value	p-value
(Intercept)	1	27	91.89526	<.0001
Sequence	1	25	0.03364	0.8560
Treatment	1	27	10.74951	0.0029
Period	1	27	0.00359	0.9527
Experiment	2	25	3.93138	0.0328

TABLE 12. Contrast between treatments for tool usage messages.

Contrast	Estimate	SE	df	t-radio	p-value
CR-SC	6.38	1.95	27	3.279	0.0029

The results of the first family [15] show that SOCIO is significantly more efficient than Creately. We acknowledge that this is mainly due to the fact that the previous version of Creately relied on Adobe Flash, which caused the software to be unstable and resulted in many participants having to quit and re-enter (this was also confirmed by the qualitative analysis, with many participants complaining about this). We noticed that the current version of Creately is no longer dependent on Adobe Flash, and we believe that, based on the experimental data, this has effectively improved Creately's efficiency.

3) VALIDATION FOR H.2.0: EFFECTIVENESS

Completeness. We measured effectiveness by the degree of task completeness.

As the violin plot (Fig. 8), descriptive statistics table (Table 13), and contrast table (Table 15) show, SOCIO V1 has a slight (0.0752) edge over Creately in terms of completeness. As we can see in the ANOVA table (Table 14), the treatment has a statistically significant impact on completeness. In sum, **SOCIO V1 outperforms Creately with respect to effectiveness**.

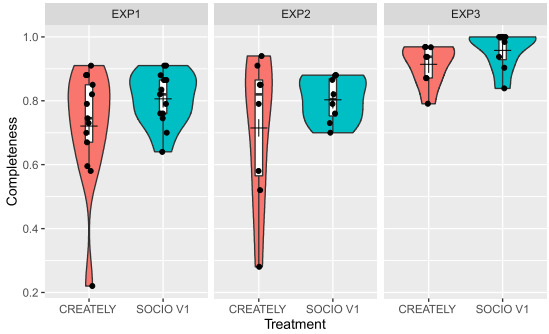


FIGURE 8. Violin plot for completeness (jitter added to the points).

TABLE 13. Descriptive statistics for completeness.

EXP	Treatment	Team	Mean	Std. Dev.	Median
EXP1	Creately	13	0.72	0.18	0.75
EXP1	SOCIO V1	13	0.81	0.08	0.82
EXP2	Creately	8	0.72	0.23	0.82
EXP2	SOCIO V1	8	0.80	0.07	0.81
EXP3	Creately	8	0.91	0.064	0.94
EXP3	SOCIO V1	8	0.96	0.06	0.99

TABLE 14. ANOVA table for completeness.

Measure	numDF	denDF	F-value	p-value
(Intercept)	1	27	2028.5640	<.0001
Sequence	1	25	0.0000	0.9970
Treatment	1	27	4.4784	0.0437
Period	1	27	0.1387	0.7125
Experiment	2	25	9.3814	0.0009

TABLE 15. Contrast between treatments for completeness.

Contrast	Estimate	SE	df	t-radio	p-value
CR-SC	-0.0752	0.0353	27	-2.128	0.0426

In sum, we reject the null hypothesis H.2.0. Compared with the results of the first family [15], completeness has improved by 7.52%, and this improvement is relevant for chatbot usability. After adding context-sensitive help to SOCIO, we observed that SOCIO V1 outperformed Creately on completeness.

4) VALIDATION FOR H.3.0: SATISFACTION

We adopted a modified SUS questionnaire to assess user satisfaction with SOCIO V1 and Creately. Each questionnaire consists of 10 SUS questions and three to four open-ended questions. In this section, we report a quantitative analysis of the responses to the SUS questions. The analysis of the responses to the open-ended questions will be reported in the qualitative analysis section.

Satisfaction Score. Fig. 9 shows the violin plot for the mean SUS scores across experiments. The respective summary of descriptive statistics is shown in Table 16, grouped by experiment and treatment.

As the violin plot (Fig. 9) and descriptive statistics table (Table 16) show, the satisfaction scores for SOCIO V1 are typically higher than for Creately. Besides, as the ANOVA table (Table 17) shows, the difference between the satisfaction scores is statistically significant. In sum, SOCIO V1 appeared to consistently satisfy participants more than Creately in the second family and widened the gap in satisfaction from 6.16 to 8.9 (see Table 18). In sum, we rejected the null hypothesis H.3.0. Compared to the first family, an improvement of 8.9 in the second family indicates that satisfaction has improved by 8.9%, and this improvement is worthwhile from the point of view of chatbot usability. Additionally, the satisfaction score of the second family is statistically significant.

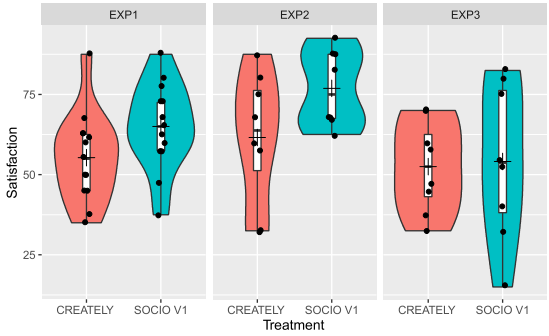


FIGURE 9. Violin plot for satisfaction (jitter added to the points).

TABLE 16. Descriptive statistics for satisfaction.

EXP	Treatment	Team	Mean	Std. Dev.	Median
EXP1	Creately	13	55.29	13.98	55.00
EXP1	SOCIO V1	13	65.00	13.58	65.00
EXP2	Creately	8	61.56	20.48	63.75
EXP2	SOCIO V1	8	76.88	11.78	75.00
EXP3	Creately	8	52.50	14.14	52.50
EXP3	SOCIO V1	8	54.06	24.24	53.75

5) VALIDATION FOR H.4.0: QUALITY

We analyzed the quality of the class diagrams using five metrics (cf. equations (2) - (6)): precision, recall, accuracy, error, and perceived success.

The violin plots for these metrics are shown in Figs. 10, 11, 12, 13, and 14, respectively. The respective summary of descriptive statistics is shown in Table 19, grouped by metric, experiment, and treatment. The summaries of the ANOVA test and contrast between treatments are shown in Tables 20 and 21, respectively.

TABLE 17. ANOVA table for satisfaction.

Measure	numDF	denDF	F-value	p-value
(Intercept)	1	27	802.2736	<.0001
Sequence	1	25	1.3088	0.2634
Treatment	1	27	4.4109	0.0452
Period	1	27	0.4925	0.4888
Experiment	2	25	3.8526	0.0348

TABLE 18. Contrast between treatments for satisfaction.

Contrast	Estimate	SE	df	t-radio	p-value
CR-SC	-8.9	4.29	27	-2.075	0.0477

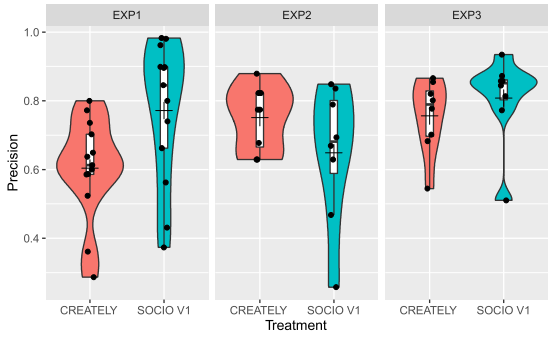


FIGURE 10. Violin plot for precision (jitter added to the points).

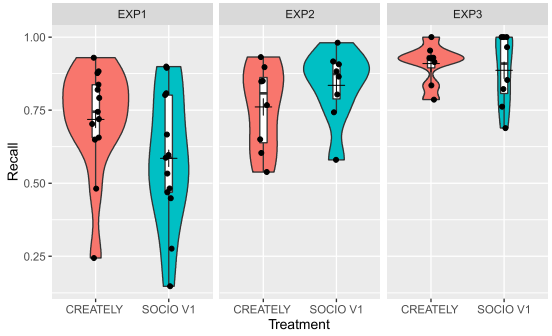


FIGURE 11. Violin plot for recall (jitter added to the points).

a: PRECISION AND PERCEIVED SUCCESS

Regarding Precision and Perceived Success, the violin plots (Figs. 10 and 14) and descriptive statistics table (Table 20) show that SOCIO V1 slightly outperforms Creately in two out of three experiments.

b: RECALL AND ACCURACY

Regarding Recall and Accuracy, the violin plots (Fig. 11, 13) and descriptive statistics table (Table 19) show

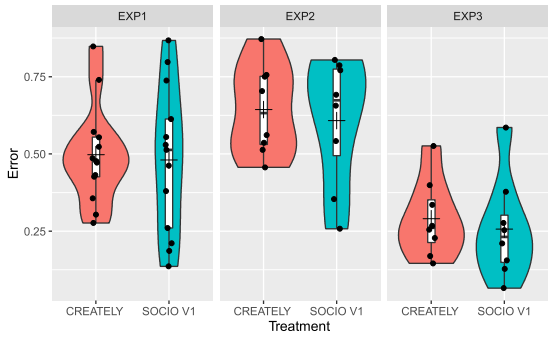


FIGURE 12. Violin plot for error (jitter added to the points).

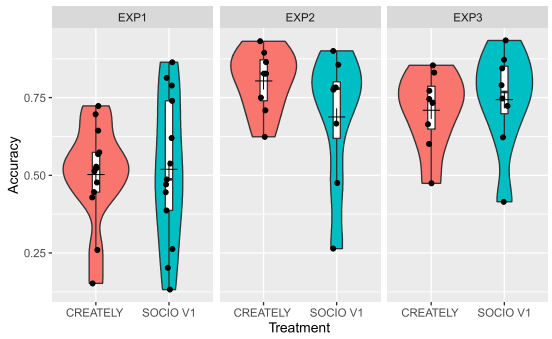


FIGURE 13. Violin plot for accuracy (jitter added to the points).

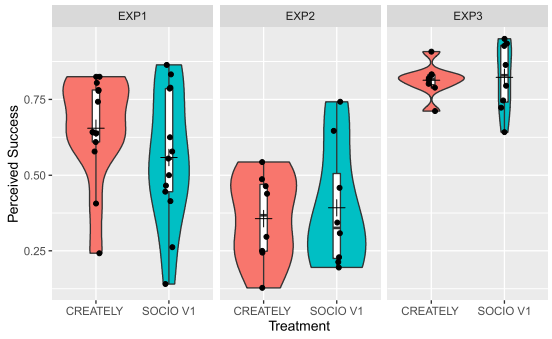


FIGURE 14. Violin plot for perceived success (jitter added to the points).

that Creately slightly outperforms SOCIO V1 in two out of three experiments.

c: ERROR

Regarding Error, the violin plots (Fig. 12) and descriptive statistics table (Table 19) show that SOCIO V1 slightly outperforms Creately across all three experiments.

However, based on the analysis of these five-quality metrics, we did not observe any statistically significant treatment. Summing up the above analysis, as the plots, descriptive statistics, ANOVA, and contrast table show, **Creately and SOCIO V1 both tend to return class diagrams of similar quality**.

TABLE 19. Descriptive statistics for quality.

Variable	EXP	Treatment	Team	Mean	Std. Dev.	Median
Precision	EXP1	Creately	13	0.60	0.15	0.61
	EXP1	SOCIO V1	13	0.77	0.21	0.85
	EXP2	Creately	8	0.75	0.09	0.77
	EXP2	SOCIO V1	8	0.65	0.20	0.68
	EXP3	Creately	8	0.76	0.11	0.79
	EXP3	SOCIO V1	8	0.81	0.13	0.85
Recall	EXP1	Creately	13	0.72	0.19	0.74
	EXP1	SOCIO V1	13	0.59	0.23	0.59
	EXP2	Creately	8	0.76	0.15	0.81
	EXP2	SOCIO V1	8	0.83	0.13	0.87
	EXP3	Creately	8	0.91	0.07	0.93
	EXP3	SOCIO V1	8	0.89	0.12	0.91
Accuracy	EXP1	Creately	13	0.50	0.16	0.52
	EXP1	SOCIO V1	13	0.52	0.24	0.49
	EXP2	Creately	8	0.80	0.10	0.83
	EXP2	SOCIO V1	8	0.69	0.22	0.78
	EXP3	Creately	8	0.71	0.13	0.74
	EXP3	SOCIO V1	8	0.74	0.16	0.77
Error	EXP1	Creately	13	0.50	0.16	0.48
	EXP1	SOCIO V1	13	0.48	0.24	0.51
	EXP2	Creately	8	0.64	0.15	0.63
	EXP2	SOCIO V1	8	0.61	0.21	0.67
	EXP3	Creately	8	0.29	0.13	0.26
	EXP3	SOCIO V1	8	0.26	0.16	0.23
Perceived Success	EXP1	Creately	13	0.65	0.17	0.64
	EXP1	SOCIO V1	13	0.56	0.22	0.55
	EXP2	Creately	8	0.36	0.15	0.37
	EXP2	SOCIO V1	8	0.39	0.21	0.33
	EXP3	Creately	8	0.81	0.06	0.82
	EXP3	SOCIO V1	8	0.82	0.11	0.83

TABLE 20. ANOVA table for quality.

Variable	Measure	numDF	denDF	F-value	p-value
Precision	(Intercept)	1	27	1333.4514	<.0001
	Sequence	1	25	0.3530	0.5578
	Treatment	1	27	0.4316	0.1306
	Period	1	27	14.6171	0.0007
	Experiment	2	25	2.1285	0.1401
Recall	(Intercept)	1	27	1546.8766	<.0001
	Sequence	1	25	0.0675	0.7972
	Treatment	1	27	1.3758	0.2511
	Period	1	27	19.5103	0.0001
	Experiment	2	25	14.6336	0.0001
Accuracy	(Intercept)	1	27	907.8200	<.0001
	Sequence	1	25	0.0727	0.7897
	Treatment	1	27	0.1277	0.7236
	Period	1	27	15.7263	0.0005
	Experiment	2	25	14.2395	0.0001
Error	(Intercept)	1	27	468.7157	<.0001
	Sequence	1	25	0.0006	0.9813
	Treatment	1	27	0.3846	0.5403
	Period	1	27	12.3023	0.0016
	Experiment	2	25	1807748	<.0001
Perceived Success	(Intercept)	1	27	814.8571	<.0001
	Sequence	1	25	0.0814	0.7778
	Treatment	1	27	0.5399	0.4688
	Period	1	27	8.5950	0.0068
	Experiment	2	25	30.7001	<.0001

In sum, we do not reject the null hypothesis H.4.0. In contrast to the result for the first family, which showed that Creately outperformed SOCIO in terms of recall and perceived success and SOCIO outperformed Creately on precision, we did not observe a statistically significant difference in the second family after both tools had been improved.

TABLE 21. Contrast between treatments for quality.

Variable	Contrast	Estimate	SE	df	t-radio	p-value
Precision	CR-SC	-0.056	0.0393	27	-1.427	0.1652
Recall	CR-SC	0.0512	0.0387	27	1.325	0.1964
Accuracy	CR-SC	0.0207	0.042	27	0.494	0.6254
Error	CR-SC	0.0215	0.0432	27	0.499	0.6219
Perceived Success	CR-SC	0.0352	0.0421	27	0.835	0.4108

6) DISCUSSION OF ANALYSIS

From the aggregation of this family of experiments result, we observed that participants seemed to have higher task completeness with SOCIO V1 compared to Creately, and they appeared to be more satisfied with SOCIO V1 than Creately. However, the two tools perform similarly in terms of efficiency and quality of class diagrams.

B. QUALITATIVE ANALYSIS–THEMATIC ANALYSIS

We enacted the thematic analysis process as follows. After each experiment session, the participants were asked to complete a modified SUS questionnaire containing three or four open-ended questions regarding (i) three positive aspects, (ii) three negative aspects, (iii) three suggestions concerning the tool they had just used, and (iv) their preference for either tool (response required only after the second session). The response to open-ended questions was transcribed into English. Due to the need to identify recurring themes to identify interesting aspects, we coded features that were mentioned more than three times in the qualitative dataset.

As shown in Figs. 15 and 16, we identified six features based on satisfaction measures for SOCIO V1 and Creately [21]: content, task, collaboration, communication, user experience, and interface. We expected these results to contribute to the development of future real-time collaboration tools, particularly chatbots, and improve user-perceived usability. Figs. 17 and 18 are bar graphs that illustrate thematic analysis sub-themes, providing a more simplified and readable analysis. The orange bars represent user suggestions for the tool, the gray bars indicate negative comments, and the blue bars are positive comments.

In general, both tools received a similar number of reviews for each of the three open-ended questions. For example, SOCIO V1 received 245 positive comments, and Creately received 244. SOCIO V1 received 198 negative comments and Creately received 177. SOCIO V1 received 72 suggestions and Creately received 63.

1) CONTENTS

SOCIO V1 outperforms Creately in terms of contents, given that it receives more positive comments (26 vs. 16), fewer criticisms (7 vs. 23), and no suggestions for improvement, whereas Creately receives 3.

Users consider both SOCIO V1 and Creately to be helpful for integrated content and design implementation purposes (e.g., ''useful for creating class diagrams'', ''useful tool for UML''). Moreover, this feature is more prominent in SOCIO V1, as it is mentioned by almost twice as many users than for Creately (11 vs. 6).

Content errors appear to be the most commonly reported faults with respect to content. Twice as many Creately users as SOCIO V1 users report errors (13 vs. 7), although they do not suggest respective improvements.

These bugs are mainly related to server and page response errors in both cases.

Apart from usefulness and errors, the remaining aspects described as positive and negative differ for the two tools. SOCIO V1 appears to be positively rated on innovation (15), as the experimental subjects regard a chatbot for building class diagrams as innovative and surprising (e.g., ''innovation in creating UML''). By contrast, Creately's content design for commercial purposes caused controversy. On versatility (5), which provides additional content for UML diagram creation and is free of charge (5), it stands out slightly compared to other tools with similar features that offer paid services. In contrast, other users also consider these features to be a weakness.

On the one hand, seven people indicate that it offers too many options that are not used for elaborating UML diagrams, and three participants even suggest reducing these options. On the other hand, three participants were also dissatisfied with the fact that, although the main functionality is free, it includes some paid features (3).

2) TASK

SOCIO V1 receives conflicting feedback on task completion, with 24 favorable and 23 negative comments, respectively. By contrast, Creately earns more negative comments (37) than positive ones (20). Compared to Creately, SOCIO V1 receives more positive feedback on task completion.

Regarding task completion, participants are most concerned with the tool's functionality and efficiency since they were mentioned most often. In terms of time efficiency, SOCIO V1 outperforms Creately, with 11 participants stating that the development of UML diagrams does not take very long (e.g., ''Creating the diagram is extremely fast,'' ''It works fast''), compared to only four comments for Creately.

However, when it came to evaluating the completeness of the tool functionality for task performance, both tools were found to have functional flaws in terms of missing class diagram elements and missing actions that need to be performed. Participants have mixed feelings about both tools; some believed the functionality was complete, while others reported flaws and suggested adding new functions.

Missing functionality was noteworthy in SOCIO V1. Whereas eight participants praised its comprehensive functionality, and five participants liked the fact that SOCIO V1 automatically and simply generates the relationships in the

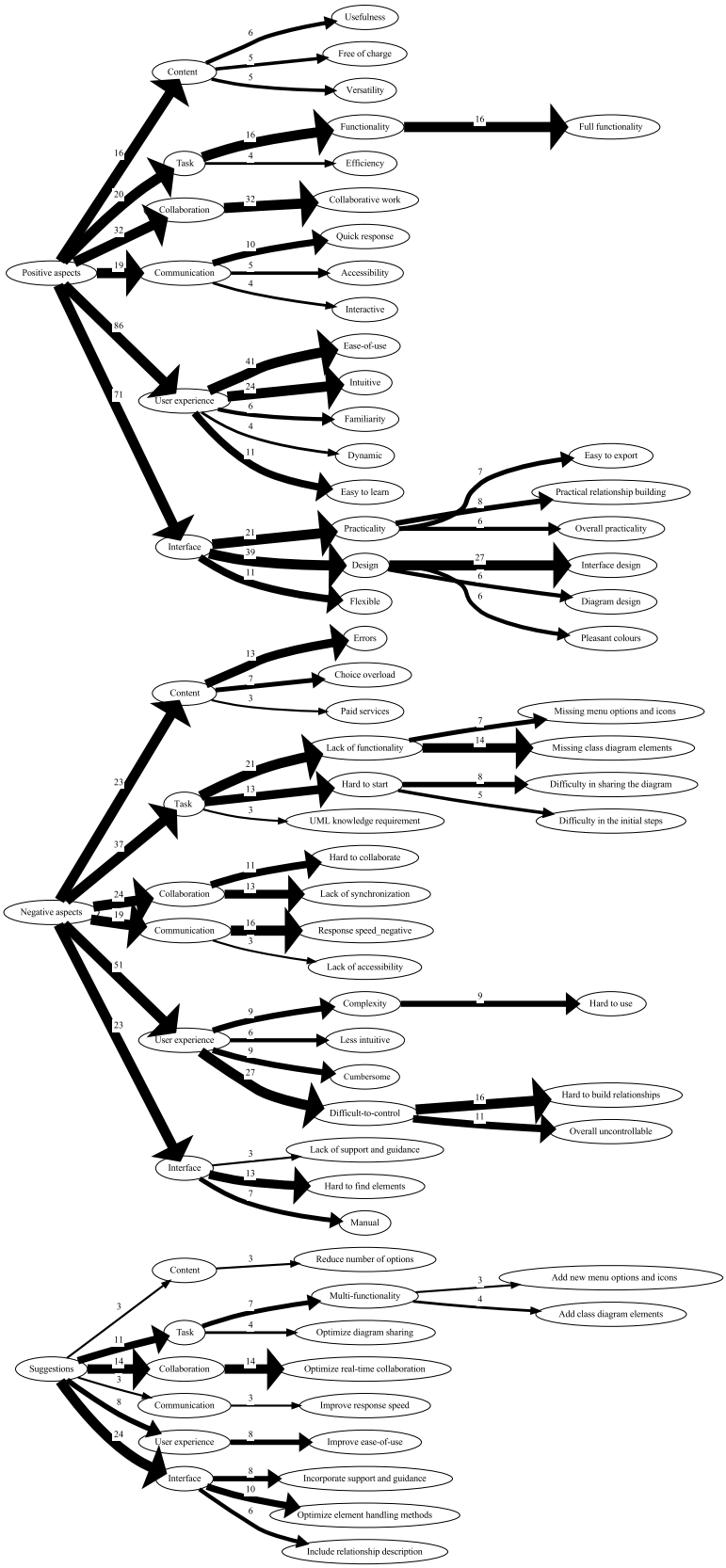


FIGURE 15. Thematic analysis for Creately.

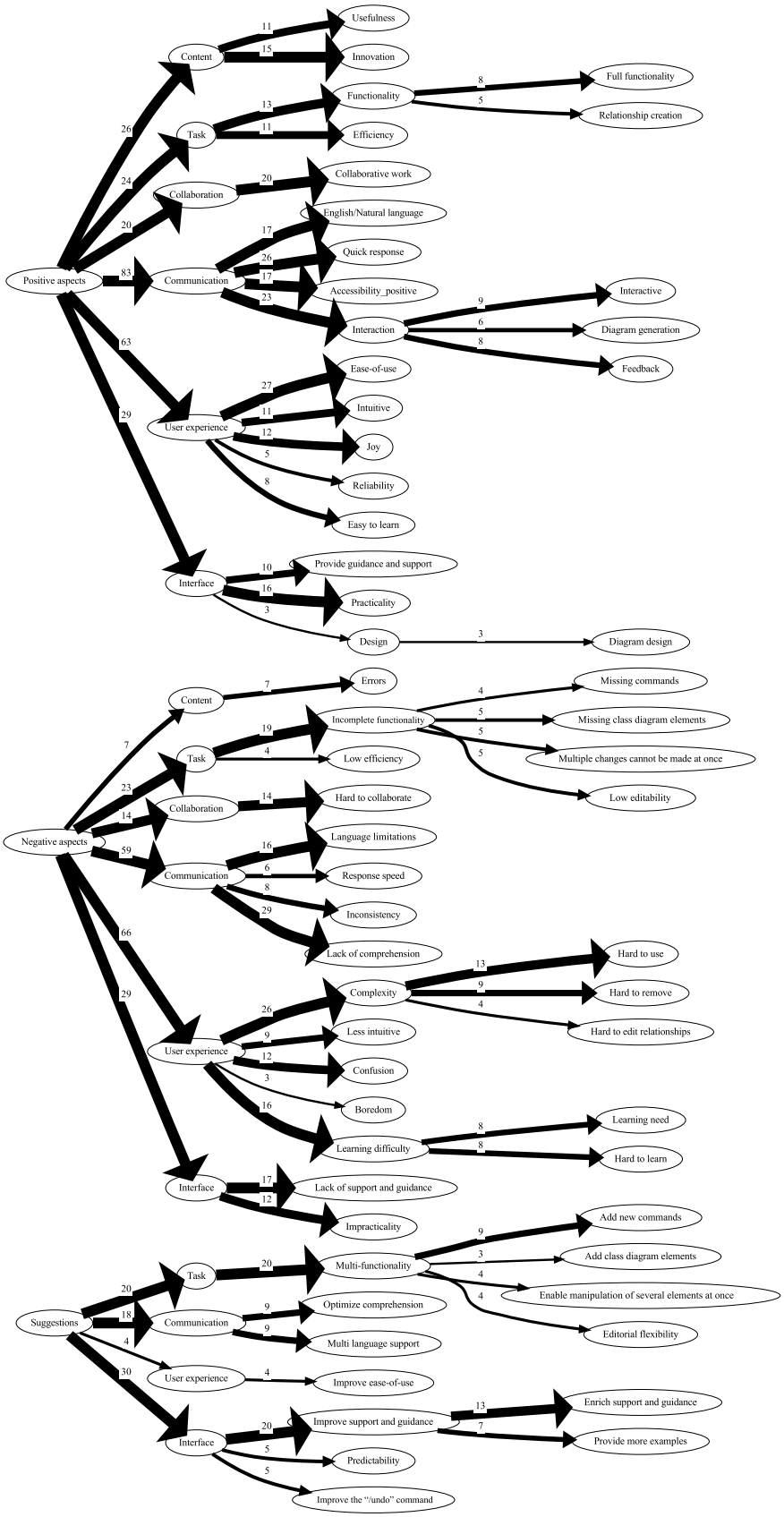


FIGURE 16. Thematic analysis for SOCIO.

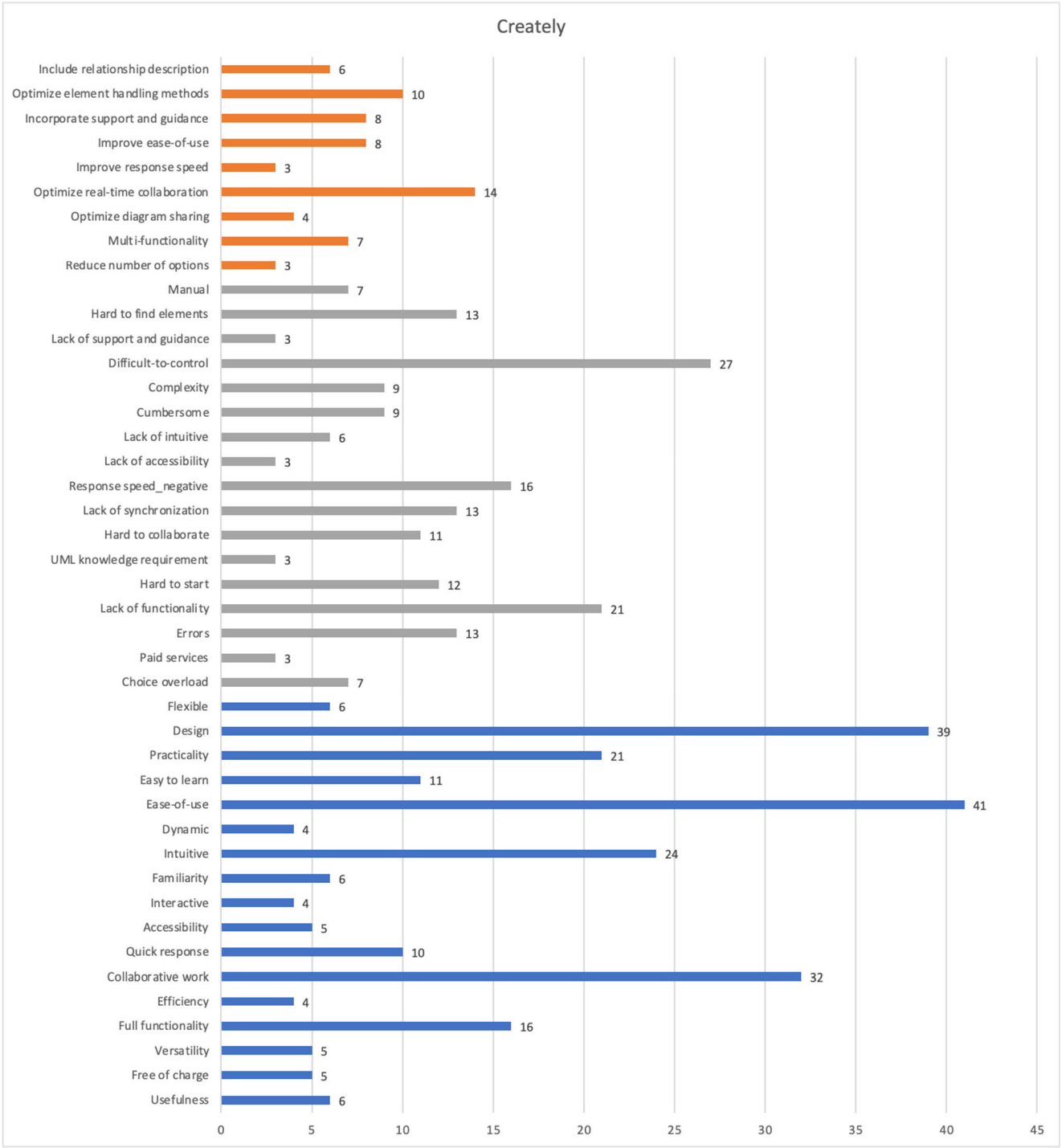


FIGURE 17. Bar graph of Creately's thematic analysis.

UML diagram (''Establishing relationships is easy,'' ''Automatically creates links between classes''), 19 participants indicated that they missed functionality, and 20 suggested that new functionalities should be added. They mentioned, for instance, that (i) it is hard to edit class diagram elements (e.g., ''You cannot modify class names,'' ''Little modification of the diagram''), (ii) it is not possible to operate many diagram pieces at once (e.g., ''I believe you cannot create several things at the same time,'' ''It does not place the attributes in a group way, it places them one by one''), and (iii) data types are missing (e.g., ''Limited attribute types,'' ''No Singleton option''). Opinions vary widely with respect to functionality completeness in Creately, with 16 participants praising its functionality and 21 participants expressing dissatisfaction with missing functionality. For example, the participants were dissatisfied with missing relationships to

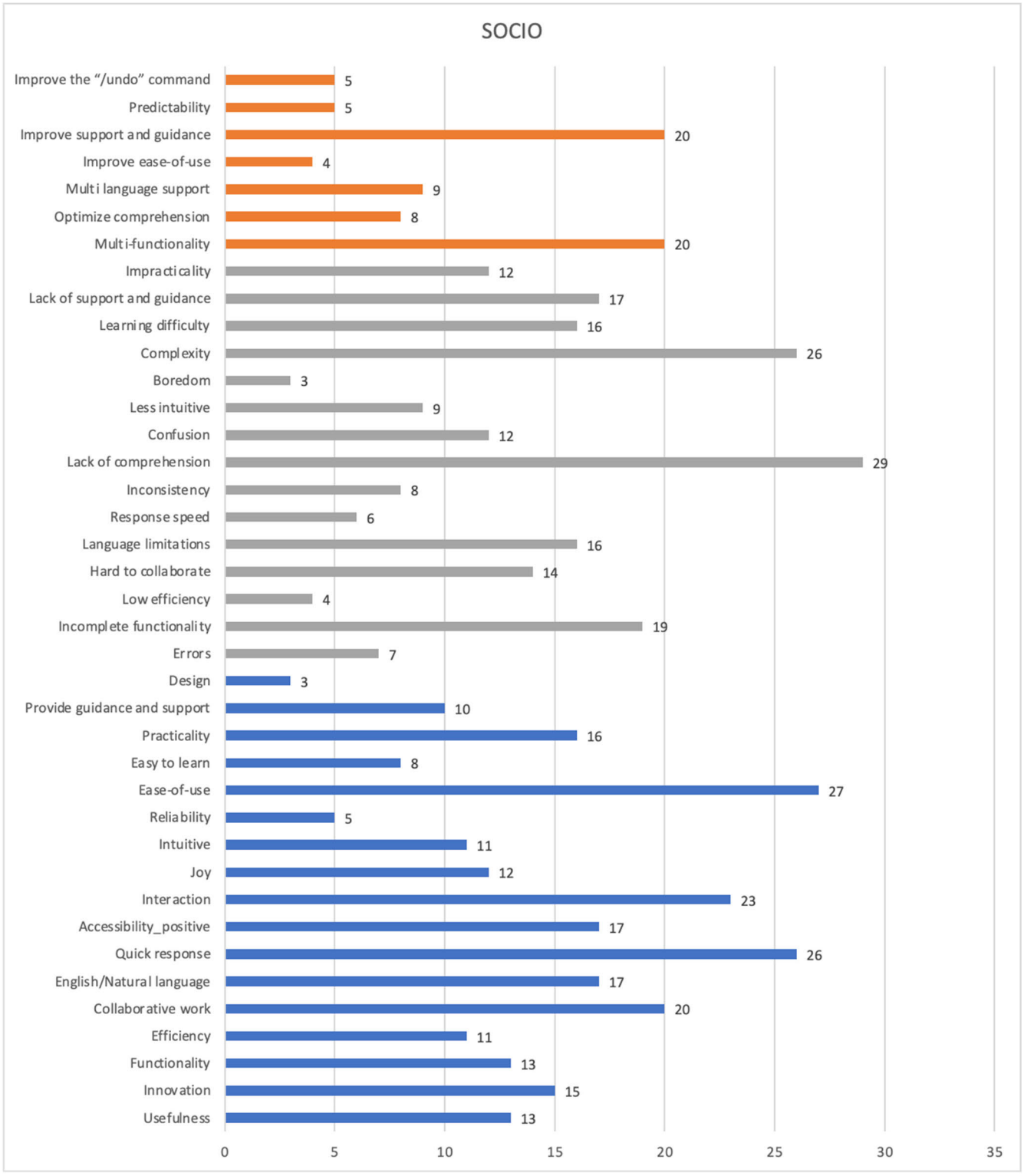


FIGURE 18. Bar graph of SOCIO's thematic analysis.

link classes (e.g., ''UML connector types,'' ''Not all UML associations''). Also, 13 Creately users had trouble signing up, logging in, and creating the document, especially when sharing the project (8).

3) COLLABORATION

The collaboration feature refers to real-time collaboration, and both tools garnered more positive than negative feedback overall. Surprisingly, on the one hand, Creately's cooperation

performance was complimented by more users (32) than SOCIO V1's (20). On the other hand, Creately earned more negative feedback in terms of collaborative capacity (24 vs. 14).

Both tools received a similar number of positive assessments for supporting real-time collaboration. However, Creately received 12 more positive reviews than SOCIO V1. We can conclude that SOCIO V1 outperforms Creately in terms of collaboration. Both tools have garnered criticism for being difficult to work with. Creately received 11 complaints (e.g., ''It's challenging to cooperate'', ''It's confusing to work with numerous individuals''), whereas SOCIO V1 received 13 (e.g., ''(It's) tough to utilize in teams,'' ''It's confusing to work in groups''). Furthermore, we observed that Creately had a specific synchronization flaw, as 13 participants found it difficult to keep up with modifications made by their teammates (e.g., ''Sometimes it takes a while to synchronize,'' ''There is a little delay when collaborating''). Based on the above, Creately received 14 suggestions on how to improve cooperation, such as integrating chat.

4) COMMUNICATION

The interaction between users and tools is referred to as communication. SOCIO V1 receives significantly more positive and negative feedback, and suggestions, than Creately on communication.

Participants provide feedback on three themes common to both tools: response time, accessibility, and interaction. SOCIO V1 outperforms Creately on each of these aspects. Roughly three times as many users praise SOCIO V1 for quick reaction time than Creately (26 vs. 10). Furthermore, only six participants consider SOCIO V1 to have a slow response time as opposed to 16 for Creately. SOCIO V1 is more accessible than Creately as it benefits from being a social media-based tool. SOCIO V1 also outperforms Creately in terms of interaction. Although both tools receive positive feedback, Creately receives four positive comments while SOCIO V1 receives 23.

Of these 23 opinions, six participants appreciated the fact that SOCIO V1 returns the updated diagram after each action (e.g., ''Shows the diagram after each command''), while eight highlighted SOCIO V1's help system in response to user errors (e.g., ''If you make a mistake in a command, it corrects you instantly,'' ''Provides good feedback'').

In addition, the experimental participants expressed positive and negative thoughts on specific chatbot aspects. Seventeen participants positively rated communication with SOCIO V1 through natural language. However, it also received a disproportionately large amount of negative feedback on natural language comprehension:

- 29 participants stated that the chatbot does not understand sentences entered to build the diagram (e.g., ''the chatbot sometimes does not understand what I enter,'' ''Limited language'').
- 16 participants complained that the chatbot only understands English sentences.
- Eight participants stated that communication with the chatbot is inconsistent because it sometimes responds differently to the same message.

These limitations are highlighted in the improvement suggestions: increase comprehension (9) and provide multi-language support.

5) USER EXPERIENCE

User experience refers to the user attitudes towards the interface and user interface experience. We observed that both tools received a lot of both positive and negative feedback in this regard.

The common sub-themes for both tools are ease of use and intuitiveness. Since it earned more positive and fewer negative comments in this respect, Creately is easier to use and more intuitive than SOCIO V1. On the other hand, SOCIO V1 was rated as more fun to use than Creately by 12 experimental participants, while Creately was rated as cumbersome or unmanageable by nine.

Both tools were praised for their wide-ranging capabilities (86 for Creately and 63 for SOCIO V1). Several people who claimed Creately is easy to use also mentioned that it is standard (6) and easy to understand (11). SOCIO V1 scores high for being fun to use (12), reliable (5), and easy to learn (8).

As already mentioned, both tools received a lot of negative feedback as well. Regarding SOCIO V1:

- Users were primarily disappointed because they found the chatbot confusing to use (12) and that it required a lot of learning (16). Since they were unfamiliar with SOCIO V1, they needed to learn how the chatbot worked (how to interact through commands and natural language sentences). Some users (8) found this taxing (''They must know the commands,'' ''You need to learn every function of every command'').

Regarding Creately:

- Some users (27) found the interface control and element management in the window frustrating (''I cannot change the position of the boxes,'' ''The control of the application with the mouse is not easy'').
- 16 subjects specifically stated that it is hard to control the elements adding relationships to link classes (''When joining or associating frames the task becomes a bit complicated,'' ''The arrows chose paths overlapping with other elements'').

When we asked participants for suggestions on how to improve the user experience for these tools, only eight and four people, respectively, suggested improving the ease of use of Creately and SOCIO V1.

6) INTERFACE

In general, Creately outperforms SOCIO V1 as it received more positive and less negative feedback. In particular, Creately was praised for being more visually appealing than SOCIO V1. Regarding interface design, 39 subjects found Creately's interface appealing, emphasizing that it is

''minimalist'' and ''simple.'' Six of them specifically positively rated the visual attractiveness of the diagrams, and six emphasized the color range used. For SOCIO V1, however, only three people referred to the diagram's design in a positive light.

The interface of both tools is suitable for developing UML diagrams since SOCIO V1 received 16 favorable comments and Creately received 21. SOCIO V1 was credited for its command usage and automatic organization of diagram elements (e.g., ''I like that all class diagram modification actions are done under the same command ((\talk),'' ''sort everything automatically''). Participants praised Creately in particular for the method of using a line to directly relate classes (8) and the ease with which diagrams can be exported using a button (7) (e.g., ''It can be easily exported''). Despite the above, some people identified issues that detract from their usefulness (12 for SOCIO V1 and 7 for Creately). For SOCIO V1, for example, they mentioned (i) the continuous use of the \talk command in Telegram or (ii) the existence of too many commands. Task performance is entirely manual in Creately, which detracts from its practicality (e.g., ''Classes and arrows are not reorganized to make it nice,'' ''Everything is written letter by letter'').

The number of suggestions for both tools for this issue (30 for SOCIO V1 and 24 for Creately) was greater than for the other five features. Participants suggested that SOCIO V1's assistance and documentation system might be improved and that Creately should incorporate help. SOCIO V1 features a help page and different responses to user input errors, which 10 participants liked, while 17 thought that the documentation was not complete enough (e.g., ''A more complete manual is missing''). There were also 20 suggestions for improvement, seven of which refer to the addition of further instances (e.g., ''More documentation,'' ''Add more examples of use''). Furthermore, five users recommended introducing predictive support into SOCIO V1 to improve the help system by providing feedback for user input errors (e.g., ''Error messages could be improved by including a hint of where the error might be in the message not understood''). On the other hand, Creately does not include any assistance or documentation, and some participants (8) requested that help be included (e.g., ''Give a tutorial or walkthrough of any tool'').

Because chatbot SOCIO V1 and web-based Creately interface interaction is different, both tools received feedback and suggestions for improving specific aspects of the interfaces. SOCIO V1 uses commands and natural-language statements, whereas Creately adopts drag and drop. With regard to the use of the \undo command in SOCIO, five people recommended that the user be allowed to specify which message to undo. Although 11 people praised Creately for its templates and the ease of diagram customization (e.g., ''Several templates available,'' ''Flexible''), 13 people complained about how hard it is to identify elements in Creately (e.g., ''The components are not easy to find''). Similarly, 10 participants suggested making it easier to manage the interface elements to overcome the control challenge, and six participants suggested including a description of the relationships as only the name is displayed when they are added.

VI. THREATS TO VALIDITY

Although we considered the question of validity during the experimental design phase to assure the validity of the experiment results, we acknowledge that several threats to validity need to be discussed. In this section, we address the main threats to the validity of our family of experiments according to Cooke and Campbell's guidelines [30].

A. CONCLUSION VALIDITY

The first threat to conclusion validity is the limited sample size (29), which may lead to low statistical power. Although we could not recruit a large enough sample size, we did our best to recruit a sample of diverse participants from different countries and regions with different cultural backgrounds, which contributed to the validity of the results. Random subject heterogeneity rules out risk.

To ensure the transparency of the experimental result and encourage the external replication of the experiments, we uploaded the original data and additional analysis in the supplementary material. In the spirit of open science, we uploaded the experimental data and all the material used in this family of experiments to https://dx.doi.org/10.21227/qzdr-nj48.

B. INTERNAL VALIDITY

On the one hand, we acknowledge that students with similar backgrounds to our family from UNIV-1 and UNIV-3 were also recruited in the first family of experiments. In order to avoid learning effects as well as to alleviate threats to the internal validity, we made sure that the same participant only participated once in either the first or second family of experiments.

On the other hand, recognizing that subjects may react differently as time passes, we limited the duration of each session to 30 minutes. We set a 10-minute break between sessions to prevent subject and experimenter fatigue or boredom.

C. CONTRUCT VALIDITY

The first limitation that we observed is English language proficiency. Through the familiarity questionnaire, participants self-assessed their English language level with mean scores of 2.64, 3.11, and 3.98 for EXP1, EXP2, and EXP3, respectively. We observed that they did not express much confidence in their English level, and when they communicated with the chatbot in English, they also used some Spanish words (Spanish is their native language). However, the SOCIO V1 chatbot only supports natural language communication in English; participants had to use English to perform the task. On the other hand, Creately does not require a lot of English communication as it is a visual tool. This may threaten the quality of communication and the experiment results. To reduce this threat, we updated the help pages in

Spanish (the native language of the subjects) and translated our materials and questionnaires into Spanish to reduce the communication effort.

The second limitation to construct validity was social threats. As mentioned before, we were forced to conduct two out of the three experiments remotely (i.e., EXP1 and EXP2) due to the COVID-19 pandemic. The remote experiment may prevent experimenters from solving misunderstandings timely. For example, two members of team 14 in EXP1 experienced network problems at the beginning of the second session of the experiment. They joined the experiment 11 minutes late. This meant that only one team member was working on task performance for the first 11 minutes. This invalidated the participation of this team, as this incident affected the experiment results.

D. EXTERNAL VALIDITY

Threats to external validity may materialize due to the use of students as experimental subjects and the adoption of toy tasks. As is common in SE experiments [6], we employed toy tasks and student subjects to measure the performance of two treatments. In addition, due to the characteristics of chatbots (using UML language), our participants had to be students of computer science or related fields. Although most of the subjects participating in the experiment were final-year computer science students and could be considered representative of novices in industry, the results of the study are applicable to an academic setting and may not be generalizable to industry.

VII. DISCUSSION OF RESULTS

Regarding the results on effectiveness and efficiency, it appears that 62.0% of the subjects tended to take as long as possible to complete and/or improve their class diagrams, while 38.0% of the subjects completed their class diagrams in as short a time as possible, i.e., they completed the task before the 30-minute time limit was up.

Of the abovementioned subjects, 62.0% completed the class diagram for the task that they were set close to the 30-minute time limit. If they had been given longer, they would have used up the allotted time. In this case, the average time taken would have been longer. However, we decided to set a time limit to be able to measure other variables, such as task completion rate.

As both tools were upgraded to varying degrees, neither of the treatments are the same as in the first family [15]. Although comparisons in data terms are meaningless, some differences should be pointed out. Creately's impressive improvements in terms of efficiency and precision referred to quality are due to a significant change: the operating environment upgrade, which no longer relies on Adobe Flash, has resulted in faster page refresh times, more efficient teamwork, and a resulting increase in the effectiveness of class diagram drawings. The changes to SOCIO mainly helped our users to create diagrams, which resulted in the improvement of satisfaction and completeness.

For future chatbots developers and researchers, we provide the following suggestions:

- Prior analysis before designing an experiment is recommended to include (1) a preliminary survey, such as an SMS, to understand the status quo of the research topic, (2) an overview study to understand better the commonalities and differences between the two comparison tools, and (3) a power analysis to reveal the best minimum sample size for acceptable results in the experiments, as well as to highlight the results that differentiate the family result from the baseline experiment result.
- When designing the family of experiments, researchers need to define the essential elements of each experiment in the family and need to control the experimental procedures. We strongly recommend developing a reproduction package to facilitate the examination of temporal results by other experimenters.
- When analyzing the first family, demographic information first needs to be collected. Second, to comprehensively analyze the experimental results, it is necessary to assess them from both quantitative and qualitative aspects. For the first family of experiments, we believe it is necessary to focus more on the results that distinguish this family of experiments from the baseline experiments. The power analysis and cumulative meta-analysis helped to compare the baseline experiments with the whole family of experiments.
- Before developing and analyzing the second family (if needed), researchers may make improvements to the tool based on the first family result. There is no explicit requirement for the second family to continue using the design of the first family. In general, the analysis method of the second family could be different or identical to the one used in the first family, depending on the research goal and changes made to the second family. However, when discussing the results of the second experiment, the differences between the previous work (e.g., the first family) and the current work (e.g., the second family) should be clearly stated, both in terms of experimental design and experimental results.

In addition, we provide a more detailed guide in the supplementary material (https://dx.doi.org/10.21227/qzdr-nj48).

VIII. CONCLUSION AND FUTURE WORK

On the one hand, based on experimental results from previous work [15], (1) we updated the help page for SOCIO by providing more than one language and more examples, (2) we provided alternative context-sensitive help when SOCIO had difficulties understanding the command that the user sent. On the other hand, the control tool Creately was also upgraded replacing Adobe Flash and including an improved interface.

To understand how to improve the usability of chatbots based on evidence, we conducted a family of three experiments with a within-subject crossover experimental design.

A total of 87 participants were recruited and divided into 29 teams. Finally, we reported pooled results, as well as quantitative and qualitative analyses.

In conclusion, we reject the null hypotheses H.2.0 and H.3.0, but not the null hypotheses H.1.0 and H.4.0.

The main results of the analysis of the data gathered in the family of experiments reveal that:

1) *With the observed quantitative results:* SOCIO V1 has better scores for effectiveness and satisfaction than updated Creately. Regarding the efficiency and quality of the class diagram, the difference between the two treatments at family level was not statistically significant.
2) *With the summary of the qualitative results:* SOCIO V1 appears to receive more positive comments and fewer criticisms than Creately regarding contents and interface. Regarding collaboration and communication, both treatments garner more positive than negative feedback. Both treatments receive conflicting feedback on task completion and user experience, with similar numbers of favorable and unfavorable comments.

This family of experiments consolidates the body of knowledge about chatbot usability improvement built on the results of the experiments. We hope our work will provide insights and different perspectives on usability evaluation for SOCIO chatbot and Creately developers.

In this research, we have found that some improvements to be implemented would be: (1) make it easier to remove/edit the elements and (2) improve the natural language ability. In the future, we will develop a second and third updated version (see background) to better understand how the usability of chatbots can be improved based on evidence.

APPENDIX A

Apart from common changes, we created three different versions of the updated the SOCIO chatbot. Here we provide details of the changes that we made to versions 2 and 3.

Updated Version 2 (SOCIO V2): With added functionalities requested by users.

1. Add a /remove command. In response to the first suggestion on modifying the /undo command to be able to undo a participant's action, we developed a new command /remove. After sending the command, users can choose the type of elements (classes, attributes, or relations) that they want to remove. The full list of elements appears, and the user can select the exact element to be removed. For instance, if a user opts to remove a relationship, the chatbot lists all the relationships in the diagram, the user selects a relationship, and the relationship is automatically deleted.
2. Add two commands /undo+ and /redo+. The user can choose how many steps to cancel or redo instead of deleting or redoing one by one. Technically, there is no need to implement anything new inside SOCIO, we merely have to add commands to Telegram with a loop that performs undo or redo as many times as requested by the user. For instance, if a novice user realizes that there is something wrong with the elements he just created, instead of deleting them one by one and creating a new project, he merely has to use the /undo+ command to delete as many steps as he wants.

Updated Version 3 (SOCIO V3): Interface preference. In order to better understand if the change of appearance affects SOCIO chatbot usability, we changed the appearance of generated class diagram using a smaller font; for example, we set the monochrome font on a black background to six.

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, Dec. 2018.

[2] S. Perez-Soler, E. Guerra, J. de Lara, and F. Jurado, "The rise of the (modelling) bots: Towards assisted modelling via social networks," in *Proc. 32nd IEEE/ACM Int. Conf. Automated Softw. Eng. (ASE)*, Oct. 2017, pp. 723–728.

[3] T. Sutikno, L. Handayani, D. Stiawan, M. A. Riyadi, and I. M. I. Subroto, "Whatsapp, viber and telegram: Which is the best for instant messaging?" *Int. J. Electr. Comput. Eng.*, vol. 6, no. 3, pp. 909–914, 2016.

[4] *Systems and Software Engineering Systems and Software Quality Requirements and Evaluation (SQuaRE) System and Software Quality Models*, Standard ISO/IEC 25010, 2011.

[5] N. Juristo and O. S. Gómez, *Replication of Software Engineering Experiments*. Berlin, Germany: Springer, 2012, pp. 60–88.

[6] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering*. Cham, Switzerland: Springer, 2012.

[7] W. F. Tichy, "Should computer scientists experiment more?" *Computer*, vol. 31, no. 5, pp. 32–40, May 1998.

[8] D. T. Lykken, "Statistical significance in psychological research," *Psychol. Bull.*, vol. 70, no. 3, pp. 151–159, 1968.

[9] V. R. Basili, F. Shull, and F. Lanubile, "Building knowledge through families of experiments," *IEEE Trans. Softw. Eng.*, vol. 25, no. 4, pp. 456–473, Jul. 1999.

[10] A. Santos, O. Gomez, and N. Juristo, "Analyzing families of experiments in SE: A systematic mapping study," *IEEE Trans. Softw. Eng.*, vol. 46, no. 5, pp. 566–583, May 2020.

[11] K. Chung, H. Y. Cho, and J. Y. Park, "A chatbot for perinatal women's and partners' obstetric and mental health care: Development and usability evaluation study," *JMIR Med. Informat.*, vol. 9, no. 3, 2021, Art. no. e18607.

[12] L. S. Nowell, J. M. Norris, D. E. White, and N. J. Moules, "Thematic analysis: Striving to meet the trustworthiness criteria," *Int. J. Qualitative Methods*, vol. 16, no. 1, pp. 1–13, 2017.

[13] M. Javadi and K. Zarea, "Understanding thematic analysis and its pitfall," *J. Client Care*, vol. 1, no. 1, pp. 34–40, 2016.

[14] V. Braun and V. Clarke, "Thematic analysis," in *APA Handbook of Research Methods in Psychology* (Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological), vol. 2, H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, and K. J. Sher, Eds. Washington, DC, USA: American Psychological Association, 2012, pp. 57–71.

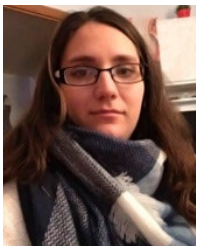
[15] R. Ren, J. W. Castro, A. Santos, O. Dieste, and S. T. Acuna, "Using the SOCIO chatbot for UML modelling: A family of experiments," *IEEE Trans. Softw. Eng.*, early access, Feb. 14, 2022, doi: 10.1109/TSE.2022.3150720.

[16] M. Reeves and J. Zhu, "Moomba—A collaborative environment for supporting distributed extreme programming in global software development," in: *Extreme Programming and Agile Processes in Software Engineering* (Lecture Notes in Computer Science), vol. 3092, J. Eckstein and H. Baumeister, Eds. Berlin, Germany: Springer, 2004, pp. 38–50.

[17] R. Ren, M. Zapata, J. W. Castro, O. Dieste, and S. T. Acuna, "Experimentation for chatbot usability evaluation: A secondary study," *IEEE Access*, vol. 10, pp. 12430–12464, 2022.

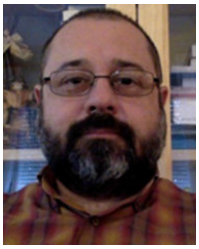
[18] R. R. Divekar, J. O. Kephart, X. Mou, L. Chen, and H. Su, "You talkin'to me? A practical attention-aware embodied agent," in *Human-Computer Interaction (INTERACT)* (Lecture Notes in Computer Science), vol. 11748, D. Lamas, F. Loizides, L. Nacke, H. Petrie, M. Winckler, and P. Zaphiris, Eds. Cham, Switzerland: Springer, 2019, pp. 760–780.
[19] *Ergonomics of Human-System Interaction—Part 11: Usability: Definitions and Concepts*, Standard ISO 9241-11, 2018.
[20] *Systems and Software Engineering Life Cycle Processes Requirements Engineering*, Standard ISO/IEC/IEEE 29148, 2018.
[21] K. Hornbæk, "Current practice in measuring usability: Challenges to usability studies and research," *Int. J. Hum.-Comput. Stud.*, vol. 64, no. 2, pp. 79–102, Feb. 2006.
[22] 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. Eval. Assessment Softw. Eng.*, Apr. 2020, pp. 260–269.
[23] A. Iovine, F. Narducci, M. de Gemmis, and G. Semeraro, "Humanoid robots and conversational recommender systems: A preliminary study," in *Proc. IEEE Conf. Evolving Adapt. Intell. Syst. (EAIS)*, May 2020, pp. 1–7.
[24] T. Fergencs and F. Meier, "Engagement and usability of conversational search—A study of a medical resource center chatbot," in *Diversity, Divergence, Dialogue (iConference)* (Lecture Notes in Computer Science), vol. 12645, K. Toeppe, H. Yan, and S. K. W. Chu, Eds. Cham, Switzerland: Springer, 2021 pp. 328–345.
[25] R. D. Riley, P. C. Lambert, and G. Abo-Zaid, "Meta-analysis of individual participant data: Rationale, conduct, and reporting," *Brit. Med. J.*, vol. 340, pp. 1–7, Feb. 2010.
[26] A. Santos, S. Vegas, M. Oivo, and N. Juristo, "A procedure and guidelines for analyzing groups of software engineering replications," *IEEE Trans. Softw. Eng.*, vol. 47, no. 9, pp. 1742–1763, Sep. 2021.
[27] 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, Feb. 2016.
[28] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Res. Psychol.*, vol. 3, no. 2, pp. 77–101, 2006.
[29] A. Whitehead, *Meta-Analysis of Controlled Clinical Trials*, 1st ed. Hoboken, NJ, USA: Wiley, 2002.
[30] T. D. Cook, D. T. Campbell, and A. Day, *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Boston, MA, USA: Houghton Mifflin, 1979.

RANCI REN received the M.S. degree in ICT research and innovation from the Universidad Autónoma de Madrid (UAM), Spain, in 2019, where she is currently pursuing the Ph.D. degree in software engineering. Her main research interests include experimental software engineering, human–computer interaction, and chatbots. She is a member of the ACM.



SARA PÉREZ-SOLER received the M.S. degree in ICT research and innovation from the Universidad Autónoma de Madrid (UAM), Spain, in 2018, where she is currently pursuing the Ph.D. degree in software engineering. She is also an Assistant Professor at the Computer Science Department, UAM. Her main research interests include chatbots, domain-specific language, and model-driven engineering.

JOHN W. CASTRO received the M.S. and Ph.D. degrees in computer science and telecommunications, specializing in advanced software development from the Universidad Autónoma de Madrid, in 2009 and 2015, respectively. He worked as a Research Assistant at the Universidad Politécnica de Madrid. He is currently an Assistant Professor at the University of Atacama, Chile. He has 15 years of experience in the area of software system development. His research interests include software engineering, software development process, and the integration of usability in the software development process.



OSCAR DIESTE received the B.S. and M.S. degrees in computing from the Universidad da Coruña and the Ph.D. degree from the Universidad de Castilla La Mancha. He is currently a Researcher with the School of Computer Engineering, UPM. He was previously with the University of Colorado at Colorado Springs (as a Fulbright Scholar), the Universidad Complutense de Madrid, and the Universidad Alfonso X El Sabio. His research interests include empirical software engineering and requirements engineering.

SILVIA T. ACUÑA received the Ph.D. degree from the Universidad Politécnica de Madrid, in 2002. She is currently an Associate Professor of software engineering at the Computer Science Department, Universidad Autónoma de Madrid. She coauthored *A Software Process Model Handbook for Incorporating People's Capabilities* (Springer, 2005), and edited *Software Process Modeling* (Springer, 2005) and *New Trends in Software Process Modeling* (World Scientific, 2006). Her research interests include experimental software engineering, software usability, software process modeling, and software team building. She is a member of the IEEE Computer Society and a member of the ACM. She was the Deputy Conference Co-Chair on the Organizing Committee of ICSE 2021.