Requirements Engineering
https://doi.org/10.1007/s00766-020-00334-0

ORIGINAL ARTICLE

SaPeer and ReverseSaPeer: teaching requirements elicitation interviews with role-playing and role reversal

Alessio Ferrari[1] · Paola Spoletini[2] · Muneera Bano[3] · Didar Zowghi[4]

Received: 14 December 2019 / Accepted: 19 May 2020

Abstract
Among the variety of the available requirements elicitation techniques, interviews are the most commonly used. Performing effective interviews is challenging, especially for students and novice analysts, since interviews' success depends largely on soft skills and experience. Despite their diffusion and their challenging nature, when it comes to requirements engineering education and training (REET), limited resources and few well-founded pedagogical approaches are available to allow students to acquire and improve their skills as interviewers. To overcome this limitation, this paper presents two pedagogical approaches, namely SaPeer and ReverseSaPeer. SaPeer uses role-playing, peer review and self-assessment to enable students to experience first-hand the difficulties related to the interviewing process, reflect on their mistakes, and improve their interview skills by practice and analysis. ReverseSaPeer builds on the first approach and includes a role reversal activity in which participants play the role of a customer interviewed by a competent interviewer. We evaluate the effectiveness of SaPeer through a controlled quasi-experiment, which shows that the proposed approach significantly reduces the amount of mistakes made by the participants and that it is perceived as useful and easy by the participants. ReverseSaPeer and the impact of role reversal are analyzed through a thematic analysis of the participant's reflections. The analysis shows that not only the students perceive the analysis as beneficial, but also that they have emotional involvement in learning. This work contributes to the body of knowledge of REET with two methods, quantitative and qualitative evaluated, respectively. Furthermore, we share the pedagogical material used, to enable other educators to apply and possibly tailor the approach.

Keywords Requirements elicitation · Interviews · REET · Peer review · Role-playing · Self-assessment

✉ Paola Spoletini
pspoleti@kennesaw.edu

Alessio Ferrari
alessio.ferrari@isti.cnr.it

Muneera Bano
muneera.bano@deakin.edu.au

Didar Zowghi
didar.zowghi@uts.edu.au

[1] ISTI-CNR: Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Consiglio Nazionale delle Ricerche Area della Ricerca di Pisa, Via Giuseppe Moruzzi, 1, 56127 Pisa, PI, Italy

[2] Department of Software Engineering and Game Development, Kennesaw State University, Building J, Office 375, 1100 South Marietta Pkwy, Marietta, GA 30060, USA

[3] School of Info Technology, Deakin University Burwood Campus/221, Burwood Hwy, Burwood, VIC 3125, Australia

[4] Faculty of Engineering and IT, University of Technology Sydney (UTS), PO Box 123, Broadway, NSW 2007, Australia

1 Introduction

Interviews between a requirements analyst and a customer, as well as other stakeholders such as domain or technical experts, are one of the most commonly used techniques to elicit requirements [2, 23, 35]. The ability of the analyst to gather correct and complete requirements from different stakeholders often depends on the analyst's experience as well as on soft skills [3, 25, 35, 45, 62, 65]. Given the multiple factors influencing the success of elicitation interviews, teaching the art of interviews to software engineering and computer science students, and young analysts in general, is particularly difficult, also due to the limited resources normally available for educational activities specifically focused on requirements engineering (RE) [32, 47].

Role-playing offers experiential learning through the simulation of real-world scenarios; for this reason, it is widely used in disciplines where soft skills and experience are relevant for the success of a task. In RE education and

Published online: 16 July 2020

training (REET), it is a recommended practice to perform role-playing activities [51, 56, 66], in which students can play the role of requirements analysts, to simulate a real-world environment in classroom settings.

Previous work has shown that students tend to commit mistakes in these simulated interviews, and has suggested that the mistakes can be leveraged to give feedback to students and make them improve their interview skills [4, 8, 27]. Other works, mostly outside RE, have shown that active involvement of students in their evaluation, through combination of peer review and self-assessment, increases their learning and understanding through reflections on their experience [12, 14, 48, 59, 60].

Our overarching research goal is *to define and evaluate novel and customizable strategies to teach requirements elicitation interviews*. To this end, this paper combines the ingredients of previous research in REET and in education, in general in two different approaches that build on each other. In particular, the first approach combines existing research on mistakes of student analysts [8, 27], role-playing [56, 66], peer review and self-assessment [12, 14, 48, 59, 60] to propose a novel approach for REET named SaPeer (role-playing, Self-Assessment and PEER review). The approach is specifically focused on improving the skills of students in requirements elicitation interviews. With SaPeer, students receive an initial lecture, followed by a *role-playing* interview experience with a fictional customer. Then, they receive a second lecture in which the typical mistakes of student analysts identified by Bano et al. [8] are listed, together with recommendations to avoid them. Based on the lecture, they are asked to listen to their own interview recording and perform *self-assessment* by evaluating the mistakes committed. Then, they are also required to *peer-review* for mistakes the interview of another student. After this activity, they perform a second interview, which can be also self-assessed and peer-reviewed. At the end of the training, the students are required to reflect on their experience through a feedback questionnaire.

We empirically evaluate the approach through a controlled quasi-experiment. Specifically, we evaluate the reduction in the number of mistakes from the first to the second interview, enabled by SaPeer. The results show that the proposed approach significantly reduces the amount of mistakes made by students. The results also show that different steps of the training may have different effects on specific mistakes, with role-playing being more effective to improve interview planning competences. Feedback from the questionnaire indicates that the steps of SaPeer are considered useful and easy, with the exception of the interview activity. This is considered useful, but also more challenging than the other steps, and students demand more preparation, with an explicit list of *right* questions to ask. Our results also suggest that more corrective feedback is needed along the training to further improve the approach.

The feedback from the evaluation of SaPeer suggests that different variations of SaPeer could be beneficial for the participants and would allow them to experience other aspects of the interviewing process. In particular, since many participants to the quasi-experiment manifested their difficulties in correcting their mistakes in formulating questions and running interviews with the right behavior, the second approach presented in this paper, called ReverseSaPeer (REVERSE role-playing, Self-Assessment and PEER review), includes a reverse role-playing activity in which the participants are interviewed by a trained analyst and experience first-hand the positive impact of being involved in a smoothly and properly run interview. The goal of experimenting with ReverseSaPeer is to understand whether reverse role-playing can be beneficial for the students. In detail, the first part of ReverseSaPeer is identical to SaPeer: students receive an initial lecture, followed by a *role-playing* interview experience with a fictional customer; then, they receive a second lecture in which they are taught the typical mistakes of student analysts and recommendations to avoid them; finally, they use this lecture, to perform *self-assessment* by evaluating the mistakes they committed. Differently from SaPeer, after these activities, they perform a second interview, in which, instead of playing the role of the interviewer the students play the role of a fictional customer interviewed by an experience analyst. The interview is recorded and can be analyzed by the students to review the performance of their interviewee and identify what went differently with respect to their own interview. At the end of this training, the students are required to reflect on this experience and share their reflections in a 500-word essay. We evaluate ReverseSaPeer qualitatively through a thematic analysis of the reflection essays to examine if the students consider it as beneficial and if there are specific benefits that they perceived. This analysis shows that not only the students identify a lot of benefits in participating in ReverseSaPeer, but they are also emotionally involved on learning through it. This is an important discovery since emotional involvement plays a fundamental role in the quality of the students' participation in the activity [11], and engaging students in learning requires consistently positive emotional involvement, which contribute to a classroom climate [43].

This paper is the extension of our previous work presented at Requirements Engineering Conference 2019 in Jeju, South Korea [29], and builds upon REET body of knowledge in general, and the work of Bano et al. [8] in particular. The work extends the original paper with ReverseSaPeer and its evaluation. Specifically, besides the update of introduction, discussion and conclusion, the current work adds Sect. 4 to the original content. The main contributions of this paper are as follows:

- We propose two novel pedagogical approaches, SaPeer and ReverseSaPeer, to teach requirements elicitation interviews through the use of role-playing, self-assessment and peer review. To support the adoption of these techniques, the support material, i.e., slides, lectures, evaluation sheets for self-assessment and peer review, is made available [30].
- We empirically evaluate effectiveness, usefulness and easiness of SaPeer with a quasi-experiment and use the results of our evaluation to reflect to possible variation of the approach.
- We evaluate ReverseSaPeer by analyzing through thematic analysis to understand how the whole approach, and the reverse role-playing activity in particular, is perceived by the students.
- We articulate a discussion on the limits of the approaches and how these limits can be overcome.

The remainder of the paper is structured as follows. In Sect. 2, we present related work and background. In Sect. 3, we describe the SaPeer approach, report the research design for the quasi-experiment we conducted and describe and discuss the results. Analogously, in Sect. 4, we describe the ReverseSaPeer approach, present our research question and research design and describe and discuss the results of our thematic analysis. Section 5 reports observation on the results of both SaPeer and ReverseSaPeer and introduces ideas on how the different components of these approaches could be combined in different ways to address different needs. Conclusion and future work are presented in Sect. 6.

2 Background

In the following, we briefly summarize background work on role-playing, both direct (i.e., the participant plays the role she is training for) and role reversal (i.e., the participants play a role which interacts with the role they are training for, e.g., the customer in the case of interviews), self-assessment and peer review, to provide the context to understand the principles underlying SaPeer and ReverseSaPeer. Then, we focus on existing research on students' analysts mistakes in RE, which is specifically used in our work, and finally we highlight our contribution to REET.

Role-playing Role-playing requires students to play a certain role, e.g., in the context of requirements interviews, the role of requirements analyst, in a simulated scenario. It is based on Dewey's *learning by doing* philosophy [24]. The technique is rooted in Moreno's *psychodrama* method [44] and is largely used for education in several fields, including nursing [20], management [33] and RE [22, 56, 61, 66]. Role-playing has been reported to improve cognitive and affective learning [34] and to be a proper support to train communication skills [28].

In software engineering education, role-playing is used for different objectives [19, 47], such as training students on software modeling and development [5], requirements inspection [61] and requirements elicitation and documentation [22]. The empirical study of Svensson and Regnell [57] has suggested that role-playing can improve students' competences in RE.

In the context of requirements elicitation interviews, with role-playing, students are required to play the role of analysts—and of customers, in case role reversal is applied [66]—in a simulated interview. While playing the role of the analyst, the participants first-hand experience all the difficulties related to the interview process and the required soft skills, while role reversal helps the participants to develop empathy and to understand what might be like to be in the other person's situation [66].

Self-Assessment In self-assessment, also known as self-evaluation [13], students evaluate and possibly grade their own work. Though traditionally self-assessment is not considered part of formal assessment methods in education, it holds a critical role in self-learning processes and to become lifelong learners. Autonomous learning [12, 46], experiential learning [31] or self-directed learning [38] all rely on the self-assessment ability, which requires the students to critically reflect on past knowledge or practice. This has been advocated to enhance students' understanding of the quality of the work and sharpen their critical analysis skills [10]. Self-assessment does not require the students to develop their own benchmarks of quality criteria for their work in isolation, rather it requires the student to analyze their work within commonly shared idea of "good" work. Self-assessment promotes a sense of responsibility on student for their own learning, which is expected of them at tertiary education level, eventually becoming independent from the need for a teacher.

Peer review In our daily lives, we interact with people and learn from them. Analogously, looking at the way in which other people do an activity helps to learn alternative solutions. A structured way of analyzing other people's work is through peer reviews. A peer review consists in evaluating the work or artifacts produced by peers in a certain working or educational environment [14]. Peer reviews in education are based on the principles of *peer learning* theories [15]. The idea behind learning from other people is that they might have been in similar situation to us and might have faced the same challenges in similar contexts [15]. Peer learning is a form of informal and collaborative learning that not just happens outside the classroom environment, but can be utilized effectively within classroom assessments [14]. There are multiple learning outcomes associated with peer learning such as enhancement in social skills, constructive

feedback, reflective learning and articulation of knowledge [14]. Peer reviews fall under collaborative learning pedagogy that are based on cognitive, social and developmental psychology [12, 17, 38]. In the software engineering practice, peer reviews are also largely used to improve the quality of artifacts such as code, requirements (specification documents) [6, 7, 41, 42] and, more recently, interviews [55].

Mistakes of Student Analysts As novices, RE students naturally tend to commit mistakes during requirements elicitation interviews. In an exploratory work, Donati *et al.* [27] identified a first set of 9 general mistake categories. Based on this work, Bano et al. [8] performed a more empirically grounded study involving 110 students divided in 28 groups and collected 34 individual mistake types, belonging to seven classes, namely question formulation (e.g., *asking vague questions*, *technical questions*, *long questions*), question omission (e.g., *not identifying stakeholders*), order of interview (e.g., *no final summary*, *opening with direct questions*), communication skills (e.g., *unnatural dialogue style*), analyst behavior (e.g., *lack of confidence*), customer interaction (e.g., *no rapport*) and planning (e.g., *lack of time management*). In the current work, we will leverage the mistakes from Bano *et al.* to define peer review and self-assessment questionnaires to be used by the students.

Contribution to REET The systematic mapping study presented by Ouhbi et al. [47] on REET shows that very few papers provide full details of the pedagogical design of their RE course or tasks along with evidence of improvement in students learning. From the mapping study, only one study from Connor et al. [21] reported the utilization of *peer learning* theory, though not formally integrated in the curriculum. The lack of studies and evidence on REET suggests that there is a need for proposing and assessing innovative pedagogy to equip graduates with the skills they need in real-world contexts [57].

To our knowledge, this is the first work that proposes pedagogical approaches for teaching requirements elicitation interviews that combine role-playing and role reversal, peer review and self-assessment through a coherent training framework. Furthermore, this work differs from that of Bano et al. [8], in that it provides an *operationalization* of their empirical results, by leveraging the identified mistake types to improve students' interview skills.

3 The SaPeer approach

This section presents the SaPeer pedagogical approach. The fundamental idea of the approach is to first foster experiential learning [31], by letting students perform a role-playing interview, which is recorded, and then stimulate learning through reflection [39, 59], by asking students to identify mistakes in their own interview and in the interview of their peers through questionnaires based on the mistakes identified in [8]. The acquired competence is then tested in a second interview. In the remainder of the section, we present the structure of SaPeer (Sect. 3.1), the research design we followed to evaluate the approach (Sect. 3.2) and its limitations (Sect. 3.3). We conclude by presenting (Sect. 3.4) and discussing the obtained results (Sect. 3.6).

3.1 SaPeer's structure

Figure 1 shows the main building blocks of SaPeer, described below, and how they are organized. All the resources associated with the approach, i.e., lecture slides, videos, questionnaires and product descriptions, are publicly available [30]. The approach can be performed entirely *online*, as we did in our case (and we describe below), or in classroom environments.

1. *Preliminary Training* the students are given a first video lecture of about 20 min on how to conduct interviews, which focuses on positive advice and best practices.
2. *First Interview* each student conducts their first one-to-one Skype interview about a product in a role-playing environment as requirements analyst. Few days before the interview, students are given a description of the product, to prepare interview questions. The role of customer is played by a tutor, and interviews are tape-recorded.
3. *Mistake-based Training* the students are given a second video lecture of 37 min, in which the student analysts' mistakes presented by Bano et al. [8] are described, and examples of erroneous behavior are given for each mistake applicable to interviews conducted online and involving a single analyst (32 out of 34). Specifically, the mistakes *looking at the laptop* and *lack of coordination and choreography* are excluded from the lecture.
4. *Self-assessment* the students are required to listen to their own interview recording and to fill a self-assessment questionnaire. The questionnaire includes 32 statements, one for each mistake type described in the mistake-based training. An example statement is: *I asked vague questions*. For each statement, the student is required to provide a degree of agreement in a 5-point Likert scale—strongly agree (5), agree (4), neutral (3), disagree (2), strongly disagree (1). Therefore, each answer produces a numerical score, which provides a quantitative indication of the occurrence of a certain mistake in the interview, based on the student's opinion.
5. *Peer review* the students are required to listen to the interview recording of another student and to fill a peer review questionnaire. This questionnaire is analogous to the self-assessment one.

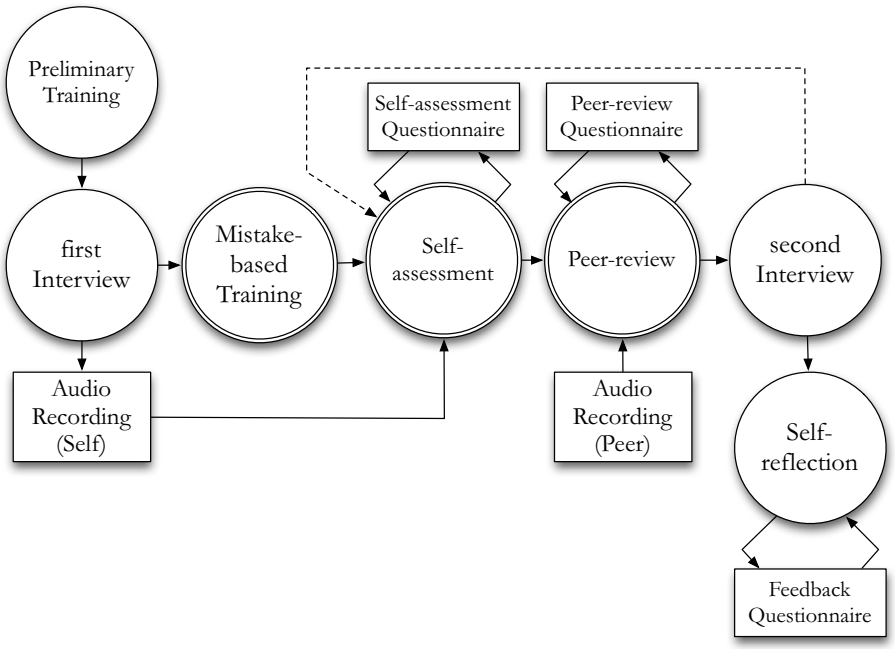

Fig. 1 Overview of the SaPeer approach

6. *Second Interview* students conduct their second Skype interview with a tutor playing the role of customer, but for a different product with respect to the first interview, so that this experience is not biased by the knowledge previously acquired. Also, in this case, students are given a product description to prepare beforehand, and the interview is tape-recorded.
7. *Self-reflection* the students are given a feedback questionnaire, in which they are asked to evaluate the usefulness and easiness of the different steps in the training (i.e., preliminary training, interviews, peer review, self-assessment and mistake-based training) using a 5-point Likert scale and to provide comments on their experience.

The design of SaPeer is modular and can be iterated based on the time available, by, for example, performing self-assessment and peer review of the second interview—dashed line in Fig. 1—or by performing additional interviews. The duration of the interviews can be tailored depending on the time resources available.

3.2 Research design

Our goal is to evaluate the learning effect of the proposed approach when teaching requirements elicitation interviews and to acquire feedback on its usefulness and easiness. To this end, we perform a controlled quasi-experiment [18, 64][1] with an experimental group and a control group. The experimental group adopts the pedagogical approach described in Sect. 3, while the control group skips the steps marked with double lines in Fig. 1 (i.e., steps 3, 4 and 5), therefore performing two interviews one after the other. The idea is both to assess the learning effect of SaPeer as a whole and to check if the effect is mainly associated with the role-playing interview activities—already used in REET, e.g., [8, 9]—or to the other practices introduced in this work (mistake-based training, self-assessment and peer review). In both cases, the results from the second interview are used to understand whether a reduction in the number of mistakes occurred with respect to the first interview. To this end, the scores from the self-assessment and peer review questionnaires are used to evaluate the amount of mistakes in each interview. The members of the control group are also later involved in an activity of self-assessment and peer review, to balance learning objectives and to acquire complementary data for the experiment.

In the following, we outline research questions, context and experimental procedure. Then, we describe the dependent variables and we formalize the hypothesis to be tested to answer the research questions, as well as the validity procedures.

[1] The design is analogous to a randomized trial, but within a sample that could not be selected considering the entire student population. The design could also be regarded as an experiment in case study settings [53].

3.2.1 Research questions

In the experiment, we want to first assess whether the approach leads to a reduction in the number of mistakes from the first to the second interview. Then, we want to check to which extent the reduction in the number of mistakes is influenced by the steps 3, 4 and 5 of the proposed approach. In the following, steps 1 to 6 are collectively referred as the SAPEER treatment.[2] Instead, we refer to the steps followed by the control group, i.e., steps 1, 2 and 6, as the *practice-only* treatment (see Fig. 2, explained later). Finally, we want to understand whether the students consider the different steps of the overall pedagogical approach useful and easy to perform. Therefore, we pose the following research questions (RQs):

- **RQ1:** *Does the* SAPEER *treatment significantly reduce the amount of mistakes made by students in requirements elicitation interviews?* To answer this RQ, we measure the amount of mistakes made by students in the experimental group in the first and second interview, and we assess whether the mistakes are reduced in the second interview.
- **RQ2:** *Is the* SAPEER *treatment significantly more effective than the practice-only treatment in reducing the amount of mistakes?* This RQ aims to assess whether a potential reduction in the number of mistakes in the second interview is due to the steps 3, 4 and 5 of SAPEER, or it is mostly due to experience acquired during the first interview. Answering this RQ requires comparison between the two groups.
- **RQ3:** *Are the steps of the* SAPEER *pedagogical approach considered useful?* The RQ evaluates the opinion of the students in terms of usefulness of each step of the proposed approach. This information is collected with the feedback questionnaire.
- **RQ4:** *Are the steps of the* SAPEER *pedagogical approach considered easy?* The RQ aims to understand whether the steps are considered easy by the students, and which steps are found more challenging. Also, this information is collected with the feedback questionnaire.

3.2.2 Study context

The experiment is conducted in the context of a RE course at Kennesaw State University, GA, USA. The 43 participants of the study are graduate students majoring in software engineering. The Master in Software Engineering (MSSWE)

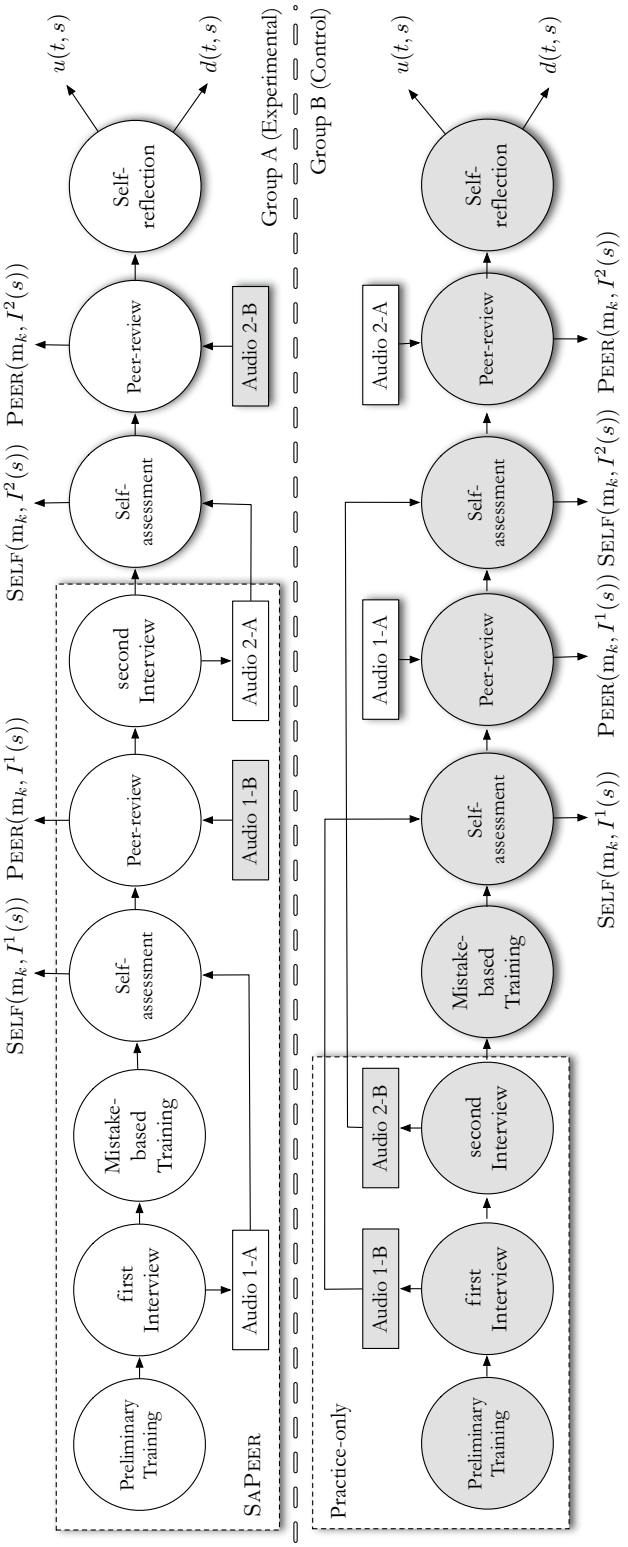

Fig. 2 Overview of the experimental procedure, with the scores collected from the different questionnaires

[2] We distinguish between SAPEER treatment (i.e., steps 1 to 6) and SAPEER approach, which is the general pedagogical approach in Sect. 3.

takes about two years and comprises 36 credits, divided into 12 courses. The MSSWE is offered both online and on campus, and each offered course can be taken in either modality. The students have very heterogeneous background, also because their program admits students who transition into computing from other disciplines. During the activity, the students were all enrolled in a RE course, which is generally taken during the first or second semester in the program. Around 50% of the students in the class had previous, mostly informal, experience with elicitation techniques, and the majority were familiar with the main topics of RE from previous courses.

The activity was part of a module on elicitation techniques offered the fourth and fifth week of class, and the students participated in it as graded part of their course workload. The students had two weeks to perform the whole activity. They also had an additional week to produce user stories about the interviews they performed—this last activity is not part of the current study. Seven tutors were involved in the role of customers for the role-playing activity. Specific countermeasures were taken to prevent ethical issues, as detailed in Sect. 3.3. To give a realistic experience to the students and at the same time keep the activity doable in the assigned time, we have chosen two case studies in which the goal was to develop apps for scheduling services and appointments for two different kinds of business. In the first case study, the interviewer was the owner of a ski resort who needed an app for managing the reservations of all the services offered by the resort. For the second case study, the interviewer was the owner of a hair saloon who wanted to develop an app to manage the saloon's reservations.[3]

3.2.3 Experimental procedure and data collection

Figure 2 summarizes the design of the study, which includes two treatments. The two treatments are SaPeer and *practice-only* and are represented in boxes with dashed lines.

The 43 participants are divided into two groups with a random assignment, group A (Experimental, 21 subjects) and B (Control, 22 subjects). Steps and information associated with group A are in white, while those associated with B are in grey in Fig. 2. Both groups perform the preliminary training activity, and the first interview with a tutor, which was constrained to last 15 min maximum. Then, group A performs mistake-based training, self-assessment on Audio 1-A (i.e., the audio recording of the first interview from group A) and peer review. This step uses the audio recording of the first interview from group B (Audio 1-B). Both groups perform the second interview (max 15 min).

The following activities are then carried out to acquire the data needed to compare the two treatments. Group B performs mistake-based training, self-assessment and peer review, using the recording of the first interviews (Audio 1-B, Audio 1-A). Then, both groups analyze the second interviews, hence self-assessing, and cross-reviewing Audio 2-A and 2-B. The questionnaires filled in all the self-assessment and peer review activities are used as a source of information to evaluate the amount of mistakes made in each interview. In turn, this information will be used to evaluate whether a reduction in the number of mistakes occurred from the first to the second interview (**RQ1** and **RQ2**). Finally, the self-reflection activity is carried out, to estimate the usefulness and easiness of the different steps of the approach according to the students (**RQ3** and **RQ4**).

3.2.4 Dependent variables

The main dependent variables, derived from the RQs, are *amount of mistakes* (RQ1), *effectiveness* (RQ2), *usefulness* (RQ3) and *easiness* (RQ4). Their formal definition is reported below.

Amount of Mistakes Let S^A and S^B be the set of students in group A and group B, respectively. A student participant $s \in \{S^A \cup S^B\}$ performs an interview $I^h(s)$, with $h \in \{1, 2\}$. The index h indicates whether it is a first or second interview. Each interview $I^h(s_j)$ receives two reviews, a self-assessment and a peer review, oriented to evaluate the mistakes. Let $\mathcal{M}$ be the set of 32 mistake types (Sect. 3). Given an interview, for each mistake type $m_k \in \mathcal{M}$, with $k \in \{1 \dots |\mathcal{M}|\}$, we have two mistake scores: $\textsc{Self}(m_k, I^h(s))$ and $\textsc{Peer}(m_k, I^h(s))$ —reported also in Fig. 2 and taking integer values in [1 ...5] according to the 5-point Likert Scale in Sect. 3.1. The amount of mistakes $\hat{M}$ for the single mistake type m_k, interview $I^h(s)$ of student s, is given by the average of the two scores:

$$\hat{M}(m_k, I^h(s)) = \frac{1}{2}(\textsc{Self}(m_k, I^h(s)) + \textsc{Peer}(m_k, I^h(s))).$$

The amount of mistakes M for a certain interview of student s is then given by averaging $\hat{M}(m_k, I^h(s))$ over all the mistake types:

$$M(I^h(s)) = \frac{1}{|\mathcal{M}|} \sum_{k \in \{1 \dots |\mathcal{M}|\}} \hat{M}(m_k, I^h(s)).$$

Both M and $\hat{M}$ take rational values in $[1 \dots 5]$, where higher values indicate higher amount of mistakes.

Effectiveness We define the effectiveness evaluated on a certain student s as the ratio between their amount of mistakes in the first and second interview (ipsative assessment):

[3] The complete description of the case studies (Product Description First Interview and Product Description Second Interview, respectively) can be found in our shared repository [30].

$E(s) = M(I^1(s)) \div M(I^2(s))$.

In the paper, we will also consider the effectiveness for single mistakes, defined as follows. The values of $\hat{M}(m_k, I^1(s))$ and $\hat{M}(m_k, I^2(s))$ indicate the amount of mistakes for the single mistake m_k in the first and second interview, respectively. We define the effectiveness $\hat{E}$ for a single mistake type m_k and student s as the ratio between the mistakes in the first and second interview:

$\hat{E}(m_k, s) = \hat{M}(m_k, I^1(s)) \div \hat{M}(m_k, I^2(s))$.

E and $\hat{E}$ take rational values in $[0.2 \dots 5]$. Values in $[0.2 \dots 1]$ indicate negative or no effectiveness, while higher values indicate increasing positive effectiveness.

Usefulness The usefulness variable is computed for each single type of step of the proposed training. As specified in Sect. 3, the types of steps are $\mathcal{T}$= {Preliminary training, interviews, mistake-based training, peer review, self-assessment}. Given a student s and a type of training step $t \in \mathcal{T}$, the usefulness score for t provided by the student is $u(t, s)$. The variable u takes integer values in $\{1, \dots, 5\}$, where higher values indicate higher usefulness.

Easiness As for usefulness, easiness is defined for each type of step and it is $d(t, s)$, i.e., the easiness score given by $s \in S$ to the type of step t. The variable d takes integer values in $\{1, \dots, 5\}$, where higher values indicate higher easiness.

3.2.5 Analysis procedure and hypotheses

The analysis procedure consists in testing a set of hypotheses derived from the RQs in Sect. 3.2.1. Below, we define the null and alternative hypotheses associated with each RQ, and we indicate the statistical tests used to test them. Parametric tests (e.g., T tests) are used to test them when their applicability conditions are satisfied. Otherwise, nonparametric tests (e.g., Wilcoxon signed rank) are used. All hypotheses are tested for confidence level 95% ($p \leq 0.05$), and we refine them considering single mistakes when appropriate.

RQ1: Does the SaPeer *treatment significantly reduce the amount of mistakes made by students in requirements elicitation interviews?* To answer RQ1, we consider paired samples from group A. Each sample includes the value of M for a certain student $s_i \in S^A$ for the first and in the second interview. More formally, we define $x_i = M(I^1(s_i))$ and $y_i = M(I^2(s_i))$, and our paired samples are $(x_1, y_1), (x_2, y_2), \dots (x_{|S^A|}, y_{|S^A|})$. The null hypothesis is $\mu_\delta \geq 0$, where $\delta = y_i - x_i$, i.e., H_{10} = "the amount of mistakes in the second interview is greater or equal than the amount of mistakes in the first interview." We perform a *one-tail* test, with alternative hypothesis: $\mu_\delta < 0$, i.e., H_{11} = "the amount of mistakes in the second interview is *lower* than the amount of mistakes in the first interview."

We also test sub-hypothesis to focus on *single* mistakes m_k. Also, in this case we have paired samples of $\hat{M}$ values for first and second interview. Given a mistake m_k, the paired samples are (x_i, y_i) where $x_i = \hat{M}(m_k, I^1(s_i))$ and $y_i = \hat{M}(m_k, I^2(s_i))$. The one-tailed null hypothesis is defined as $\mu_\delta \geq 0$ as in the previous case, i.e., $H_{10}^{m_k}$ = "the amount of mistakes of type m_k in the second interview is greater or equal than the amount of mistakes in the first interview." Again, a one-tail test is performed with $\mu_\delta < 0$, i.e., $H_{11}^{m_k}$ = "the amount of mistakes of type m_k in the second interview is *lower* than the amount of mistakes of type m_k in the first interview."

RQ2: Is the SaPeer *treatment significantly more effective than the practice-only treatment in reducing the amount of mistakes?* To answer RQ2, we consider independent samples of the effectiveness variable E from group A and group B. Specifically, we have $E_A = \{E(s_i), i = 1 \dots |S^A|\}$ and $E_B = \{E(s_j), j = 1 \dots |S^B|\}$. The one-tailed null hypothesis is H_{20} = "the effectiveness of SaPeer treatment is lower or equal than the one of the practice-only treatment" (i.e., $\mu_{E_A} \leq \mu_{E_B}$). The one-tail alternative hypothesis that we consider is H_{21} = "the effectiveness of SaPeer treatment is *greater* than the one of the practice-only treatment" ($\mu_{E_A} > \mu_{E_B}$).

As for RQ1, we also consider sub-hypotheses associated with single mistakes m_k. We have independent samples $\hat{E}_A = \{\hat{E}(m_k, s_i), i = 1 \dots |S^A|\}$ and $\hat{E}_B = \{\hat{E}(m_k, s_j), j = 1 \dots |S^B|\}$. The null hypothesis is $H_{20}^{m_k}$ = "the average effectiveness for mistake m_k of the SaPeer treatment is lower or equal than the one of the practice-only treatment" ($\mu_{\hat{E}_A} \leq \mu_{\hat{E}_B}$), and the one-tail alternative is $H_{21}^{m_k}$ = "the effectiveness of the SaPeer treatment for mistake m_k is *greater* than the one of the practice-only treatment" ($\mu_{\hat{E}_A} > \mu_{\hat{E}_B}$).

RQ3: Are the steps of the SaPeer *pedagogical approach considered useful?* This RQ is answered separately for each group, as the groups are applying the steps in a different order, and their judgment may be influenced by that. Hence, given a group of students $S = s_1 \dots s_{|S|}$ and a step of type $t \in \mathcal{T}$, our samples are $u(t, s_i)$, with $i = 1 \dots |S|$. The null hypothesis is H_{30}^t = "the usefulness of the step of type t is lower or equal to the midpoint of the scale, i.e., 3 = Moderately useful" ($\mu_u \leq 3$). The one-tail alternative hypothesis is H_{31}^t = "the usefulness of the step of type t is greater than the midpoint of the scale" ($\mu_u > 3$)—hence leaning toward higher level of usefulness. This evaluation is based on Carver et al. [61].

RQ4: Are steps of the SaPeer *pedagogical approach considered easy?* As for usefulness, for each type of step $t \in \mathcal{T}$ and for each student group S, we have one sample of the easiness variable $d(t, s_i)$ with $i = 1 \dots |S|$. The null hypothesis is H_{40}^t = "the easiness of the step of type t is lower or equal to the midpoint of the scale, i.e., 3 = Neither easy nor

difficult" ($\mu_d \leq 3$), while the one-tail alternative hypothesis is H^t_{41} = "the easiness of the step of type t is greater than the midpoint of the scale" ($\mu_d > 3$)—higher levels of easiness.

3.3 Validity procedure

Construct Validity

The main variable of the study from which the other variables are derived, i.e., the amount of single mistakes $\hat{M}$ (Sect. 3.2.4), has been evaluated through students' scores, which are subjective and may be biased. To mitigate these threats, $\hat{M}$ is computed as average between self-assessment and peer review scores. Furthermore, a tutor not originally involved in the experiment reviewed a sample of 20 interviews, five for each type (1-A, 1-B, 2-A, 2-B, Fig. 2), and assessed them with the peer review questionnaire. The Spearman's rank correlation test between the scores given by the tutor and the average $\hat{M}$ estimated by the students indicates a statistically significant and *medium* correlation, with rho = 0.3129139 and $p = 5.27\text{e}{-16}$. This linear correlation indicates that the $\hat{M}$ values can be regarded as an approximation of the score of the tutor, as our analyses are all based on differences or ratios between scores.

Internal Validity To address problems related to possible imbalance of competence between the groups, the mistakes committed in the first interview can be considered as a *pretest* to assess that the students actually start from the same level of competence, i.e., the same amount of mistakes. To this end, we test the null hypothesis that there is no significant difference between group A and B when considering their average amount of mistakes in the first interview. Formally, let $M_A = \{M(s_i), i = 1 \dots |S^A|\}$ and $M_B = \{M(s_j), j = 1 \dots |S^B|\}$, we define a two-tail null hypothesis $\mu_{M_A} = \mu_{M_B}$. As data are normally distributed (Shapiro–Wilk's test, $W = 0.96396, p = 0.7338$ for group A and $W = 0.97, p = 0.7977$ for group B) and variance is the same for the samples (F test, $F = 0.81545$, num $df = 15$, denom $df = 17, p = 0.6971$), we perform an unpaired, two-sample T test. The null hypothesis is not rejected, as $t = -0.62359, df = 32, p = 0.5373$. Therefore, we can consider that both groups start from approximately the same level of competence. It should be noted that this assessment also addresses *experimental mortality* [18], as the values are based on the sample used to produce the results, i.e., after part of the students retired from the experiment—see values of actual participants in Sect. 3.4. It is worth noticing that we could not entirely address issues related to experimental mortality for what concerns RQ3 and RQ4. As some of the participants did not respond to the self-reflection questionnaire, there is the risk that we collected the opinion of only highly motivated participants.

Another threat of internal validity may be the influence of 7 tutors acting as customers. To have uniform treatments, tutors received common instructions, participated in a 2-h meeting to discuss details of the project, were monitored by the course instructor, and exchanged information through a Slack channel. Furthermore, it was ensured that each student met a different tutor in each interview, so to reduce any bias due to a previous contact.

To prevent *ethical issues* [36, 54], potentially impacting internal validity, the following countermeasures were taken. The steps of the activities were clearly explained upfront and also the general context of the study. The explanation was available in written and video form. Students participated in the activities as part of the class, but the consent to have the data analyzed was collected on a volunteering basis. Students were not graded based on the questionnaires and interviews, but only on the final list of user stories. The forms were administered through the class learning environment in a set of posts with the description of the activity and the different links to the material. The students filled them online using an ID chosen the beginning of the experiment. All the information was collected and analyzed using non-identifiable IDs.

External Validity As a quasi-experiment, external validity is limited, since the opportunistically selected sample comes from a specific course in RE. However, by applying principles of case-based generalization [63], there are architectural aspects of the study that can be used as a term of comparison to generalize the results: Participants are graduate students in software engineering, all the training activities were performed online, and all interviews were *first* interviews with a customer performed by one student analyst. We argue that our results may be applicable for analogous educational contexts.

3.4 Execution and results

The experiment was conducted in September 2018. The students who completed the experiment and produced usable data for RQ1 and RQ2, i.e., peer reviews and self-assessment, are 16 for group A and 18 for group B. Among these students, 12 from A and 10 from B also responded to the feedback questionnaires, hence producing data for RQ3 and RQ4.

3.4.1 RQ1: Mistake reduction

As shown in Fig. 3, the amount of mistakes M is reduced from the first to the second interview for group A. As both samples passed the Shapiro–Wilk's test of normality ($W = 0.96396, p = 0.7338$ for interview 1, and $W = 0.93371, p = 0.2789$ for interview 2), we performed a paired T test to check whether the amount of mistakes in interview 2 is lower than the amount of mistakes in interview 1 (H_{11}). The difference is significant, with

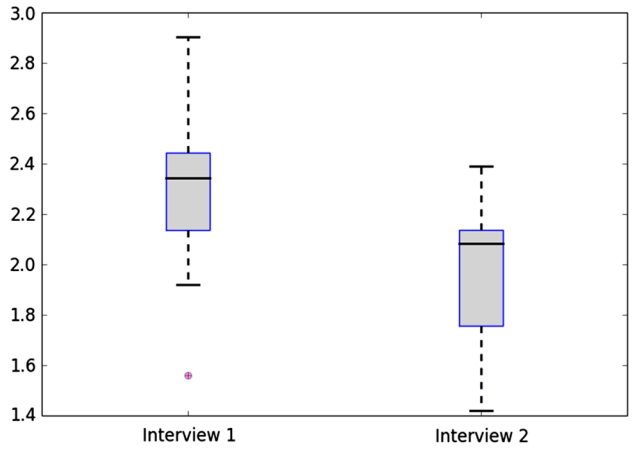

Fig. 3 Amount of mistakes M in first and second interview for group A

$t = -4.7721, df = 15, p = 0.0001235$; hence, H_{10} is rejected in favor of H_{11}. The Cohen's d is -1.037, indicating a *large* effect size.

To understand for which mistakes we had a major improvement in the second interview, it is useful to look at Fig. 4. The figure reports the average over the students of the variable $\hat{M}$ for each type of mistake m_k and compares these values for interview 1 and 2. The darker areas, related to interview 1, can be used as a reference to understand how much improvement—in terms of mistakes reduction—was obtained. For m_k, we used a paired Wilcoxon signed-rank[4] test to check whether the average amount of mistakes in interview 2 is significantly lower than the average amount of mistakes in interview 1 ($H_{11}^{m_k}$). Cases that resulted significant and for which $H_{10}^{m_k}$ can be rejected are marked with * in Fig. 4. We see that there is a general reduction in mistakes for each class, and for each type of mistake. We also see that the most common mistakes in interview 1 are in the classes of question formulation, question omission and order of interview. The major improvement after the training was obtained for the mistake *no final summary*: Suggesting the students to provide a summary at the end of the interview is a quite simple guideline that the students appeared to have followed in interview 2. Similarly, suggesting them to ask for probing questions is another recommendation that was correctly followed (see *no probing questions*). For other cases of frequent mistakes in interview 1, such as *not identifying success criteria*, or *not asking about feature prioritization*, the improvement is notably smaller. These are areas in which the training should be improved, as it appears to have been not sufficiently successful. It is also interesting to notice the improvements obtained in the planning class. In the second interview, the students appeared to have a better time management, better preparation in the domain and better planning. With few exceptions, less improvement was observed on mistakes belonging to communication skills, analyst behavior and customer interaction. These are also the classes in which less mistakes were already committed during the first interview (as the dark area is lower with respect to the other classes).

3.5 RQ2: Effectiveness

The SAPEER treatment appears to be slightly more effective than the practice-only treatment, as shown in Fig. 5. Both samples of effectiveness passed the Shapiro–Wilk's test of normality ($W = 0.95739, p = 0.6148$ for group A, and $W = 0.95284, p = 0.4713$ for group B). Furthermore, the variances of the samples are equal, according to the F test ($F = 1.3787$, num $df = 15$, denom $df = 17, p = 0.5206$). Given that both conditions are satisfied, an unpaired, two-sample T test is performed to assess whether the effectiveness of the SAPEER treatment is greater than the practice-only treatment (H_{21}).

When performing the test, we have $t = 1.4712, df = 32$, and $p = 0.0755$. This indicates that the difference in terms of effectiveness is not significant, and H_{20} cannot be rejected.

It is now useful to compare the effectiveness for the two groups, considering each single mistake. Figure 6 provides a plot of the difference between the average effectiveness for group A and group B, considering each mistake type, i.e., difference between average of $\hat{E}_A$ and average of $\hat{E}_B$ for each m_k according to the definitions in Sect. 3.2. Darker bars indicate higher effectiveness for group A, while white bars indicate higher effectiveness for group B. For each mistake, we performed an unpaired two-sample Wilcoxon test (i.e., a Mann–Whitney test),[5] to check whether the effectiveness of SAPEER treatment is greater than the practice-only treatment[6] ($H_{21}^{m_k}$). Significant cases, for which $H_{20}^{m_k}$ is rejected, are marked with * in Fig. 4. Although most of the cases are not statistically significant, it is useful to discuss the results.

In the majority of the cases, effectiveness is higher for group A, and especially for the mistakes in the class order of interview, in which *no final summary* clearly appears as the mistake in which students of group A improved more with respect to those of group B. Interestingly, there are

[4] We could not apply the T test, as the samples for each mistake did not pass the Shapiro–Wilk's test of normality in most of the cases.

[5] We could not apply the unpaired T test, as the samples for each mistake did not pass the Shapiro–Wilk's test of normality in most of the cases.

[6] For those mistakes in which the practice-only treatment is clearly more effective (white bars in Fig. 6), we performed the same type of test, but to verify whether the effectiveness of SAPEER is significantly *lower* than the practice-only treatment. Results were not significant.

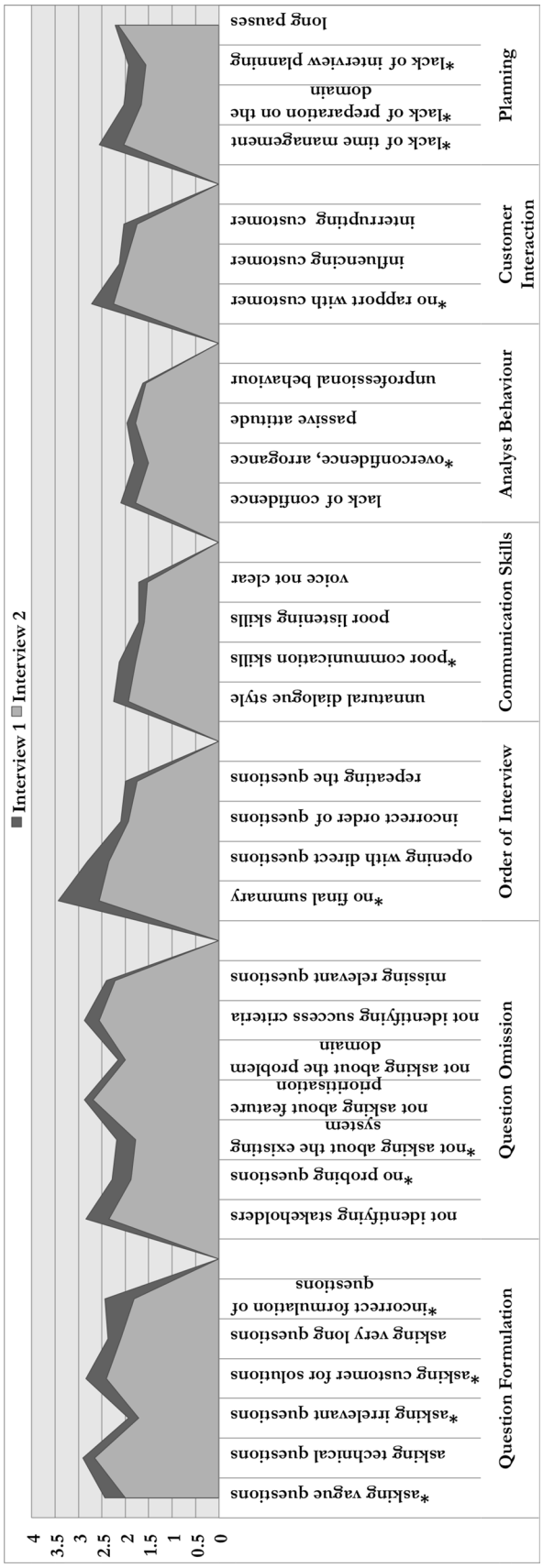

Fig. 4 Average over students of the amount of single mistakes $\hat{M}$ in first and second interview for group A

also cases of mistakes in which group B improved more than group A, such as *asking irrelevant questions*, *missing relevant questions* and *unprofessional behavior*, and most of the mistakes are related to planning. In Sect. 3.4.1, we have shown that planning was a relevant area of improvement already for group A. However, the improvement in average is less than in group B. This suggests that improvement in terms of planning may be mostly due to the actual experience of eliciting requirements during interview 1—in which the students may have directly experienced the consequences of poor planning, as, for example, running out of time—rather than the application of all the steps of the SAPEER treatment.

3.5.1 RQ3: Usefulness

Students were required to evaluate the degree of usefulness of the different steps of the approach. Figure 7 reports the results for group A and group B. Table 1 reports the average of $u(t, s_i)$ over s_i, i.e. the average usefulness rating for step t, denoted as U. For each characteristic, we determine whether the mean response is significantly greater than the midpoint of the scale, i.e., 3 = moderately useful (H^t_{31}), by applying the Wilcoxon signed-rank test. Nonsignificant cases for which H^t_{30} is not rejected are marked in bold.

From Fig. 7, we see that both groups considered most of the steps Moderately to extremely useful, with group A more oriented toward a positive judgment, as none of the respondents selected slightly useful or not at all useful. This happened for group B, in which the students are more negative about the usefulness of the self-assessment and peer review steps. The discrepancy is understandable, as group A performed the steps in the order planned by the SAPEER approach, while group B had to perform multiple review activities, without having the possibility of a second interview after the training. In this sense, group B did not follow the approach, but executed its steps without following the appropriate order, and this is why the usefulness of its steps is less appreciated. It is worth noting, however, that also students from group B appreciated the usefulness of interviews and mistake-based training. These results are evident when looking at Table 1, which shows that while for group A the usefulness score is always significantly higher than moderately useful, this is true for both groups when asked about the interviews and the mistake-based training step.

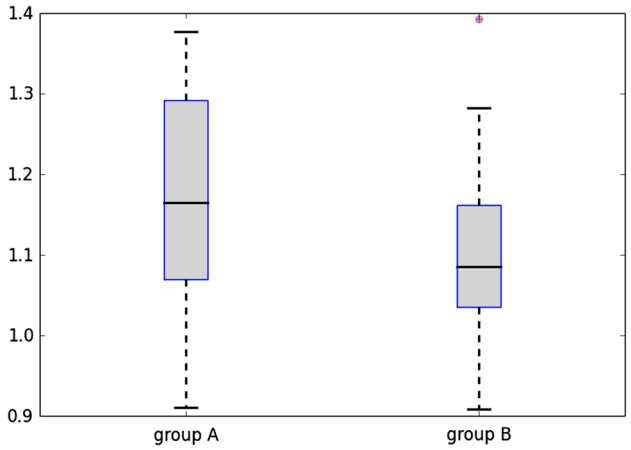

Fig. 5 Effectiveness E of the SAPEER treatment (group A) with respect to the practice-only treatment (group B)

3.5.2 RQ4: Easiness

As for usefulness, students were required to give feedback about the easiness of the steps of the approach. Figure 8 reports descriptive statistics for the two groups, while Table 2 reports the average of $d(t, s_i)$ over s_i, i.e., the average easiness rating for each step (denoted as D), together with the V and p values from the Wilcoxon signed-rank test performed to determine whether the mean response is significantly greater than the midpoint of the scale, i.e., 3 = neither easy nor difficult, (H^t_{41}). Nonsignificant cases (H^t_{40} not rejected) are marked in bold.

From Fig. 8, we see that both groups considered most of the steps neither easy nor difficult to very easy. One exception is the interviews, which have been considered more difficult, especially by group B. This group performed the second interview without the mistake-based training, and this absence of guidance may have been one of the reasons for the increased difficulty with respect to group A. This difficulty with interviews is confirmed by Table 2, in which we see that the average easiness D is 3 for group A and 2.4 for group B. With some differences, also in terms of significance, the other steps of the approach received, in average, a score between 3.5 and 4 (moderately easy).

3.6 Takeaways

The main takeaway messages from our study are:

1. SAPEER enables a reduction in mistakes already from the first to the second interview (Sect. 3.4.1);
2. The steps of SAPEER, and in particular interviews and mistake-based training, are considered useful (Sect. 3.5.1);

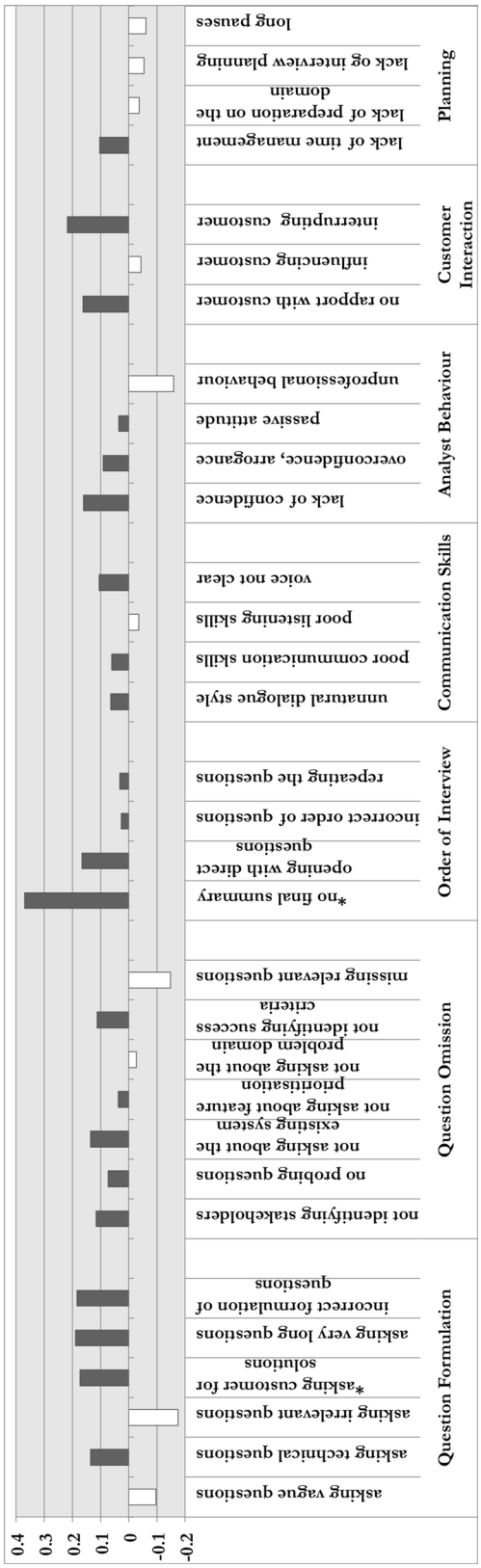

Fig. 6 Difference between group A and B in terms of average of the effectiveness for single mistakes $\hat{E}$

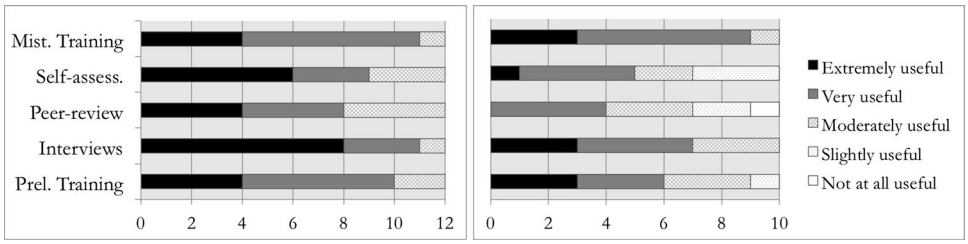

Fig. 7 Results for the usefulness variable for groups A (left) and B (right)

3. Although interviews are considered among the most useful steps, they are also considered as more challenging than the other steps, which are in general evaluated as moderately easy (Sect. 3.5.2);
4. The primary usefulness of interviews is confirmed by the fact that the improvement obtained through the SAPEER treatment is not significantly higher than the improvement with the practice-only treatment (Sect. 3.5).

From Sect. 3.5, we also see that the impact of mistake-based training, peer review and self-assessment in mistake reduction is not significant, except for a few mistakes: *asking customer for solutions* and *no final summary*. These are mistakes with a more well-defined perimeter, which can be corrected with simple recommendations as the ones given in our lectures. Other mistakes are more behavioral and systemic, such as those related to communication skills, analyst behavior and customer interaction. We argue that these mistakes are harder to correct with recommendations and may require more exposure to practice, experience and time. For mistakes related to planning, significant improvement was observed in students following the SAPEER treatment (Fig. 4). However, the improvement was even higher for students following the practice-only treatment (Fig. 6). This suggests that the actual act of *interviewing* may be the one with the highest positive effect for improving the interview planning competences of the students. Therefore, instructors are highly recommended to stress the importance of practicing interviews. On the other hand, it would be useful to have replication studies focused on the planning aspect, to show that the observed difference is also significant.

Overall, these results suggest that the different steps of the SAPEER approach have different impact on specific mistakes. Further research is needed to better understand this diverse impact, thus profiting from the complementarity of the steps.

If time is also crucial, given the results from Sect. 3.5, students can in principle skip the peer review and self-assessment steps, hence focusing on the interview activities. If instead time is not an issue, the process can be extended with further interviews and associated review activities.

Table 1 Average usefulness U and Wilcoxon signed-rank test results

Gr.	Prel. Train.	Interv.	Peer review	Self-assess.	Mist. Train.
A	$U = 4.17$ $V = 55$ $p = 0.002$	$U = 4.58$ $V = 66$ $p = 0.001$	$U = 4$ $V = 36$ $p = 0.006$	$U = 4.25$ $V = 45$ $p = 0.003$	$U = 4.25$ $V = 66$ $p = 0.001$
B	$U = 3.8$ $V = 25.5$ $p = 0.028$	$U = 4$ $V = 28$ $p = 0.010$	$\mathbf{U = 3}$ $\mathbf{V = 14}$ $\mathbf{p = 0.536}$	$\mathbf{U = 3.3}$ $\mathbf{V = 24}$ $\mathbf{p = 0.203}$	$U = 4.2$ $V = 45$ $p = 0.003$

Bold cells represent nonsignificant cases for which $Ht30$ is not rejected

4 The REVERSESAPEER approach

As mentioned in Sect. 3.6, several mistakes in the area of question omission appeared hard to correct with SAPEER. The lack of guidance in asking proper questions emerged also

Table 2 Average easiness D and Wilcoxon signed-rank test results

Gr.	Prel. Train.	Interv.	Peer review	Self-assess.	Mist. Train.
A	$D = 4$ $V = 36$ $p = 0.006$	$\mathbf{D = 3}$ $\mathbf{V = 27.5}$ $\mathbf{p = 0.521}$	$D = 3.7$ $V = 38$ $p = 0.032$	$D = 3.8$ $V = 63$ $p = 0.026$	$D = 3.9$ $V = 45$ $p = 0.003$
B	$\mathbf{D = 3.6}$ $\mathbf{V = 6}$ $\mathbf{p = 0.074}$	$\mathbf{D = 2.4}$ $\mathbf{V = 3.5}$ $\mathbf{p = 0.976}$	$\mathbf{D = 3.6}$ $\mathbf{V = 22}$ $\mathbf{p = 0.096}$	$D = 3.8$ $V = 38$ $p = 0.032$	$\mathbf{D = 3.5}$ $\mathbf{V = 17}$ $\mathbf{p = 0.099}$

Bold cells represent nonsignificant cases for which $Ht30$ is not rejected

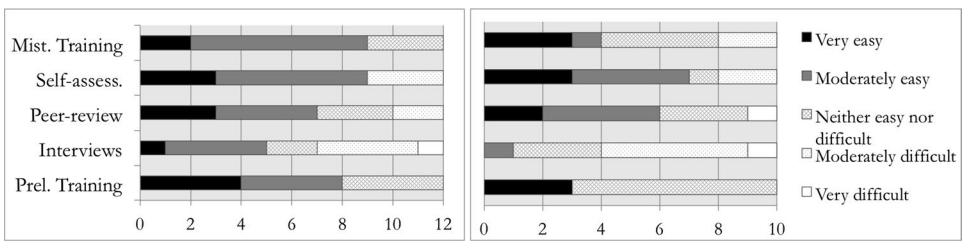

Fig. 8 Results for the easiness variable for groups A (left) and B (right)

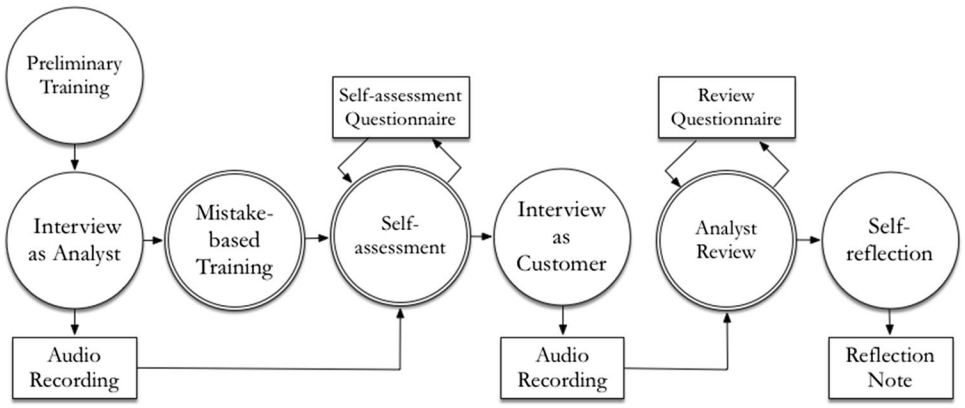

Fig. 9 Overview of the ReverseSaPeer approach

from some students' comments provided through the feedback questionnaire (e.g., "Not having examples of questions, only examples of the types of questions not to ask, it was difficult to formulate question"; "It would be helpful to have a few examples of questions themselves"). This suggests that a list of *right* questions to ask may be beneficial. Questions to start an interview, and to identify missing stakeholders, were suggested by Donati et al. [27]. Other questions can be defined based on the studies of Pitts and Browne on procedural prompts [49], and studies about interviews from other fields such as journalism [1] or social sciences [52]. However, a static list of questions for students to study does not show how to use them in the appropriate context and how to create the correct flow and climate. Furthermore, in Sect. 3.6 we observed that several mistakes in the area of analyst behavior might require more guidance than just recommendations to be corrected. Given the relevant impact of the role-playing activities in SaPeer, we hypothesize that the use of *role reversal* [66], i.e., students acting as customers and interviewed by an experienced analyst, could overcome these limitations. Experiencing the other side of an interview and seeing an example of *right* questions in context could be more effective to correct the behavioral mistakes and to help students in improving their interviewing skills. It is worth noting that ReverseSaPeer does not aim at systematically overcoming all the limitations of SaPeer. Instead, it is designed to address the legitimate need of experiencing a well-performed interview and explore to which extent this is considered beneficial. To this end, feedback will be collected to understand the learning benefits observed by the students.

Following this intuition, we propose the ReverseSaPeer pedagogical approach. ReverseSaPeer builds on SaPeer and modifies it by substituting the second interview with a role reversal activity. Moreover, to manage the student's time efficiently, given the results from Sect. 3.5 and considering that the participants will still review an interview conducted by others while reviewing the role reversal activity, the peer review of the first interview is removed.

In the remainder of the section, we first present the components and the structure of ReverseSaPeer (Sect. 4.1), we then present our research question and research design (Sect. 4.2) and the results of our analysis (Sect. 4.3). We conclude with summarizing the takeaway messages (Sect. 4.4) and with the discussion of threats to validity (Sect. 4.5).

4.1 ReverseSaPeer's structure

Figure 9 shows the main building blocks of ReverseSaPeer, many of which are in common with SaPeer. In particular, as in SaPeer, students participate in the following activities:

1. The vision of a 20-min *Preliminary Training* on how to conduct interviews;
2. An *Interview as Analyst* in which each student plays the role of requirements analyst in the same settings used in SaPeer;
3. The vision of a 37 min *Mistake-based Training* which presents the mistake types commonly done by novice analysts;
4. A *Self-assessment* performed by listening to the recording of their own interview and filling a self-assessment questionnaire.

After the self-assessment, in order to help students to correct behavioral mistakes, ReverseSaPeer substitutes the peer review and the second interview as analyst (part of SaPeer) with a role reversal activity. In particular, the remaining steps of ReverseSaPeer are:

5. *Interview as Customer* students conduct their second interview with a research student assistant who has been trained as requirements analyst. In this interview, students play the role of the *customer*. As for the case of the first interview, students are given a product description to prepare beforehand, and the interview is tape-recorded.

6. *Analyst Review* the students are required to listen to the recording of the interview conducted by the graduate research assistant and fill out the review questionnaire. This questionnaire is analogous to the self-assessment one, except for the formulation of the statements, which in this case, as in the case of the peer review questionnaire in SaPeer, are in third person (*The analyst asked vague questions*, etc.).
7. *Self-reflection* the students are required to reflect on the ReverseSaPeer experience and all its phases and to write up their reflection in an essay of maximum 500 words.

The material used for the training and the case studies were the same used in SaPeer, and, also in this case, the approach can be performed entirely online or in the classroom.

4.2 Research design

SaPeer's evaluation has shown that the role-playing activity is the one that mostly affects the improvement in student's performance in conducting interviews. However, despite the improvements obtained in the second interview, students still complained the lack of training in asking correct questions and avoiding the mistakes committed in the first interview. In ReverseSaPeer, students have the possibility to learn through experiencing best practices to conduct interviews and the qualities of good analysts. For this reason, while evaluating ReverseSaPeer, we are mainly interested in learning which is the students' perception of the role reversal activity. Formally, our research question is:

RQ5: *What are the benefits and challenges of* ReverseSaPeer *from the viewpoint of the students?* The RQ evaluates the opinion of the students on ReverseSaPeer by analyzing their input and grouping it in different emerging relevant concepts.

The students' opinions have been collected in the context of a RE course at Kennesaw State University, GA, USA. Analogously to the 43 participants in the SaPeer's quasi-experiment, the 41 participants in this study are graduate students majoring in software engineering and they come from very heterogeneous backgrounds. However, all the participants were introduced to the main topics of RE as part of their previous education. Moreover, in this edition of the RE course, almost 40% of the students had a direct experience in using elicitation interviews.

The considered RE course is the same course (offered the subsequent academic year) in which the students who use the quasi-experiment for evaluating SaPeer were enrolled. ReverseSaPeer was part of a graded activity in a module on elicitation techniques that offered the fourth and fifth week of class. Four tutors were involved in the role of customers for the first interview. The tutors have mixed background and were not trained to be requirements analysts, but they prepare together on the topic of the interview to have a consistent preparation and offer an analogous experience to all the participants. The interviewer in the role reversal activity was instead a graduate research assistant, trained to conduct and analyze interviews.

All the participants in the course performed the activities in ReverseSaPeer, and, for each participant, the following data were produced: audio of both the role-playing activities, self-assessment of the first interview and review of the second interview, and a final self-reflection on the approach and its activities. The collected material is part of the pedagogical method, and it is functional to meet the activity learning outcomes. In addition, the final self-reflection notes are collected to answer to **RQ5** by means of thematic analysis [16].

The students were asked to provide their opinions based on their experience on whether they improved their learning through ReverseSaPeer and what benefits and challenges they encountered during the process. Two researchers (third and fourth authors) carried out the thematic analysis of the self-reflection notes. They sat together to read them all and analyzed the emerged themes that indicate learning, benefits and challenges that students faced during the overall ReverseSaPeer experience. The classification presented in Fig. 10 arose from two rounds of thematic analysis. In the first round of coding, the researchers coded the quotes from students that referred to benefits/challenges but also indicated whether students have expressed positive or negative feelings about ReverseSaPeer, which later emerged as a separate theme regarding the "Affects" [65]. Students also expressed their opinions on whether ReverseSaPeer increased their learning of requirements elicitation interviews or not, and hence the authors coded those quotes separately for the theme of "Learning." In the second round, the authors merged all positive experiences into "Benefits" and difficulties expressed by students into "Challenges," along with the themes of "Learning" and "Affects" of ReverseSaPeer. "Benefits" theme had 9 codes, "Challenges" had 16 codes, and 7 codes were in "Affect" theme. The "Learning" theme was analyzed to see whether students agreed that ReverSaPeer helped in their learning, and hence this theme had only two codes "Positive learning experience" and "Irrelevant learning experience." In the Results section, we will be providing some of the quotes from student reflections in relation to these themes.

4.3 Results

Figure 10 schematizes the main themes emerged during the thematic analysis grouped in four main categories, which highlight the benefits and the challenges perceived by the students while participating in ReverseSaPeer, in general,

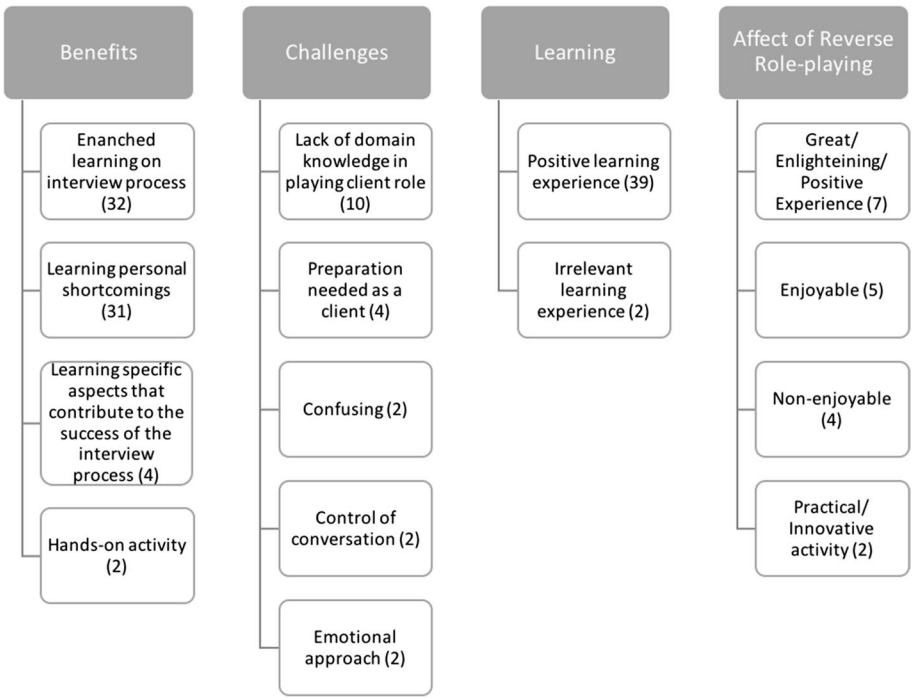

Fig. 10 Overview of the themes emerged from the reflections on the ReverseSaPeer approach

and in the role reversal activity, in particular, and the emotions the students specifically had about the experience.

4.3.1 Benefits

The majority of the students (over 85%) identified benefits related to ReverseSaPeer in their reflections. All the extracted themes provide an answer to **RQ5**.

The most emerging theme was that the approach *enhanced learning on interview process*. This emerged in 78% of the essays that listed whose specific aspects of conducting interviews were enhanced by participating in ReverseSaPeer, but also made general considerations about the benefits of the overall learning experience. In the notes, students highlighted as both the role-playing activities represented, in different ways, enhanced the learning of the interview process. A recurrent theme about the first interview is playing the role of the interviewer help to understand the skills needed in the overall process. For example, one of the students highlights that, when she was the interviewer, she "learned immediately how important it is to be flexible in an interview." One of the students mentioned that he "learned from this experience conducting the interview as the interviewer that without proper planning, an interview can be very difficult and costly."

The role reversal activities were mainly perceived as an enhanced learning process to understand how to correctly conduct an interview. For example, a student highlighted that "[His] interview for the Cool Ski Resorts definitely gave [him] a great example of how to properly conduct an interview." Analogously, another student said that "the second interview served as a good way of seeing how a good interview should be held."

Another benefit of ReverseSaPeer, emerged in more than 75% of the notes, is that it allows students to learn personal shortcomings. Being an interviewer "gave a very good perspective of just how unprepared [he] was for conducting interviews" is a recurrent comment in the notes. Sometimes, students refer to specific shortcomings such as the incapability of promptly react in the conversation (e.g., "I found myself being silent some few times, while trying to orientate the conversation …," "There were a few periods of awkward silence when …"), insufficient preparation (e.g., "As I started the interview, I quickly realized that I did not know what I was doing"), and inadequate questions (e.g., "I did not have any questions to really understand the full scope of the problem," "I did not know what to ask to understand more about the need of my customer after asking three or four questions").

Notice that students realized their shortcomings also during the self-assessment or the role reversal activity. For example, one of the students, even if he put effort in not using jargon, during the self-assessment "did notice that [he] made some technical questions unconsciously, and even though [he] did manage the conversation to flow, [he] missed some important questions." Analogously, being part of a well-conducted interview made some students aware of the limitations of their performance (e.g., "It was not until the second interview, where I was the interviewee, that I understood how important it is to make it seem as if it is a

conversation and as such, to let it flow") and made them wish to be able to "go back to do better in first interview and ask better questions like the graduate research assistant did."

Another (less frequent worth mentioning) benefits emerged in the analysis of the students essays is that ReverseSaPeer is a *hands-on activity* (e.g., "My experience as an interviewee was more realistic and good practical exposure on requirement elicitation process"); and it helps to *learn specific aspects* that contribute to the success of the interview process, otherwise difficult to grasp such as

- the critical role of effective listening in probing questions (and of "digging deeper on responses with probing questions");
- the effectiveness of teamwork in interview ("my first thought after the interview was that it would have been very beneficial to have a team member to help me conduct the interview.");
- the importance of having a respectful relationship with the client (e.g., from a comment in one of the essays about the role reversal activity: "The analyst also showed a great deal of care and respect for my time, making sure not to let the interview drag on without end.");
- the importance of the order of the interview (e.g., "This experience taught me that having my questions in order before the interview will make it go smoother");
- the importance of the client's correct answers (e.g., "both interviewer and interviewee need to put same amount of effort to get correct system built").

4.3.2 Challenges

Besides the benefits, ReverseSaPeer presents also some challenges. The analysis highlighted that 56% of the students identified at least one challenge in participating in the proposed activities.

Three of these challenges are related to some minor design choices of the approach and, so, can be used to improve it. The most commonly encountered challenge, identified in 25% of the cases, is the *lack of domain knowledge in playing client role*. Some students "felt a little overwhelmed trying to put together a backstory," others "felt slightly uncomfortable," because they "had to make stuff up." Being interviewed has been perceived problematic also for other reasons (still partially related to the lack of domain knowledge) by around 12% of the students. Some found it *confusing* (e.g., "all those questions can become confusing or perhaps intimidating") and others did *not feel prepared as a customer* and "felt a little overwhelmed trying to put together a backstory for Jim/Mary."

As it happened for SaPeer, while role-playing is considered a positive experience by the majority of the students, it is also considered *complex*, especially when students play the role of the interviewer. Indeed, there are many aspects that even with extensive preparation are difficult "on the spot" (e.g., "responsiveness" and "simultaneously taking notes and being engaging" for the customer). Also, acting as a customer can be challenging since the *control of the interview* is in the hands of the analysts ("I think I enjoyed interviewing more than being interviewed since I had the control of the structure of the conversation").

Other challenges that emerged in the analysis are related to the participants' *emotional approach* to the interview. Because of nervousness, a student felt that he needed "to run more practice interview sessions" to better exploit the experience, while overconfidence "made [some student's] performance suffers."

The emerged themes related to challenges add information to the answer to **RQ5**. Indeed, they show that while many benefits are perceived by the students, ReverseSaPeer presents also some challenges that compromise these benefits, especially the challenges related to the lack of domain knowledge in the role reversal activity. Luckily, this challenge can be mitigated by offering the students multiple alternatives of project so that they can choose the one that they are more comfortable with, and by providing more information about the domain. Giving the students the possibility to choose their own process would also solve this problem, but would require them to create the one-page description needed by the analyst to prepare for the interview and would add additional workload for the analyst who will need to do many preparations.

4.3.3 Learning

A theme that emerged in all the notes is the impact on learning. Only 2 of 41 participants (less than 5%) did not identify ReverseSaPeer as a positive learning experience. Between the two, one did not give any connotation to the experience, while the other student was fairly negative about ReverseSaPeer ("it did not really add much to my knowledge"), but also constructively analyzed the problems ("The second interview felt too far removed from my experience to provide much at all") and what could have helped him ("I think it would have been provided more insight if we were able to see a review from the interviewee on how we did").

The remaining students were all positive on the learning experience ("I was not expecting this much learning"), and each essay mentioned on average a couple of specific skills that they learnt from ReverseSaPeer(e.g., "how to ask proper questions in a logical order," "not jump into the technical questions," "communication skills").

This emerged result adds to the answer to **RQ5**: Students perceive ReverseSaPeer as a positive learning experience.

4.3.4 "Affect"

Half of the students' essays included a description of the emotions that the students had in participating in ReverseSaPeer and, among them, almost 80% are positive.

Students *enjoyed* the experience and consider it *good* ("It was a great opportunity which increased my knowledge and experience in the interview field"). For another student, "it was a very *enlightening experience*, *and I am sure the interviewing part of elicitation takes many years to master as a requirements analyst"*. Other positive emerged affects are *fantastic*, *great*, or *interesting* experience. In particular, for one of the student "the two interviews [were] very helpful and interesting to do." Some of the positive effects were focused on the reverse role activity and the flow of the approach ("I am really glad I got to be the interviewee the second time so that I could see what kind of things I needed to improve on. Another If they took placed reversely, I might just imitate the graduate research assistant in the second interview and not realize how important it is."

This adds to the response to **RQ5** by highlighting the positive effects and the emotional involvement in the learning process that students perceived while taking part in ReverseSaPeer.

4.4 Takeways

The main takeaway messages from our study are:

1. The themes that emerged from the majority of the essays are related to benefits of participating in ReverseSaPeer.
2. The main perceived benefits are that ReverseSaPeer provides enhanced learning on interview process and helps to learn personal shortcomings. These benefits are perceived as a consequence of both role-playing activities and self-assessment. It is important to note that these benefits cannot be experienced just going through a list of correct questions. This confirms that adding the role reversal activity is a more effective solution to the problems related to not knowing what to ask to avoid mistakes.
3. Another important benefit that emerged by the analysis is that ReverseSaPeer is strongly perceived as a positive learning experience. This reinforces the intuition that proposing an approach based on active learning is perceived as effective by the students.
4. Positive effects and emotional involvement are experienced by the students while participating in ReverseSaPeer. This means that students not only learned, but had a pleasant and somewhat surprising experience, profiting from all the benefits of learning by "playing."
5. The analysis of the essay indicates that students were perceived as challenging being interviewed because of the lack of domain knowledge. This has a negative impact on the benefits of ReverseSaPeer, and indicates that not only acting as analyst requires preparation, but also acting as customer. Luckily, this problem can be mitigated as explained in Sect. 4.3.

Overall, ReverseSaPeer is perceived as beneficial by students participating in it. Different benefits were identified, and, among them, some suggested that the role reversal activity is a key element in the approach.

4.5 Threats to validity

Reliability The essence of reliability for qualitative research lies with consistency and verifiability of the findings. Despite the inherent subjectivity of any qualitative study, as shown in Sect. 4.2, we have analyzed the students' reflections through a systematic and rigorous procedure, which merges the analysis of two experts. Moreover, we share our data,[7] our derived themes, as well as a large amount of representative fragments for the themes. Finally, the students were provided clear instructions of what include in their reflections, which are hence, in general, comparable in terms of quality and informative content. Therefore, we believe that the link between the data and the findings is sufficiently reliable.

Validity The research design was oriented to gather feedback from the students. While students were not evaluated for the reflection documents, they were evaluated for the activity that was the subject of the reflection. This might have affected their feedback, which is the source of our conclusions. To mitigate this threat, which could not be entirely avoided given the constraints of the course in which the study was conducted, we clearly stated that their opinions would have not been part of the evaluation, so they were invited to be honest. As for the data extraction from the reflections, two researchers participated to it, and this form of triangulation further supports results validity.

Generalizability The results can be applicable to similar contexts of English-speaking classrooms of graduate students in computing disciplines who have already some academic background in software development (e.g., an introductory course on software engineering). Other results may emerge if different contexts are considered.

5 Discussion

We organize our discussion by first relating our results to existing work in REET, and we then outline ideas for improving and tailoring the approach to different classroom environments.

[7] The students' reflection are made available in our repository [30].

Results in Relation to Education Literature

The current work confirms the utility of role-playing in education in general [28, 34] and REET in particular [22, 57, 66] and, to our knowledge, is the first work that empirically shows that role-playing helps to improve interviewing skills in RE. Results about the *usefulness* of peer review and self-assessment activities are partially in line with the literature on these educational practices, as students appear to have had a positive learning experience, possibly thanks to their involvement in the assessment process [48, 60].

Our results also confirm that playing the reverse role in interviews is a useful learning experience [66]. Experiencing the "other side" of an activity helps to both understand the customer's perceptions during the interview and thus better relate to her, and to observe the correct behavior of an analyst and learn from it. To our knowledge, this work is the first attempt to systematically analyze how role-playing activities both in the main and in the reverse role are perceived by participants. The results of our analysis show that students perceive it as positive and beneficial.

However, the effect of self-assessment and peer review practices, although positive (Fig. 5), is not statistically significant for what concerns mistake reduction. Given that these are well-established practices [26, 60], with a long history found on philosophical and pedagogical theories of constructivism and community learning [37], further experimentation is needed, possibly based on an improved version of the approach.

Our analysis also shows difficulties in dealing with domain knowledge while acting as interviewee. Even if this perception comes from the reverse role activity, it confirms the importance of domain knowledge in the interview process and has the students experiences how the lack of it creates issue in the conversation.

Finally, from the analysis of ReverseSaPeer emerged that students are emotionally involved in the learning process when participating in it. This is something that was not possible to observe through the feedback questionnaire used in SaPeer since it did not collect any information about this aspect, but this consideration applies also to it. This is an important result because emotional involvement positively impacts on the quality of the students' participation [11] and contributes to engage students and to improve the classroom climate [43]. Also, this result suggests that approaches analogous to ReverseSaPeer (and SaPeer) could be used to teach other complex-to-teach software engineering topics (e.g., project management) of which the success is influenced by many factors, including soft skill.

Improving the Approach As mentioned, some behavioral mistakes are hard to correct through recommendations. However, we have seen that students did not significantly improve on several mistakes related to the area of question omission (see Fig. 4, the $*$ symbol marks significance), for which suggestions can be provided. This shows that students need further guidance of this aspect. ReverseSaPeer is our first attempt to address this limitation. However, in the approach we did not consider the problems connected with the lack of domain knowledge of the students while performing the reverse role activity. Moreover, we choose to have students learning from a positive role reversal activity with a trained interviewer and neglect what students could learn from a negative reverse role activity during which they could first-hand experience the impact of mistake on the interviewee's attitude.

Another aspect to improve in the approaches would be to provide students feedback on their first interview, e.g., by giving them the results of their peer review questionnaire. This was not possible in the context of the study due to timing issues—first interviews for group A were reviewed after second ones were performed—and feedback from peers may have also some drawbacks as recently noticed by To and Panadero [58]. However, we argue that this form of corrective feedback, possibly complemented by tutor's feedback, may be particularly helpful, as also suggested by some students for both SaPeer and ReverseSaPeer (e.g., "I would replace the self-assessment or at least add an assessment from the professor"; "Getting the feedback from the first interview before doing the second may have helped").

Tailoring the Approach SaPeer and ReverseSaPeer are designed to be modular and adaptable, and, although the steps should be preferably performed in the recommended order to prevent difficulties (see Sect. 3.5.2), teaching contexts may vary in number of students and resources, hence requiring adaptation of SaPeer. Specifically, in case scale is a major issue, students can conduct interviews in groups. If tutors are not sufficient to handle all the students, *role reversal* [66], with students acting as customers, can be applied. Notice that this presents the limitation that different students will experience different quality activities because of the different attitude and preparation of the students.

Furthermore, if time is also crucial, given the results from Sect. 3.5, students can in principle skip the peer review and self-assessment steps, hence focusing on the interview activities. A shorter version of the approach could also compact the experience including only the two training and the first interview. If instead time is not an issue, the process can be extended with further interviews, and associated review activities. SaPeer and ReverseSaPeer are also specifically oriented to novices, with pre-defined questionnaires for peer review and self-assessment. Experienced learners may be expected to design the criteria or rubric for assessment themselves [50].

6 Conclusion and future work

This paper presents and evaluates SaPeer and ReverseSaPeer, two novel pedagogical approaches, for teaching requirements elicitation interviews. The approaches follow the active learning teaching modality [40] and are based on role-playing, reverse role-playing, peer review and self-assessment, and leverage previous research on mistakes of student analysts in RE [8, 27]. The material developed for both the approaches can be used to deliver them either face-to-face or online. Both teaching modalities provide students the same experience and should be used in accordance with the students familiarity to the modalities. The produced material is explicitly developed for students majoring in computing-related fields because it takes into account their specific background, the proposed interviews have the goal of eliciting requirements for new technological solutions, and the considered mistakes have been extracted analyzing requirements elicitation interviews.

The quasi-experiment conducted to assess SaPeer shows that students following the approach significantly reduce the amount of mistakes made. Major reductions are observed for mistakes that can be corrected with well-defined actions, such as providing a summary at the end of the interview, or asking probing questions. Mistakes more related to behavioral aspects are harder to correct, and some mistakes in the area of question omission are not correctly addressed at the moment. Furthermore, we also observed that the control group (group B in our experiments), who performed two interviews in a row, was also able to reduce part of the mistakes in the second interview. This confirms the intuition that the actual practice of interviewing, even in a role-playing context, may be the crucial one to improve students' interview skills.

The analysis of ReverseSaPeer, introduced to both address the problem of asking the correct questions and have the students learn from a well-conducted interview, shows that the overall approach is perceived as beneficial and helpful.

Future work will focus on further improvement and dissemination of SaPeer and ReverseSaPeer. We plan to include suggestions of possible example questions to ask, to address problems of question omission, as well as corrective feedback activities, which are lacking in the current approach. Moreover, we plan to consider the effect of a negative reverse role experience with an interviewer who commits mistakes in the interview. Experiments will be performed to assess the effectiveness of the modified approaches and to better understand the relationship between the steps of the training and the reduction in specific types of mistakes. We also plan to create off-the-shelves modules (one for each activity) with recommendations on how to combine them depending on the available resources, the set learning outcomes, and the audience.

Acknowledgements This work was partially supported by the National Science Foundation under grant CCF-1718377.

References

1. Adams S (2001) Interviewing for journalists. Psychology Press, London
2. Agarwal R, Tanniru MR (1990) Knowledge acquisition using structured interviewing: an empirical investigation. JMIS 7(1):123–140
3. Aranda AM, Dieste O, Juristo N (2016) Effect of domain knowledge on elicitation effectiveness: an internally replicated controlled experiment. TSE 42(5):427–451
4. Argyris C, Schon DA (1974) Theory in practice: increasing professional effectiveness. Jossey-Bass, San Francisco
5. Auriol G, Baron C, Fourniols JY (2008) Teaching requirements skills within the context of a physical engineering project. In: REET 2008. IEEE, pp 6–11
6. Aurum A, Petersson H, Wohlin C (2002) State-of-the-art: software inspections after 25 years. Softw Test Verif Reliab 12(3):133–154
7. Bacchelli A, Bird C (2013) Expectations, outcomes, and challenges of modern code review. In: ICSE'13. IEEE, pp 712–721
8. Bano M, Zowghi D, Ferrari A, Spoletini P, Donati B (2018) Learning from mistakes: an empirical study of elicitation interviews performed by novices. In: 2018 IEEE 26th international requirements engineering conference (RE). IEEE, pp 182–193
9. Bano M, Zowghi D, Ferrari A, Spoletini P, Donati B (2019) Teaching requirements elicitation interviews: an empirical study of learning from mistakes. Requirements Eng 24:259–289. https://doi.org/10.1007/s00766-019-00313-0
10. Black P, Wiliam D (1998) Assessment and classroom learning. Assess Educ Princ Policy Pract 5(1):7–74
11. Boekaerts M (2010) The crucial role of motivation and emotion in classroom learning. In: The nature of learning: using research to inspire practice, pp 91–111. https://doi.org/10.1787/9789264086487-6-en
12. Boud D (2012) Developing student autonomy in learning. Routledge, London
13. Boud D (2013) Enhancing learning through self-assessment. Routledge, London
14. Boud D, Cohen R, Sampson J (1999) Peer learning and assessment. Assess Eval High Educ 24(4):413–426
15. Boud D, Cohen R, Sampson J (2014) Peer learning in higher education: learning from and with each other. Routledge, London
16. Braun V, Clarke V (2012) Thematic analysis. In: Cooper H, Camic PM, Long DL, Panter AT, Rindskopf D, Sher KJ (eds) APA handbooks in psychology®. APA handbook of research methods in psychology, vol 2. Research designs: Quantitative, qualitative, neuropsychological, and biological. American Psychological Association, p 57–71. https://doi.org/10.1037/13620-004
17. Bruffee KA (1993) Collaborative learning: higher education, interdependence, and the authority of knowledge. The Johns Hopkins University Press, Baltimore
18. Campbell DT, Stanley JC (2015) Experimental and quasi-experimental designs for research. Cambridge, Cambridge
19. Carver J, Jaccheri L, Morasca S, Shull F (2004) Issues in using students in empirical studies in software engineering education. In: Proceedingsof the 5th international workshop on enterprise networking and computing in healthcare industry (IEEE Cat. No. 03EX717). IEEE, pp 239–249

20. Christiaens G, Baldwin JH (2002) Use of dyadic role-playing to increase student participation. Nurse Educ 27(6):251–254
21. Connor AM, Buchan J, Petrova K (2009) Bridging the research-practice gap in requirements engineering through effective teaching and peer learning. In: 2009 Sixth international conference on information technology: new generations. IEEE, pp 678–683
22. Damian D, Al-Ani B, Cubranic D, Robles L (2005) Teaching requirements engineering in global software development: a report on a three-university collaboration. In: REET 2005. IEEE, pp 685–690
23. Davis A, Dieste O, Hickey A, Juristo N, Moreno AM (2006) Effectiveness of requirements elicitation techniques: empirical results derived from a systematic review. In: RE'06. IEEE, pp 179–188
24. Dewey J (1986) Experience and education. In: The educational forum, vol 50. Taylor & Francis, pp 241–252. https://doi.org/10.1080/00131728609335764
25. Distanont A, Haapasalo H, Vaananen M, Lehto J (2012) The engagement between knowledge transfer and requirements engineering. IJKL 1(2):131–156
26. Dochy F, Segers M, Sluijsmans D (1999) The use of self-, peer and co-assessment in higher education: a review. Stud High Educ 24(3):331–350
27. Donati B, Ferrari A, Spoletini P, Gnesi S (2017) Common mistakes of student analysts in requirements elicitation interviews. In: International working conference on requirements engineering: foundation for software quality. Springer, pp 148–164
28. Doorn N, Kroesen JO (2013) Using and developing role plays in teaching aimed at preparing for social responsibility. Sci Eng Ethics 19(4):1513–1527
29. Ferrari A, Spoletini P, Bano M, Zowghi D (2019) Learning requirements elicitation interviews with role-playing, self-assessment and peer-review. In: 2019 IEEE 27th international requirements engineering conference (RE), pp 28–39. https://doi.org/10.1109/RE.2019.00015
30. Ferrari A, Spoletini P, Bano M, Zowghi D (2020) SaPeer and ReverseSaPeer approaches for training students in requirements elicitation interviews—educational material (version 2.0). https://doi.org/10.5281/zenodo.2625706
31. Fowler J (2008) Experiential learning and its facilitation. Nurse Educ Today 28(4):427–433
32. Gabrysiak G, Giese H, Seibel A, Neumann S (2010) Teaching requirements engineering with virtual stakeholders without software engineering knowledge. In: REET'10. IEEE, pp 36–45
33. Greenberg J, Eskew DE (1993) The role of role playing in organizational research. J Manag 19(2):221–241
34. Greenblat CS (1973) Teaching with simulation games: a review of claims and evidence. Teach Sociol 1(1):62–83
35. Hadar I, Soffer P, Kenzi K (2014) The role of domain knowledge in requirements elicitation via interviews: an exploratory study. REJ 19(2):143–159
36. Hall T, Flynn V (2001) Ethical issues in software engineering research: a survey of current practice. Empir Softw Eng 6(4):305–317
37. Hamer J, Cutts Q, Jackova J, Luxton-Reilly A, McCartney R, Purchase H, Riedesel C, Saeli M, Sanders K, Sheard J (2008) Contributing student pedagogy. ACM SIGCSE Bull 40(4):194–212
38. Hammond M, Collins R (1991) Self-directed learning: critical practice. ERIC
39. Harris LR, Brown GT (2013) Opportunities and obstacles to consider when using peer-and self-assessment to improve student learning: case studies into teachers' implementation. Teach Teach Educ 36:101–111
40. Johnson RT, Johnson DW (2008) Active learning: cooperation in the classroom. Annu Rep Educ Psychol Jpn 47:29–30
41. MacLeod L, Greiler M, Storey MA, Bird C, Czerwonka J (2018) Code reviewing in the trenches: challenges and best practices. IEEE Softw 35(4):34–42
42. McIntosh S, Kamei Y, Adams B, Hassan AE (2016) An empirical study of the impact of modern code review practices on software quality. Empir Softw Eng 21(5):2146–2189
43. Meyer DK, Turner JC (2006) Re-conceptualizing emotion and motivation to learn in classroom contexts. Educ Psychol Rev 18(4):377–390
44. Moreno JL (1946) Psychodrama, vol 1. Beacon House, New York
45. Niknafs A, Berry DM (2013) An industrial case study of the impact of domain ignorance on the effectiveness of requirements idea generation during requirements elicitation. In: RE'13. IEEE, pp 279–283
46. Nunan D (1996) Towards autonomous learning: some theoretical, empirical and practical issues. In: Pemberton R et al. (eds) Taking control: autonomy in language learning, Hong Kong University Press. pp 13–24
47. Ouhbi S, Idri A, Fernández-Alemán JL, Toval A (2015) Requirements engineering education: a systematic mapping study. Requir Eng 20(2):119–138
48. Pearce J, Mulder R, Baik C (2009) Involving students in peer review: case studies and practical strategies for university teaching. https://people.eng.unimelb.edu.au/jonmp/pubs/Praze/Student_Peer_Review.pdf. Accessed 14 July 2020
49. Pitts MG, Browne GJ (2007) Improving requirements elicitation: an empirical investigation of procedural prompts. Inf Syst J 17(1):89–110
50. Race P (2001) The lecturer's toolkit: a practical guide to learning, teaching and assessment. Psychology Press, London
51. Regev G, Gause DC, Wegmann A (2009) Experiential learning approach for requirements engineering education. REJ 14(4):269–287
52. Ritchie J, Lewis J, Nicholls CM, Ormston R et al (2013) Qualitative research practice: a guide for social science students and researchers. Sage, Thousand Oaks
53. Runeson P, Host M, Rainer A, Regnell B (2012) Case study research in software engineering: guidelines and examples. Wiley, New York
54. Singer J, Vinson NG (2002) Ethical issues in empirical studies of software engineering. IEEE Trans Softw Eng 28(12):1171–1180
55. Spoletini P, Ferrari A, Bano M, Zowghi D, Gnesi S (2018) Interview review: an empirical study on detecting ambiguities in requirements elicitation interviews. In: International working conference on requirements engineering: foundation for software quality. Springer, pp 101–118
56. Svensson RB, Regnell B (2017) Is role playing in requirements engineering education increasing learning outcome? Requirements Eng 22:475–489. https://doi.org/10.1007/s00766-016-0248-4
57. Svensson RB, Regnell B (2017) Is role playing in requirements engineering education increasing learning outcome? Requir Eng 22(4):475–489
58. To J, Panadero E (2019) Peer assessment effects on the self-assessment process of first-year undergraduates. Assess Eval High Educ 44(6):920–932. https://doi.org/10.1080/02602938.2018.1548559
59. Topping K (2003) Self and peer assessment in school and university: reliability, validity and utility. In: Segers M, Dochy F, Cascallar E (eds) Optimising new modes of assessment: in search of qualities and standards. Innovation and change in professional education, vol 1. Springer, Dordrecht. https://doi.org/10.1007/0-306-48125-1_4
60. Van Zundert M, Sluijsmans D, Van Merriënboer J (2010) Effective peer assessment processes: research findings and future directions. Learn Instr 20(4):270–279

61. Walia GS, Carver JC (2013) Using error abstraction and classification to improve requirement quality: conclusions from a family of four empirical studies. Empir Softw Eng 18(4):625–658
62. Wang C, Cui P, Daneva M, Kassab M (2018) Understanding what industry wants from requirements engineers: an exploration of re jobs in Canada. In: Proceedings of the 12th ACM/IEEE international symposium on empirical software engineering and measurement. ACM, p 41
63. Wieringa R, Daneva M (2015) Six strategies for generalizing software engineering theories. Sci Comput Program 101:136–152
64. Wohlin C, Runeson P, Höst M, Ohlsson MC, Regnell B, Wesslén A (2012) Experimentation in software engineering. Springer, Berlin
65. Zowghi D, Coulin C (2005) Requirements elicitation: a survey of techniques, approaches, and tools. In: Aurum A, Wohlin C (eds) Engineering and managing software requirements. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-28244-0_2
66. Zowghi D, Paryani S (2003) Teaching requirements engineering through role playing: lessons learnt. In: RE'03. IEEE, pp 233–241