Learning Requirements Elicitation Interviews with Role-playing, Self-assessment and Peer-review

***Abstract*—Interviews are largely used in the practice of requirements elicitation. Performing an effective interview often depends on soft-skills, and on knowledge acquired through experience. 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. This paper presents a novel pedagogical approach that combines role-playing, peer-review and self-assessment to enable students to reflect on their mistakes, and improve their interview skills. We evaluate the approach through a controlled quasi-experiment. The study shows that the approach significantly reduces the amount of mistakes made by the students. Feedback from the participants confirms the usefulness and easiness of the proposed training. This work contributes to the body of knowledge of REET with an empirically evaluated method for teaching interviews. Furthermore, we share the pedagogical material used, to enable other educators to apply and possibly tailor the approach.**

I. Introduction

Requirements elicitation is normally performed by means of interviews between a requirements analyst and a customer, as well as other stakeholders such as domain or technical experts [1]–[3]. The ability of the analyst to gather correct and complete requirements from these subjects often depends on the analyst's experience as well as on natural soft-skills [2], [4]–[8]. Given the multiple factors influencing the success of elicitation, teaching the art of interviews to software engineering and computer science students is particularly difficult, also due to the limited resources normally available for educational activities specifically focused on requirements engineering (RE) [9], [10]. In RE education and training (REET) it is recommended practice to perform role-playing activities [11]–[13], in which students can play the role of requirements analysts, so to have a hands-on experience with interviews. Previous work has shown that students tend to commit mistakes in these simulated interviews, and have suggested that the mistakes can be leveraged to give feedback to students and make them improve their interview skills [14]–[16]. 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 [17]–[21].

This paper combines existing research on mistakes of student analysts [15], [16], role-playing [11], [12], peer-review and self-assessment [17]–[21] to propose a novel approach for REET named SAPEER (role-playing, Self-Assessement 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.* [16] 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 of mistakes from the first to the second interview, enabled by SAPEER. Our results show that the proposed approach significantly reduces the amount of mistakes made by students. The results also suggest 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 contribution of this paper is twofold. We present a novel pedagogical approach, with all the slides, modules and information needed to apply it in other contexts. Furthermore, we present an empirical assessment of the proposed approach, showing its effectiveness in terms of mistakes' reduction.

The remainder of the paper is structured as follows. In Sect. II we present related work and background. In Sect. III we describe the proposed approach. In Sect. IV we report the research design for the experiment and in Sect V we describe the results. Sect. VI reports observation on the results. Conclusion and future work are presented in Sect. VII.

II. Background

The software engineering discipline is required to produce graduates with the ability of learning from reflections, either on personal competence or those of their peers in collaborative environments, eventually turning them into reflective practitioners in their field [21]. Role-playing, which requires students to play a certain role—e.g., requirements analyst—in a simulated scenario, is based on Dewey's *learning by*

doing philosophy [22], and is recommended in socio-technical areas such as RE and software engineering in general [12], [23], in which experiential learning is critical. Peer-review and self-assessment are pedagogical practices that emphasize on learning through reflection [20]. Cognitively and socially, peer-review and self-assessment differ from traditional teacher-lead assessments, and rely on the judgment and skills of students [24]. Pedagogical designs that combine peer-review along with self-assessment in a collaborative learning environment provide significant increase in students' learning outcomes [25] and have been reported to help students to improve their skills in areas of communication, self-evaluation, observation, and self-criticism [26]. In the following, we briefly summarise background work on role-playing, self-assessment and peer-review, to provide the context to understand the principles underlying SAPEER. 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.

a) Role-playing: Role-playing is a technique rooted in Moreno's *psychodrama* method [27], and is largely used for education in several fields, including nursing [28], management [29], and RE [11], [12], [30], [31]. Role-playing has been reported to improve cognitive and affective learning [32], and to be a proper support to train communication skills [33]. In software engineering education, role-playing is used for different objectives [10], [34], such as training students on software modelling and development [35], requirements inspection [31], and requirements elicitation and documentation [30]. The empirical study of Svensson and Regnell [36] have suggested that role-playing can improve students' competences in RE.

b) Self-Assessment: Self-assessment, also known as self-evaluation [25], is a critical skill for becoming lifelong learners. Though traditionally self-assessment is not considered part of formal assessment methods in education, it holds a critical role in self-learning processes. Autonomous learning [17], [37], experiential learning [38], or self-directed learning [39] all rely on the self-assessment ability of an individual. It requires the students to develop the ability of critical reflections on past knowledge or practice. In self-assessment, the students grade their own work, by actively participating in the process of assessment. This has been advocated to enhance students' understanding of the quality of the work or knowledge, sharpen their critical analysis skills and assist them in becoming self-learners [40], [41].

c) Peer-review: Peer-review consists in evaluating the work or artifacts produced by peers in a certain working or educational environment [18]. Peer-review in education is based on the principles of *peer learning* theories. 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 [42]. Peer learning is a form of informal and collaborative learning that not just happens outside the classroom environment, but can be utilised effectively within classroom assessments [18]. There are multiple learning outcomes associated with peer learning such as enhancement in social skills, constructive feedback and critical analysis by observing peers, reflective learning, and articulation of knowledge [18]. Peer-review falls under collaborative learning pedagogies that are based on cognitive, social and developmental psychology [17], [39], [43]. In the software engineering practice, peer-review is also largely used to improve the quality of artifacts such as code, requirements [44]–[47], and, more recently, interviews [48].

d) Mistakes of Student Analysts: As novices, RE students naturally tend to commit mistakes during requirements elicitation. In an exploratory work, Donati *et. al.* [15] identified a first set of 9 general mistake categories. Based on this work, Bano et al. [16] performed a more empirically grounded study involving 110 students, 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 Behaviour (e.g., *lack of confidence*), Customer Interaction (e.g., *no rapport*), 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.

e) Contribution to REET: The systematic mapping study presented by Ouhbi *et al.* [10] 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 of students learning. From the mapping study, only one study from Connor *et al.* [49] reported the utilisation of *peer learning* theory, though not formally integrated in the curriculum. The lack of studies and evidence on REET suggests that is a need for proposing and assessing innovative pedagogies to equip graduates with the skills they need in real-world contexts [36].

Formalising peer-review and self-assessment within a study curriculum can increase the level of understanding regarding the process and the outcomes to be achieved [50] and create a positive learning experience for students due to their involvement in the assessment process [19], [21]. To our knowledge, this is the first work that proposes a pedagogical approach for teaching requirements elicitation interviews that combines role-playing, peer-review and self-assessment through a coherent training framework. Furthermore, this work differs from that of Bano *et al.* [16], in that it provides an *operationalisation* of their empirical results, by leveraging the identified mistake types to improve students' interview skills.

III. THE SAPEER APPROACH

This section presents the SAPEER pedagogical approach. The fundamental idea of the approach is to first foster experiential learning, by letting students perform a role-playing interview, and then stimulate learning through reflection, by asking students to identify mistakes in their own interview and in the interview of their peers. The acquired competence is then tested in a second interview. Fig. 1 shows the main building blocks of SAPEER, described below. All the resources

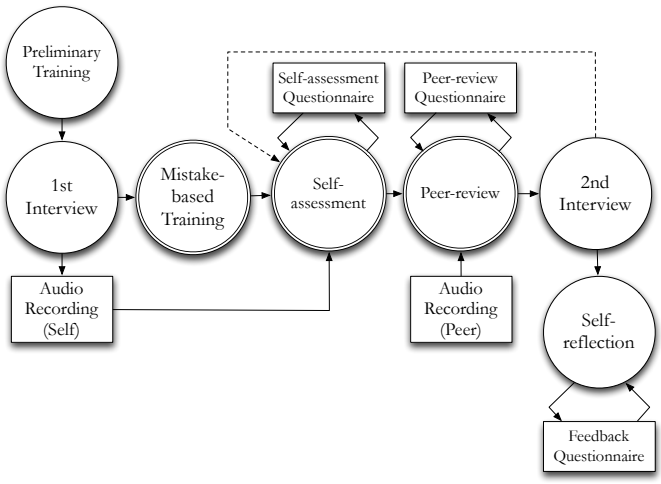

Fig. 1: Overview of the SAPEER approach.

associated to the approach, i.e., lecture slides (videos not currently shared for double-blind policy), questionnaires, and product descriptions, are publicly available [51]. The approach is designed to be performed entirely *online*, but can be adapted for classroom environments.

1) **Preliminary Training:** the students are given a first video lecture of about 20 minutes on how to conduct interviews, which focuses on positive advice.
2) **1st Interview:** each student conducts their 1st 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 minutes, in which the student analysts' mistakes presented by Bano *et al.* [16] are described, and examples of erroneous behaviour are given for each mistake. The lecture presents the mistake types that are applicable to the current context (32 out of 34), in which interviews are conducted online and involve a single analyst. 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. Example statements are: *I asked vague questions*, *I did not ask for additional stakeholders*, etc. 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, according to 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, except for the formulation of the statements, which in this case are in third person (*The analyst asked vague questions*, etc.). Again, a score is produced for each statement.
6) **2nd Interview:** every student conduct their 2nd Skype interview with a tutor playing the role of customer, but for a different product with respect to the 1st 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.
7) **Self-reflection:** the students are given a feedback questionnaire, in which they are asked to evaluate the usefulness and easiness of the different types of steps of the training, and to provide comments on their experience. Specifically, the types of steps to evaluate are Preliminary Training, Interviews, Peer-review, Self-assessment and Mistake-based Training. The students are asked to evaluate their usefulness on a 5-point Likert Scale—Extremely useful (5), Very useful (4), Moderately useful (3), Slighly useful (2), Not at all useful (1). Similarly for easiness–Very easy (5), Moderately easy (4), Neither easy nor difficult (3), Moderately difficult (2), Very difficult (1).

The design of SAPEER is modular, and can be iterated based on the time available, by, e.g., performing self-assessment and peer-review of the 2nd 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.

IV. 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 [52], [53][1] with an experimental group and a control group. The experimental group adopts the pedagogical approach described in Sect. III, 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.

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

A. Research Questions

In the experiment, we want to first assess whether the approach leads to a reduction of mistakes from the 1st to the

[1]The design is analogous to a randomised 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 a case-study settings [54].

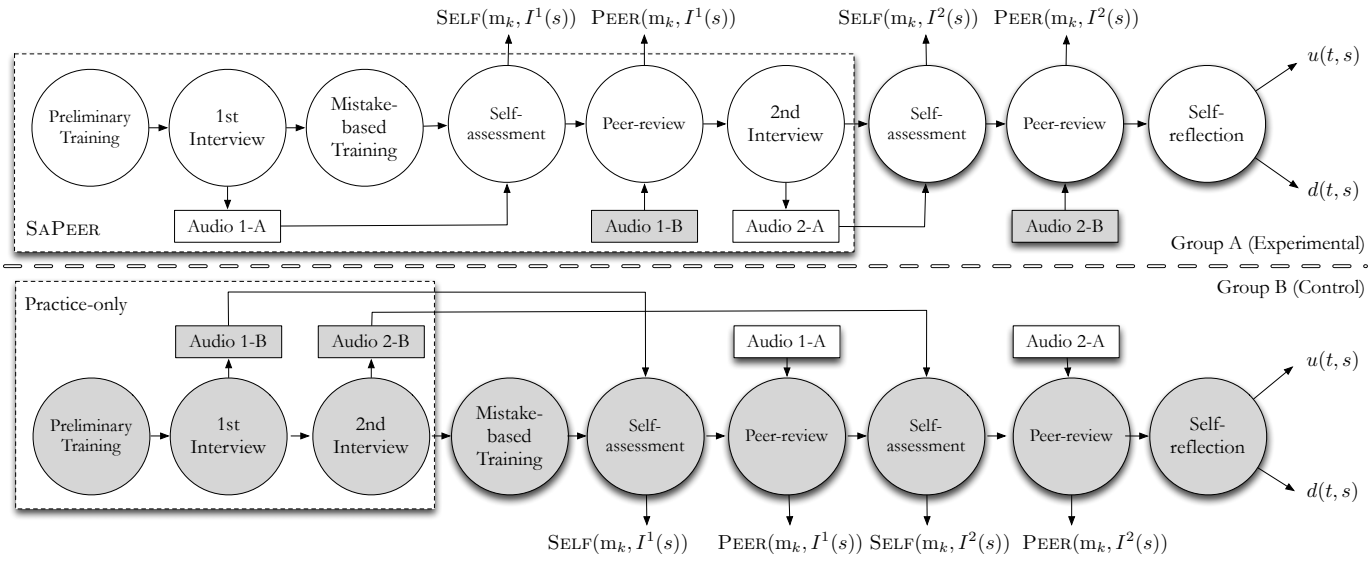

Fig. 2: Overview of the Experimental Procedure, with the scores collected from the different questionnaires.

2nd interview. Then, we want to check to which extent the reduction 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 following the SAPEER treatment in the 1st and 2nd interview, and we assess whether the mistakes are reduced in the 2nd interview. Answering this RQ requires to focus solely on the mistakes of the experimental group.
- **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 of mistakes in the 2nd interview is due to the steps 3, 4, and 5 of SAPEER, or it is mostly due to experience acquired during the 1st 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.
- **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.

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

B. Study Context

The experiment is conducted in the context of a RE course. The 43 participants of the study are graduate students majoring in software engineering. The students have very heterogeneous background, also because their program admits students who transition into computing from other disciplines. 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.

C. Experimental Procedure and Data Collection

Fig. 2 summarises 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 to group A are in white, while those associated to B are in grey in Fig. 2. Both groups perform the preliminary training activity, and the 1st interview with a tutor, which was constrained to last 15 minutes maximum. Then, group A performs mistake-based training, self-assessment on Audio 1-A (i.e., the audio recording of the 1st interview from group A), and peer-review. This step uses the audio recording of the 1st interview from group B (Audio 1-B). Both groups perform the 2nd interview (max 15 minutes).

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 1st interviews (Audio 1-B, Audio 1-A). Then, both groups analyse the 2nd 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 of mistakes occurred from the 1st to the 2nd 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**).

D. Study Variables

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

1) Amount of Mistakes: Let S^A and S^B 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 1st or 2nd interview. Each interview $I^h(s_j)$ receives two reviews, a self-assessment and a peer-review, oriented to evaluate the mistakes. Let $\mathcal{M}$ the set of 32 mistake types (Sect. III). 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. 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.

2) Effectiveness: We define the effectiveness evaluated on a certain student s as the ratio between their amount of mistakes in the 1st and 2nd interview:

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

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 1st and 2nd interview, respectively. We define the effectiveness $\hat{E}$ for a single mistake m_k and student s as the ratio between the mistakes in the 1st and 2nd interview:

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

Higher values of E and $\hat{E}$ indicate more effectiveness.

3) Usefulness: The usefulness variable is computed for each single type of step of the proposed training. As specified in Sect. III, 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.

4) 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.

E. Analysis Procedure and Hypotheses

The analysis procedure consists in testing a set of hypotheses derived from the RQs in Sect. IV-A. Below, we define the null and alternative hypotheses associated to each RQ. 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% ($p \leq 0.05$).

a) 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 1st and in the 2nd 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 2nd interview is greater or equal than the amount of mistakes in the 1st interview". We perform a *one tail* test, with alternative hypothesis: $\mu_\delta < 0$, i.e., H_{11} = "the amount of mistakes in the 2nd interview is *lower* than the amount of mistakes in the 1st 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 1st and 2nd 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 2nd interview is greater or equal than the amount of mistakes in the 1st 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 2nd interview is *lower* than the amount of mistakes of type m_k in the 1st interview".

b) 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 to 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" ($\mu_{\hat{E}_A} > \mu_{\hat{E}_B}$).

c) 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 towards higher level of usefulness. This evaluation is based on Carver *et al.* [31].

d) 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_{41}^t = "the easiness of the step of type t is greater than the midpoint of the scale" ($\mu_d > 3$)—higher levels of easiness.

F. Validity Procedure

a) 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. IV-D), 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 = 5.27e-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.

The mistake types were derived by Bano et al. [16] in the context of group interviews, so some mistakes typical of individual interviews may not be part of our list. To overcome this issue the feedback questionnaire asked students to include additional mistakes not originally listed. No additional cases were identified that could not be traced to the original types.

b) Internal Validity: To address problems related to possible imbalance of competence between the groups, the mistakes committed in the 1st interview can be considered as a *pre-test* 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 1st 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-samples 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* [52], 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. V.

Another threat of internal validity may be the influence of 7 tutors acting as customers. To have uniform treatments, tutors received common instructions, 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.

c) 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 generalisation [55], there are architectural aspects of the study that can be used as a term of comparison to generalise the results: participants are MSc students in software engineering, all the training activities were performed online, 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.

V. 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.

A. RQ1: Mistake Reduction

As shown in Fig. 4, the amount of mistakes M is reduced from the 1st to the 2nd interview for group A. As both samples passed the Shapiro-Wilk's test of normality (W = 0.96396, p

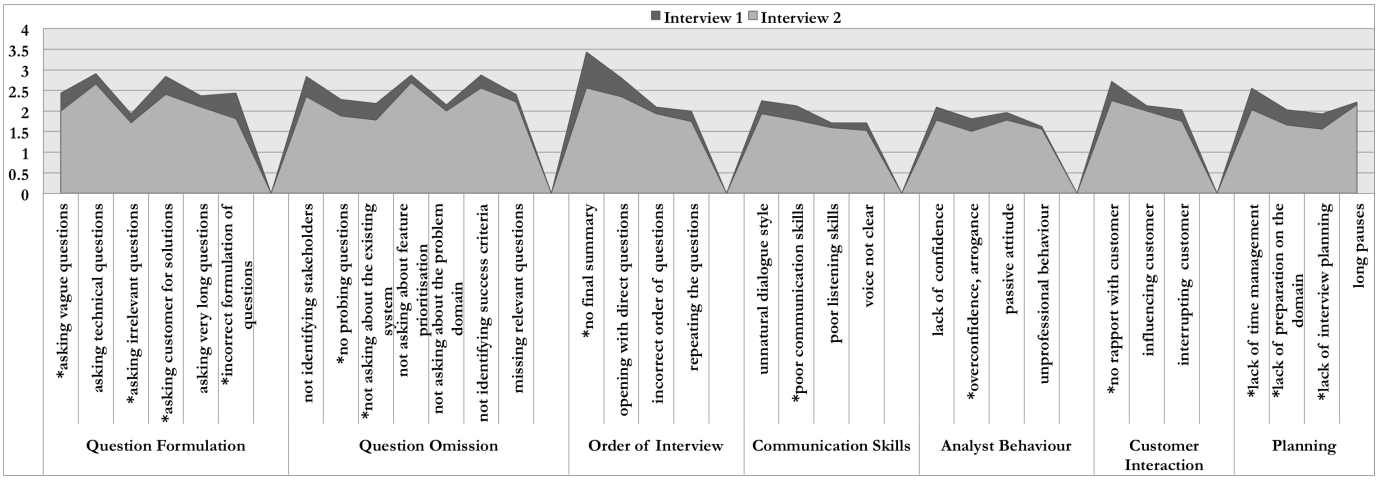

Fig. 3: Average over students of the amount of single mistakes $\hat{M}$ in 1st and 2nd interview for group A.

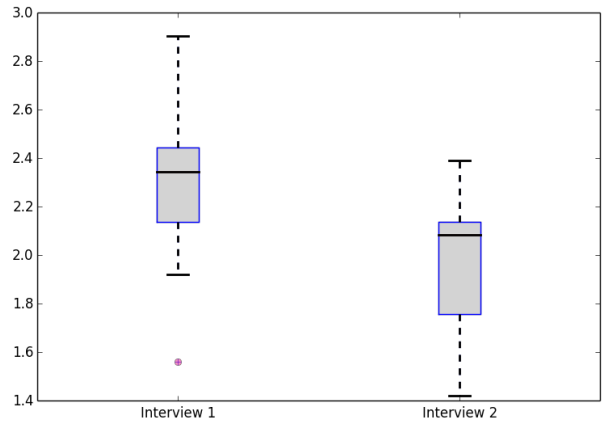

Fig. 4: Amount of mistakes M in 1st and 2nd interv. for group A.

= 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 t = -4.7721, df = 15, p = 0.0001235, hence H_{10} is rejected in favour of H_{11}.

To have further insight, and understand for which mistakes we had a major improvement in the 2nd interview, it is useful to look at Fig. 3. 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[3] 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. 3. We see that there is a general reduction of 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 prioritisation*, 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 2nd 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 Behaviour and Customer Interaction. These are also the classes in which less mistakes were already committed during the 1st interview (as the dark area is lower with respect to the other classes).

[3]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.

B. RQ2: Effectiveness

The SAPEER treatment appears to be slightly more effective than the practice-only treatment, as shown by Fig. 6. 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-samples 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.

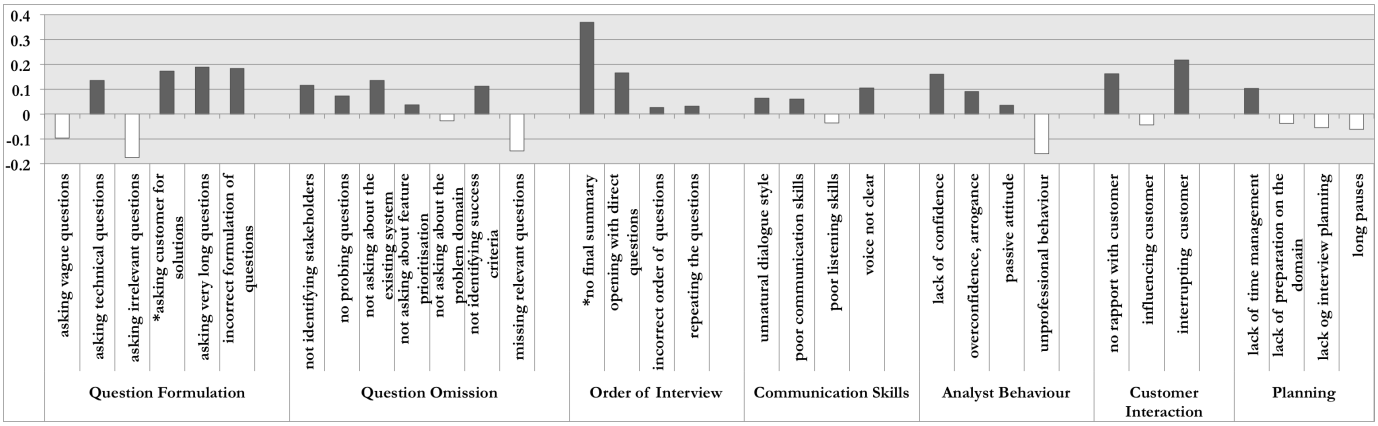

Fig. 5: Difference between group A and B in terms of average of the effectiveness for single mistakes $\hat{E}$.

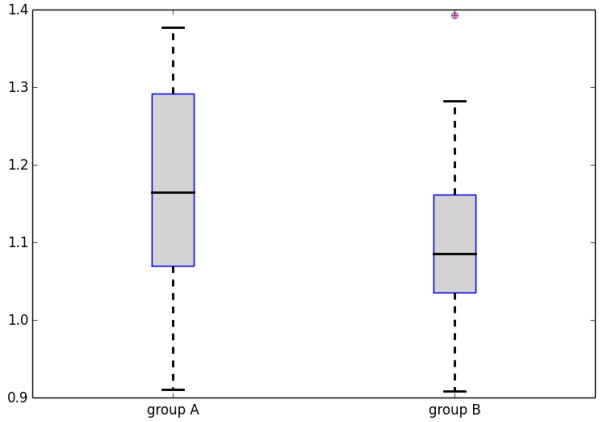

Fig. 6: Effectiveness E of the SAPEER treatment (group A) with respect to the practice-only treatment (group B).

It is now useful to compare the effectiveness for the two groups, considering each single mistake. Fig. 5 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. IV. 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-samples Wilcoxon test (i.e., a Mann-Whitney test), to check whether the effectiveness of SAPEER treatment is greater than the practice-only treatment[4] ($H_{21}^{m_k}$). Significant cases, for which $H_{20}^{m_k}$ is rejected, are marked with * in Fig. 3. 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 also cases of mistakes in which group B improved more than group A, such as *asking irrelevant questions*, *missing relevant questions* and *unprofessional behaviour*, and most of the mistakes related to Planning. In Sect. V-A 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, e.g., running out of time—, rather than the application of all the steps of the SAPEER treatment.

[4]For those mistakes in which the practice-only treatment is clearly more effective (white bars in Fig. 5), 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.

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	U = 3 V = 14 p = 0.536	U = 3.3 V = 24 p = 0.203	U = 4.2 V = 45 p = 0.003

TABLE I: Average usefulness U and Wilcox. Sign. Rank test results.

C. RQ3: Usefulness

Students were required to evaluate the degree of usefulness of the different steps of the approach. Fig. 7 reports the results for group A and group B. Table I 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 characteristics we determine whether the mean response is significantly greater than the midpoint of the scale, i.e., 3 = Moderately useful (H_{31}^t), by applying the Wilcoxon Signed Rank test. Non-significant cases for which H_{30}^t is not rejected are marked in grey.

From Fig. 7, we see that both groups considered most of the steps Moderately to Extremely useful, with group A more oriented towards a positive judgment, as none of the respondents selected Slighly 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

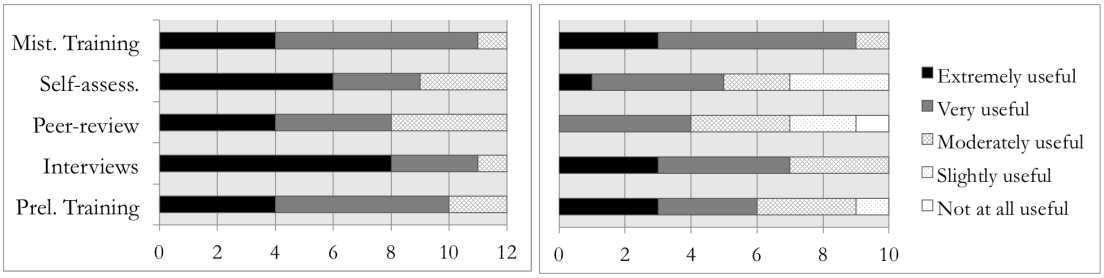

Fig. 7: Results for the Usefulness variable for group A (left) and B (right).

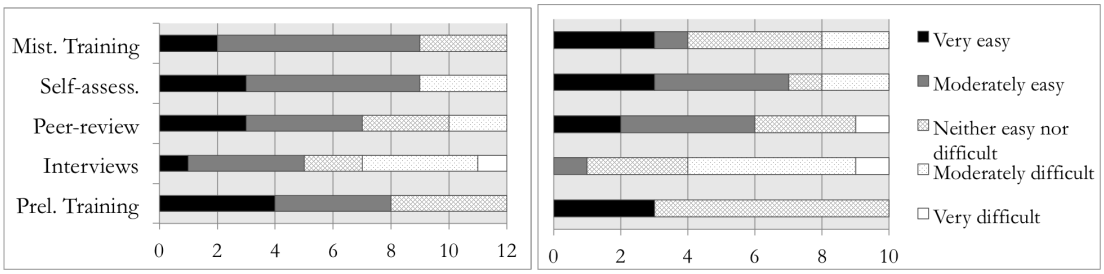

Fig. 8: Results for the Easiness variable for group A (left) and B (right).

Gr.	Prel. Train.	Interv.	Peer-review	Self-assess.	Mist. Train.
A	D = 4 V = 36 p = 0.006	D = 3 V = 27.5 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	D = 3.6 V = 6 p = 0.074	D = 2.4 V = 3.5 p = 0.976	D = 3.6 V = 22 p = 0.096	D = 3.8 V = 38 p = 0.032	D = 3.5 V = 17 p = 0.099

TABLE II: Average easiness D and Wilcoxon Sign. Rank test results.

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 2nd 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 I, 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.

D. RQ4: Easiness

As for usefulness, students were required to give feedback about the easiness of the steps of the approach. Fig. 8 reports descriptive statistics for the two groups, while Table II 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}). Non-significant cases (H^t_{40} not rejected) are marked in grey.

From Fig. 8, we see that both groups considered most of the steps Neither easy nor difficult to Very easy. One exception are the interviews, which have been considered more difficult, especially by group B. This group performed the 2nd 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 II, 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 3.8 and 4 (Moderately easy).

VI. DISCUSSION

We organise our discussion by first listing the main take-away messages from our results and then discussing the complimentary role of role-playing and the other steps of SAPEER. We relate our results with existing work in REET, and we finally outline ideas for improving and tailoring the approach to different classroom environments.

a) Take-away Messages: The main take-away messages from our study are: 1) SAPEER enables a reduction of mistakes already from the first to the second interview (Sect. V-A); 2) the steps of SAPEER, and in particular interviews and mistake-based training, are considered useful (Sect. V-C); 3) although interviews are considered among the most useful steps, they are also considered as more challenging than the other steps, generally evaluated as moderately easy (Sect. V-D); 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. V-B).

b) Complementarity of the Steps: From Sect. V-B, we 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 behavioural and systemic, such as those related to Communication Skills, Analyst Behaviour and Customer interaction. We argue that these mistakes are harder to correct with recommendations, and may require more experience and time. For mistakes related to Planning, significant improvement was observed in students following the SAPEER treatment (Fig. 3). However, the improvement was even higher for students following the practice-only treatment (Fig. 5). 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. 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.

c) Results in Relation to Education Literature: The current work confirms the utility of role-playing in education in general [32], [33] and REET in particular [12], [30], [36], 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 [19], [21]. However, the effect of self-assessment and peer-review practices, although positive (Fig. 6), is not statistically significant for what concerns mistake reduction. Given that these are well established practices [21], [56], with a long history founded on philosophical and pedagogical theories of constructivism and community learning [57], further experimentation is needed, possibly based on an improved version of the approach.

d) Improving the Approach: As mentioned, some behavioural 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. 3, the $*$ symbol marks significance), for which suggestions can be provided. We argue that students need further guidance on this aspect, and list of *right* questions to ask may be beneficial, as also suggested by some students' comments from 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*). Questions to start an interview, and to identify missing stakeholders, were suggested by Donati et al. [15]. Other questions can be defined based on the studies of Pitts and Browne on procedural prompts [58], and studies about interviews from other fields such as journalism [59] or social sciences [60].

Another aspect to improve the training 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—1st interviews for group A were reviewed after 2nd ones were performed—and feedback from peers may have also some drawbacks as recently noticed by To and Panadero [61]. 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 (e.g., *I would replace the self-assessment or at least add an assessment from the professor and even the students who played the interviewee roles*; *getting the feedback from the first interview before doing the second may have helped*).

e) Tailoring the Approach: SAPEER is designed to be modular and adaptable, and, although the steps should preferably performed in the recommended order to prevent difficulties (see Sect. V-D), 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* [12], with students acting as customers, can be applied. Furthermore, if time is also crucial, given the results from Sect. V-B, 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. SAPEER is 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 [62].

VII. Conclusion and Future Work

This paper presents and evaluates SAPEER, a novel pedagogical approach for teaching requirements elicitation interviews. The approach is based on role-playing, peer-review and self-assessment, and leverages previous research on mistakes of student analysts in RE [15], [16]. 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 asking for a summary at the end of the interview, or asking probing questions. Mistakes more related to behavioural aspects are harder to correct, and some mistakes in the area of question omission are not correctly addressed at the moment. Future work will focus on further improvement and dissemination of SAPEER. 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. Experiments will be performed to assess the effectiveness of the improved approach, and to better understand the relationship between the steps of the training and the reduction of specific types of mistakes.

References

[1] A. Davis, O. Dieste, A. Hickey, N. Juristo, and A. M. Moreno, "Effectiveness of requirements elicitation techniques: Empirical results derived from a systematic review," in *RE'06*. IEEE, 2006, pp. 179–188.

[2] I. Hadar, P. Soffer, and K. Kenzi, "The role of domain knowledge in requirements elicitation via interviews: an exploratory study," *REJ*, vol. 19, no. 2, pp. 143–159, 2014.

[3] R. Agarwal and M. R. Tanniru, "Knowledge acquisition using structured interviewing: an empirical investigation," *JMIS*, vol. 7, no. 1, pp. 123–140, 1990.

[4] A. Distanont, H. Haapasalo, M. Vaananen, and J. Lehto, "The engagement between knowledge transfer and requirements engineering," *IJKL*, vol. 1, no. 2, pp. 131–156, 2012.

[5] D. Zowghi and C. Coulin, "Requirements elicitation: A survey of techniques, approaches, and tools," in *Engineering and managing software requirements*. Springer, 2005, pp. 19–46.

[6] A. M. Aranda, O. Dieste, and N. Juristo, "Effect of domain knowledge on elicitation effectiveness: An internally replicated controlled experiment," *TSE*, vol. 42, no. 5, pp. 427–451, 2016.

[7] A. Niknafs and D. M. Berry, "An industrial case study of the impact of domain ignorance on the effectiveness of requirements idea generation during requirements elicitation," in *RE'13*. IEEE, 2013, pp. 279–283.

[8] C. Wang, P. Cui, M. Daneva, and M. Kassab, "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, 2018, p. 41.

[9] G. Gabrysiak, H. Giese, A. Seibel, and S. Neumann, "Teaching requirements engineering with virtual stakeholders without software engineering knowledge," in *REET'10*. IEEE, 2010, pp. 36–45.

[10] S. Ouhbi, A. Idri, J. L. Fernández-Alemán, and A. Toval, "Requirements engineering education: a systematic mapping study," *Requirements Engineering*, vol. 20, no. 2, pp. 119–138, 2015.

[11] R. B. Svensson and B. Regnell, "Is role playing in requirements engineering education increasing learning outcome?" *REJ*, pp. 1–15, 2016.

[12] D. Zowghi and S. Paryani, "Teaching requirements engineering through role playing: Lessons learnt," in *RE'03*. IEEE, 2003, pp. 233–241.

[13] G. Regev, D. C. Gause, and A. Wegmann, "Experiential learning approach for requirements engineering education," *REJ*, vol. 14, no. 4, pp. 269–287, 2009.

[14] C. Argyris and D. A. Schon, *Theory in practice: Increasing professional effectiveness.* Jossey-Bass, 1974.

[15] B. Donati, A. Ferrari, P. Spoletini, and S. Gnesi, "Common mistakes of student analysts in requirements elicitation interviews," in *International Working Conference on Requirements Engineering: Foundation for Software Quality*. Springer, 2017, pp. 148–164.

[16] M. Bano, D. Zowghi, A. Ferrari, P. Spoletini, and B. Donati, "Learning from mistakes: An empirical study of elicitation interviews performed by novices," in *2018 IEEE 26th International Requirements Engineering Conference (RE)*. IEEE, 2018, pp. 182–193.

[17] D. Boud, *Developing student autonomy in learning*. Routledge, 2012.

[18] D. Boud, R. Cohen, and J. Sampson, "Peer learning and assessment," *Assessment & evaluation in higher education*, vol. 24, no. 4, pp. 413–426, 1999.

[19] J. Pearce, R. Mulder, and C. Baik, "Involving students in peer review: Case studies and practical strategies for university teaching," Centre for the Study of Higher Education, University of Melbourne, 2009. [Online]. Available: https://people.eng.unimelb.edu.au/jonmp/pubs/Praze/Student_Peer_Review.pdf

[20] K. Topping, "Self and peer assessment in school and university: Reliability, validity and utility," in *Optimising new modes of assessment: In search of qualities and standards*. Springer, 2003, pp. 55–87.

[21] M. Van Zundert, D. Sluijsmans, and J. Van Merriënboer, "Effective peer assessment processes: Research findings and future directions," *Learning and instruction*, vol. 20, no. 4, pp. 270–279, 2010.

[22] J. Dewey, "Experience and education," in *The Educational Forum*, vol. 50, no. 3. Taylor & Francis, 1986, pp. 241–252.

[23] T. R. Henry and J. LaFrance, "Integrating role-play into software engineering courses," *Journal of Computing Sciences in Colleges*, vol. 22, no. 2, pp. 32–38, 2006.

[24] L. R. Harris and G. T. Brown, "Opportunities and obstacles to consider when using peer-and self-assessment to improve student learning: Case studies into teachers' implementation," *Teaching and Teacher Education*, vol. 36, pp. 101–111, 2013.

[25] D. Boud, *Enhancing learning through self-assessment*. Routledge, 2013.

[26] F. J. Dochy and L. McDowell, "Introduction: Assessment as a tool for learning." *Studies in educational evaluation*, vol. 23, no. 4, pp. 279–98, 1997.

[27] J. L. Moreno, "Psychodrama, first volume." 1946.

[28] G. Christiaens and J. H. Baldwin, "Use of dyadic role-playing to increase student participation," *Nurse educator*, vol. 27, no. 6, pp. 251–254, 2002.

[29] J. Greenberg and D. E. Eskew, "The role of role playing in organizational research," *Journal of management*, vol. 19, no. 2, pp. 221–241, 1993.

[30] D. Damian, B. Al-Ani, D. Cubranic, and L. Robles, "Teaching requirements engineering in global software development: a report on a three-university collaboration," in *REET 200)*. IEEE, 2005, pp. 685–690.

[31] G. S. Walia and J. C. Carver, "Using error abstraction and classification to improve requirement quality: conclusions from a family of four empirical studies," *Empirical Software Engineering*, vol. 18, no. 4, pp. 625–658, 2013.

[32] C. S. Greenblat, "Teaching with simulation games: a review of claims and evidence." *Teaching Sociology*, vol. 1, no. 1, pp. 62–83, 1973.

[33] N. Doorn and J. O. Kroesen, "Using and developing role plays in teaching aimed at preparing for social responsibility," *Science and Engineering Ethics*, vol. 19, no. 4, pp. 1513–1527, 2013.

[34] J. Carver, L. Jaccheri, S. Morasca, and F. Shull, "Issues in using students in empirical studies in software engineering education," in *Proceedings. 5th International Workshop on Enterprise Networking and Computing in Healthcare Industry (IEEE Cat. No. 03EX717)*. IEEE, 2004, pp. 239–249.

[35] G. Auriol, C. Baron, and J.-Y. Fourniols, "Teaching requirements skills within the context of a physical engineering project," in *REET 2008*. IEEE, 2008, pp. 6–11.

[36] R. B. Svensson and B. Regnell, "Is role playing in requirements engineering education increasing learning outcome?" *Requirements Engineering*, vol. 22, no. 4, pp. 475–489, 2017.

[37] D. Nunan, "Towards autonomous learning: some theoretical, empirical and practical issues," in *Taking Control: Autonomy in Language Learning*. Hong Kong University Press, 1996.

[38] J. Fowler, "Experiential learning and its facilitation," *Nurse Education Today*, vol. 28, no. 4, pp. 427–433, 2008.

[39] M. Hammond and R. Collins, *Self-Directed Learning: Critical Practice.* ERIC, 1991.

[40] P. Black and D. Wiliam, "Assessment and classroom learning," *Assessment in Education: principles, policy & practice*, vol. 5, no. 1, pp. 7–74, 1998.

[41] D. Sluijsmans, "Establishing learning effects with integrated peer assessment tasks," 2002.

[42] D. Boud, R. Cohen, and J. Sampson, *Peer learning in higher education: Learning from and with each other*. Routledge, 2014.

[43] K. A. Bruffee, *Collaborative learning: Higher education, interdependence, and the authority of knowledge.* ERIC, 1993.

[44] A. Aurum, H. Petersson, and C. Wohlin, "State-of-the-art: software inspections after 25 years," *Software Testing, Verification and Reliability*, vol. 12, no. 3, pp. 133–154, 2002.

[45] S. McIntosh, Y. Kamei, B. Adams, and A. E. Hassan, "An empirical study of the impact of modern code review practices on software quality," *Empirical Software Engineering*, vol. 21, no. 5, pp. 2146–2189, 2016.

[46] A. Bacchelli and C. Bird, "Expectations, outcomes, and challenges of modern code review," in *ICSE'13*. IEEE, 2013, pp. 712–721.

[47] L. MacLeod, M. Greiler, M.-A. Storey, C. Bird, and J. Czerwonka, "Code reviewing in the trenches: Challenges and best practices," *IEEE Software*, vol. 35, no. 4, pp. 34–42, 2018.

[48] P. Spoletini, A. Ferrari, M. Bano, D. Zowghi, and S. Gnesi, "Interview review: An empirical study on detecting ambiguities in requirements elicitation interviews," in *International Working Conference on Requirements Engineering: Foundation for Software Quality*. Springer, 2018, pp. 101–118.

[49] A. M. Connor, J. Buchan, and K. Petrova, "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, 2009, pp. 678–683.

[50] D. Saunders, "Peer tutoring in higher education," *Studies in Higher Education*, vol. 17, no. 2, pp. 211–218, 1992.

[51] Anonymous, "SaPeer Approach for Training Students in Requirements Elicitation Interviews—Educational Material," Apr. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2625706

[52] D. T. Campbell and J. C. Stanley, *Experimental and quasi-experimental designs for research*. Ravenio Books, 2015.
[53] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in software engineering*. Springer Science & Business Media, 2012.
[54] P. Runeson, M. Host, A. Rainer, and B. Regnell, *Case study research in software engineering: Guidelines and examples*. John Wiley & Sons, 2012.
[55] R. Wieringa and M. Daneva, "Six strategies for generalizing software engineering theories," *Science of computer programming*, vol. 101, pp. 136–152, 2015.
[56] F. Dochy, M. Segers, and D. Sluijsmans, "The use of self-, peer and co-assessment in higher education: A review," *Studies in Higher education*, vol. 24, no. 3, pp. 331–350, 1999.
[57] J. Hamer, Q. Cutts, J. Jackova, A. Luxton-Reilly, R. McCartney, H. Purchase, C. Riedesel, M. Saeli, K. Sanders, and J. Sheard, "Contributing student pedagogy," *ACM SIGCSE Bulletin*, vol. 40, no. 4, pp. 194–212, 2008.
[58] M. G. Pitts and G. J. Browne, "Improving requirements elicitation: an empirical investigation of procedural prompts," *Information systems journal*, vol. 17, no. 1, pp. 89–110, 2007.
[59] S. Adams, *Interviewing for journalists*. Psychology Press, 2001.
[60] J. Ritchie, J. Lewis, C. M. Nicholls, R. Ormston *et al.*, *Qualitative research practice: A guide for social science students and researchers*. sage, 2013.
[61] J. To and E. Panadero, "Peer assessment effects on the self-assessment process of first-year undergraduates," *Assessment & Evaluation in Higher Education*, pp. 1–14, 2019.
[62] P. Race, *The lecturer's toolkit:a practical guide to learning, teaching assessment*. Psychology Press, 2001.