Comparing expert and peer assessment of pedagogical design in integrated STEAM education

Kyriaki Vakkou[1], Tasos Hovardas[1], Nikoletta Xenofontos[1], and Zacharias Z. Zacharia[1]

[1]Research in Science and Technology Education Group, Department of Education, University of Cyprus

Author note

This study was partly supported by the Erasmus + project BRIDGES – Broadening Recognition Initiatives Developing Gender Equity in Sciences (ERASMUS +, 2019-1-FR01-KA203-062515) and partly co-funded by the European Union and the Republic of Cyprus through the Research and Innovation Foundation (Project: INNOVATE/0719/0098). We are grateful to all colleagues in these two projects for having fruitful discussions related to the scope and objectives of the present study. We are also grateful to all pre-service teachers who participated in the study. The authors have no conflicts of interest to disclose.

Correspondence concerning this manuscript should be addressed to Kyriaki Vakkou, Research in Science and Technology Education Group, Department of Education, University of Cyprus, 75 Kallipoleos St., PO 20537, 1678, Nicosia, Cyprus. Email: vakkou.kyriaki@ucy.ac.cy

Abstract

The objective of this exploratory study was to implement peer assessment for pedagogical design in integrated STEAM education and to compare expert and peer feedback, in this regard. We engaged pre-service teachers registered in an undergraduate programme for primary education in a formative/reciprocal peer assessment arrangement, where they had the chance to act as both peer assessors and peer assessees. Although global measures of validity (correlations between total scores of expert and peer assessors) and reliability (correlations between total scores of different peer assessors for the same pedagogical scenario) were satisfactory, there were assessment criteria for which peer assessment failed to be valid and/or reliable and which should deserve more attention in future training sessions. We found peer over-scoring across all assessment criteria. There were also indications of participant preference of expert feedback over peer feedback, where the later included fewer justifications of quantitative scores and suggestions for change.

Keywords: Integrated STEAM Education, expert assessment, expert feedback, peer assessment, peer feedback, pedagogical design, female engagement in STEAM

Introduction

Peer assessment aims to actively involve peers in employing their knowledge and skills to assess peer work (Cestone et al., 2008; Van Gennip et al., 2010). This may include providing peers with quantitative feedback, for instance, scores across assessment criteria, and/or qualitative feedback, with any justification of scores as well as recommendations for improving peer work (Hovardas et al., 2014; Tsivitanidou et al., 2011). The later would be decisive for letting peer assessees benefit from peer feedback. In education, a quite effective peer assessment format has been the formative/reciprocal one (Tsivitanidou et al., 2011), which engages students in both the roles of peer assessor and peer assessee. Usually this starts with all students undertaking the same set of learning activities to deliver a set of learning products to be assessed. Learning products are any physical or virtual artefacts created by students themselves as they go through a learning activity sequence (Hovardas, 2016; Hovardas et al., 2018). Having created the learning products to be assessed later on, the formative/reciprocal peer assessment procedure should be able to familiarize students with the main requirements and characteristics of the work needed to produce the objects of assessment and shape their background knowledge and skills to be able to act as peer assessors. To better support students in their peer assessor role, a training session is often preceding peer assessment (van Zundert et al., 2010; Xiao & Lucking, 2008). In the peer assessee role, peers screen peer feedback and use it constructively to rework and improve their learning products. The formative/reciprocal peer assessment arrangement lets students gain from multiple reflection processes, for example, when peer assessors compare their own learning products with those of their peers, and when peer assessees are about to rework their learning products taking into account peer feedback (Anker-Hansen & Andree, 2019; Hovardas et al., 2014).

Although peer assessment has been practiced quite often with pre-service teachers (Topping, 2021), there are too few studies engaging pre-service teachers in peer assessment

for pedagogical design[1] (Fang et al., 2021; Lin, 2018; Ng, 2016; Tsai et al., 2002) and, with the exception of Tsai et al. (2002), who reported that peer assessment was not valid across all dimensions studied, no previous study reported either on the validity or the reliability of peer assessment for pedagogical design. What is more, peer assessment has not been yet implemented in pedagogical design for integrated Science, Technology, Engineering, Arts, and Mathematics (STEAM)[2] education (Margot & Kettler, 2019; Thibaut et al., 2018). Integrated STEAM education is understood as the inclusion of at least two STEAM subjects in designing learning activity sequences, whole lesson plans or even projects, with a concentration on real-world problems (Tasiopoulou et al., 2020). Peer assessment would be especially valuable in this case, where teacher collaboration for pedagogical design is indispensable (Margot & Kettler, 2019). STEAM integration seems to be quite demanding and challenging for primary and secondary teachers (Brown & Bogiages, 2019), despite the fact that STEAM education should already presuppose some interdisciplinarity. The silo approach, which compartmentalizes each STEM discipline within its own confines, is still

[1] Pedagogical design begins with planning learning activities, which includes class arrangement (i.e., if activities will be performed by individual students, groups of students or the entire class), the description of learning products, and time needed for students to undertake each activity. Pedagogical design also involves the orchestration of separate activities into sequences of activities, lesson plans or projects. Pedagogical design should align with curriculum standards (e.g., learning objectives, assessment), while it depends on the pedagogical theories and instructional strategies to be chosen (see de Jong et al., 2021).

[2] STEM has been extended to also involve “Arts” (STEAM) and highlight the innovation and creativity of the concept; the “A” in STEAM is interpreted by some scholars as “All”, which wishes to denote the inclusiveness of the approach (Iacovou, 2021). We will refer to “STEM” whenever we present findings of previous research, which also referred to “STEM”.

prevailing in many curricula and in everyday school practice in most educational systems, presenting substantial barriers for promoting integrated STEAM education (Kelly & Knowles, 2016). To address these barriers, pre-service teachers need to be familiarized with good practice in pedagogical design in integrated STEAM education and to work with their peers to design learning activity sequences, lesson plans and projects based on STEAM integration (Hovardas et al., 2020; Tasiopoulou et al., 2020). Using peer assessment for that purpose would allow pre-service teachers develop the competences and mindset needed to provide insightful feedback to peers as well as gain from such input. If peer assessment proves valid and reliable in the context of pedagogical design for integrated STEAM education, and if peer feedback can include constructive input, for instance, justifications for quantitative scores given by peers and suggestions for improving peer work, then it may be instrumental for pre-service teacher training.

Another considerable challenge for pedagogical design in integrated STEAM education, which could be tackled by peer assessment, is female engagement (Zacharia et al., 2020). Previous research has shown that female students in primary education do not differ from their male peers in their attitudes towards STEM (McGuire et al., 2020; Zhou et al., 2019). Moreover, girls in primary education tend to receive higher STEM grades than boys (O'Dea et al., 2018) and tasks related to ICT literacy (Siddiq & Scherer, 2019). It is quite interesting that career beliefs of female students in STEM do not correspond at all to their attitudes and ability in primary education (Sadler et al., 2012; Selimbegović et al., 2019). Indeed, girls do not expect to be as successful as boys in STEM-related careers, which results eventually in fewer girls than boys being interested in pursuing a STEM career at the beginning of high school. This mismatch between female attitudes and performance in STEM, on the one side, and female STEM career beliefs, on the other, is a distinguishing feature in the transition from primary to secondary education and marks female field-specific ability beliefs (Wang & Degol, 2017). What we confront here is a type of "bottleneck

effect", where the overall decrease of students interested in following STEM careers is accompanied by a sharp decrease in the gender diversity of students who still remain interested. This bottleneck effect may be held responsible for any further reduction in female enrolment in STEM subjects and degrees in higher education (Zacharia et al., 2020). It would be crucial to examine if the implementation of peer assessment in pedagogical design for integrated STEAM education could offer input and insight for addressing female engagement. Specifically, qualitative feedback provided by peers can include justification of scores (quantitative part of feedback) and suggestions for improving pedagogical design in this direction. Female engagement will be one of the design dimensions on which we will focus in present study.

Our objective was to implement peer assessment for pedagogical design in integrated STEAM education and to compare expert and peer feedback, in this regard. To our knowledge, this is the first study to investigate if peer assessment can be employed for improving pedagogical design in integrated STEAM education. We engaged pre-service teachers registered in an undergraduate programme for primary education in a formative/reciprocal peer assessment arrangement, where they had the chance to act as both peer assessors and peer assessees. Participants delivered a short but comprehensive pedagogical scenario concentrating on educational robotics, where they had to refer to at least two STEAM subjects, describe a real-world problem to be solved by primary students through thinking critically and creatively, include problem-solving activities for educational robotics, and engage girls as much as boys. Following a training session, participants acted as peer assessors providing quantitative feedback (scores) and qualitative feedback (justification of their scores; suggestions for improving pedagogical scenarios) to their peers. An expert also assessed each pedagogical scenario, and, based on these scores, we awarded badges to a number of participants for recognition of excellence in developing pedagogical scenarios. Moreover, we awarded assessment badges to pre-service teachers based on

deviations of peer assessor scores from expert assessor scores for the same pedagogical scenario.

We first investigated if pre-service teachers were able to respond to expert assessment by improving their pedagogical design (Research question 1). This would provide a solid indication of understanding assessment criteria, grasping the dimensions of pedagogical design and working productively to improve pedagogical scenarios along these dimensions. Then, we examined if peer assessment was valid and reliable (Research question 2). If it was, then it could be exploited in pre-service teacher training for pedagogical design in integrated STEAM education. Our next objective was to compare between expert and peer feedback and outline the weaknesses of peer feedback, if any, for instance, where peer feedback was inferior to expert feedback (Research question 3). This would give us the opportunity to target such weaknesses in training sessions for peer assessment. Finally, we investigated the main determinants that led groups of peer assessees to choose a pedagogical scenario that they would then fully develop into a lesson plan. Here we aimed to explore if performance badges would feature out as significant determinants (Research question 4), which would imply that pre-service teacher training may benefit from exploiting performance badges and letting pre-service teachers use them in their social media and networks.

Methods

Participants

Participants were pre-service teachers (5 males and 20 females) who registered as undergraduate students in the compulsory course “Science Teaching Methods” offered in the fourth semester of the undergraduate programme for primary education at the Department of Education, University of Cyprus. The course content involved a strong component on integrated STEAM education. Participation in the study was part of an assignment given to students, which counted, upon completion of all related activities,

towards 10% of their final mark in the course. Student performance in the assignment did not influence their final grade but, to receive the 10%, they had to submit all deliverables related to the assignment on time. Although all 29 students enrolled in the course agreed to take part in the study, only 25 managed to conclude all tasks and be included in the sample. All participants were guaranteed anonymity. They were informed that their deliverables would be used within the frame of the current study and they provided their informed consent for using them as data sources. Participants were notified that they were free to withdraw at any time from the study, if they felt inclined to do so, without providing any further explanation and without their withdrawal having any impact on the allocation of the 10% of their grade. No participant had any prior experience in peer assessment.

Procedure

Overview

All participants followed an introductory session to the study and a training session on peer assessment, where the first and second authors acted as instructors (see Figure 1 for a presentation of the whole procedure). Participants then developed pedagogical scenarios for integrated STEAM education concentrating on educational robotics. Each scenario was assessed twice by an expert and once by a peer (the second round of expert assessment was accompanied by peer assessment as well). The first round of expert assessment was planned to check if pre-service teachers would respond to expert feedback and improve their scenarios. This would also provide some additional guidance to pre-service teachers in terms of good practice in pedagogical design, concentrating on the first version of the scenarios they delivered. The second round of expert assessment was used to estimate the validity and reliability of peer assessment and investigate differences between expert and peer feedback. Based on expert scores for pedagogical scenarios in the second round, and overlap of peer scores with expert scores, two types of performance badges were granted to a selection of participants, namely, a scenario badge and an assessment badge. Participants

then were randomly assigned to groups and they had to choose one scenario to fully develop into a lesson plan among the ones that group members had already delivered for assessment. The focus here was on whether performance badges were decisive for scenario selection.

Introductory session

In the introductory session, the aim and scope of the study was presented, specifications of participation were discussed and the participants granted their informed consent for the use of the data sources, which will be presented in the next section. Participants were informed that they would take part in a procedure of developing pedagogical scenarios for integrated STEAM education, which would involve two rounds of expert assessment and one round of peer assessment. Each participant would deliver one pedagogical scenario using the GINOBOT for designing learning activities for primary students (https://www.engino.com/w/index.php/products/innolabs-robotics/ginobot). The introductory session included a component of educational robotics focusing on the GINOBOT, the basic functionalities and capabilities of the robot, the KEIRO software for programming the GINOBOT (https://enginoeducation.com/downloads/), and prototype lesson plans concentrating on the GINOBOT. Pedagogical scenarios were meant to be comprehensive descriptions of pedagogical design that should meet four requirements: First, scenarios should address at least two STEAM subjects, which was used as an approach to operationalize integrated STEAM education. Each scenario should describe a real-world problem to be solved by primary students using the GINOBOT in problem-solving activities through thinking critically and creatively. Apart from these requirements, pedagogical scenarios should also seek to engage girls as much as boys to address the gender gap in STEAM education. The introductory session included examples of good practice in pedagogical design for all these dimensions.

Another objective of the introductory session was to familiarize participants with Open Badges, specifically, Open Badge Factory (https://openbadgefactory.com/en/), which is used by competent organizations to create, issue and manage Open Badges, and Open Badge Passport (https://openbadgepassport.com/), where badge owners can obtain and store a pdf certificate of their badge and share it with other users in their social media accounts. Open Badges can be issued for recognizing either intention (e.g., intention to enter a community of practice, intention to communicate a message) or performance (e.g., knowledge, achievements, competences, skills, abilities). They have the form of a digital artefact with malleable visual identity and they carry relevant metadata. Open Badges can be employed in social media accounts to increase visibility of intention or performance of badge owners and shape interaction with other social media users accordingly. All participants created an account in Open Badge Passport and this infrastructure was employed for issuing participant badges for recognizing excelling performance in pedagogical design and peer assessment.

Training session

The training session on peer assessment focused on formative/reciprocal peer assessment. It started with all participants creating an account in HumHub (https://www.humhub.com/en), which was used by the expert assessor and peer assessors to rate pedagogical scenarios and submit expert and peer feedback to peer assessees. Participants rated two different ready-made scenarios provided by the instructors using four assessment criteria, which followed closely the good practice requirements given to participants for developing scenarios:

Criterion 1: The scenario refers explicitly to the STEAM subjects involved;

Criterion 2: The scenario describes a real-world problem to be solved through thinking critically and creatively;

Criterion 3: The scenario includes problem-solving activities with the GINOBOT robot;

Criterion 4: The scenario seeks to engage girls as much as boys.

After rating the first ready-made scenario, participants discussed with instructors their scores and justifications for these scores. A comparison with expert scores followed and an elaboration upon deviations between expert and participant scores concluded that part of the training session. Then, participants rated the second ready-made scenario, and in this case, they were requested to provide justifications for their scores as well as suggestions for changes to improve the scenario. Another round of discussion followed, which involved all above aspects.

Delivery of scenarios, expert assessment, peer assessment, performance badges and student group formation

Each participant delivered one pedagogical scenario, which was first rated by an expert assessor (Senior Research Associate at the University of Cyprus holding a PhD in Science Education and having participated in five European research projects in STEAM education during the last decade). The expert assessor used the same assessment criteria which participants had used in the training session. The expert assessor also provided qualitative feedback to participants with justification of scores across criteria and changes proposed for improving pedagogical scenarios. Reworked scenarios were again assessed by the expert assessor in a second expert assessment round, as well as by participants themselves who acted as peer assessors. Each participant used the same four assessment criteria they had used in the training session to rate two peer pedagogical scenarios chosen randomly and provide qualitative feedback to peer assessees with justification of scores for each assessment criterion and suggestions for changes. The identity of all assessors and asseesees was known to all participants. Excelling performance in pedagogical design (i.e., scenarios with the three highest total expert assessor scores, which belonged to 7 participants) as well as excelling performance in peer assessment (i.e., the three lower

ranked deviations of total peer scores from the expert assessor for the same pedagogical scenario, which belonged to 10 participants) were recognized by being awarded specific badges (scenario badge and assessment badge, respectively). The identity of all pre-service teachers who received badges was known to all peers. Participants then were randomly assigned to groups and selected one pedagogical scenario from those that group members had already submitted for assessment, to further develop it into a lesson plan in integrated STEAM education.

Data sources and coding

Pedagogical scenarios, quantitative scores for each assessment criterion and qualitative feedback (justification of scores; suggestions for changes for improving pedagogical scenarios) provided by the expert assessor and peer assessors in HumHub were the data sources for the study. We coded expert and peer qualitative feedback for items justifying scores and changes proposed for improving pedagogical scenarios. An additional coding process focused on how different STEAM disciplines were integrated in pedagogical scenarios. The first and third author acted as independent coders for 10% of all data. Inter-rater reliability amounted to over 85% and the rest of the cases were resolved after a discussion between coders.

Statistical analyses

We employed non-parametric statistics for all data we collected, since data distributions were non-normal. Specifically, we used Wilcoxon Signed Ranks Tests to ascertain whether size (word count) of pedagogical scenarios provided by participants, total expert assessor scores and expert scores for each assessment criterion differed significantly between the first and second round of expert assessment. These analyses would reflect if participants were able to respond to expert assessment and improve their pedagogical scenarios. To estimate the validity of peer assessment, we computed Spearman's *rho* correlation coefficients for total scores between expert and peer assessors as well as for

scores given for each assessment criterion between expert and peer assessors. Another set of Spearman's *rho* correlation coefficients were computed for total scores and scores for each assessment criterion between the two different peer assessors who were assigned the same pedagogical scenario. This second correlational analysis concentrated on the reliability of peer assessment. Differences in the characteristics of expert and peer feedback, including size (word count) of feedback, scores for each assessment criterion, number of items justifying scores, and number of changes proposed to peer assessees, were examined by means of Mann-Whitney Tests. Finally, we employed tree modeling to investigate if performance badges would be significant determinants for the selection of pedagogical scenarios by peer assessees for developing a lesson plan in integrated STEAM education. In this analysis, we used as independent variables the following parameters: Participants' gender, whether they had been granted a scenario badge and/or an assessment badge, total expert assessor score in the first and second round of expert assessment, the difference in total scores between the first and second round of expert assessment, total peer assessor scores, and the absolute value of the difference in total scores between peer assessors.

Results

Pre-service teacher responsiveness to expert assessment

Average word count of pre-service teachers' pedagogical scenarios increased from the first to the second round of expert assessment from 107.28 to 160.12 words (Wilcoxon Signed Ranks Test $Z = -3.58$, $p < 0.001$). This increase in the size of scenarios was accompanied by an analogous increase in the average value of the total score of the expert assessor from 5.88 (min = 4, max = 9; standard deviation = 1.33), in the first round, to 7.32 (min = 4, max = 10; standard deviation = 1.60) in the second round (Wilcoxon Signed Ranks Test $Z = -3.67$, $p < 0.001$). These results suggest that pre-service teachers, overall, responded to the suggestions of the expert assessor and were able to enrich the descriptions of their pedagogical scenarios and improve their scores. Examining each assessment criterion separately (Table 1), there was

significant improvement of scenarios in three out of four criteria (Criterion 1: The scenario refers explicitly to the STEAM subjects involved, Wilcoxon Signed Ranks Test $Z = -2.71$, $p < 0.01$; Criterion 2: The scenario describes a real-world problem to be solved through thinking critically and creatively, Wilcoxon Signed Ranks Test $Z = -2.89$, $p < 0.01$; Criterion 4: The scenario seeks to engage girls as much as boys, Wilcoxon Signed Ranks Test $Z = -2.83$, $p < 0.01$). For problem-solving activities with the GINOBOT robot (Criterion 3), improvement was not significant. In this case, there was probably a ceiling effect, with the average expert score being already quite high in the first round of expert assessment. We need to highlight the rather low average scores for Criterion 4 ("The scenario seeks to engage girls as much as boys"). Despite the improvement that was recorded, most scenarios failed to effectively address female engagement after the first round of expert assessment.

Validity and reliability of peer assessment

Spearman's *rho* correlation coefficients between total expert scores and total peer scores (global measure of the validity check for peer assessment) as well as between total scores provided by different peer assessors (global measure of the reliability check for peer assessment) revealed that, overall, peer assessment was valid (Spearman's *rho* correlation coefficient = 0.48, $p < 0.001$; $N = 50$) and reliable (Spearman's *rho* correlation coefficient = 0.70, $p < 0.001$; $N = 25$). Spearman's *rho* correlation coefficients for the validity and reliability check for each criterion separately are shown in Table 2. Peer assessment proved to be valid in three out of four assessment criteria (Criterion 2: The scenario describes a real-world problem to be solved through thinking critically and creatively, Spearman's *rho* correlation coefficient = 0.47, $p < 0.001$; $N = 50$; Criterion 3: The scenario includes problem-solving activities with the GINOBOT robot, Spearman's *rho* correlation coefficient = 0.42, $p < 0.01$; $N = 50$; Criterion 4: The scenario seeks to engage girls as much as boys, Spearman's *rho* correlation coefficient = 0.39, $p < 0.01$; $N = 50$). Reliability revealed somewhat worse results, with two out of the four assessment criteria having significant coefficients (Criterion 2: The

scenario describes a real-world problem to be solved through thinking critically and creatively; Spearman's *rho* correlation coefficient = 0.61, $p < 0.01$, $N = 25$; Criterion 3: The scenario includes problem-solving activities with the GINOBOT robot; Spearman's *rho* correlation coefficient = 0.87, $p < 0.001$, $N = 25$). All the above findings indicate that peer assessment did not succeed in providing valid and reliable quantitative feedback across all assessment criteria, despite the training session that pre-service teachers had attended.

Comparison between expert and peer feedback

Average scores for each assessment criterion in expert and peer feedback are presented in Table 3. All scores in peer feedback were higher than expert assessor scores and in three out of four criteria these differences were found to be significant (Criterion 1: The scenario refers explicitly to the STEAM subjects involved, Mann-Whitney $Z = -2.84$, $p < 0.01$; Criterion 2: The scenario describes a real-world problem to be solved through thinking critically and creatively, Mann-Whitney $Z = -2.90$, $p < 0.01$; Criterion 4: The scenario seeks to engage girls as much as boys, Mann-Whitney $Z = -4.79$, $p < 0.001$). The fact that there was no significant difference for Criterion 3 (The scenario includes problem-solving activities with the GINOBOT robot) should be linked to the ceiling effect that was underlined for this criterion in the section on "Pre-service teacher responsiveness to expert assessment" above (see also Table 1, in this regard). Overall, the consistently higher average scores of peers as compared to expert scores may indicate some type of positive bias towards peers.

Differences in average scores (quantitative feedback) combined with difference in feedback size (word count) can help us trace and interpret further differences in the qualitative elements of expert and peer feedback, i.e., items provided for justification of scores and changes proposed to peer assessees for improving their pedagogical scenarios. The size of expert feedback (average word count = 168 words; standard deviation = 27 words) was significantly larger compared to the size of peer feedback (average word count = 91 words; standard deviation = 24 words) (Mann-Whitney $Z = -6.73$, $p < 0.001$). At the same time, the

average number of items justifying scores (Table 4) as well as the average number of changes proposed to peer assessees (Table 5) were, for all assessment criteria, higher in expert feedback as compared to peer feedback. Although peer assessors were able to provide at least one item for justifying their quantitative scores for each assessment criterion, changes proposed to peer assessees were too few, with no change included in peer feedback for Criterion 3 ("The scenario includes problem-solving activities with the GINOBOT robot"). Taken together, the above findings imply that lower average scores across all assessment criteria in expert feedback were accompanied by more items to justify scores and more changes proposed to peer assessees, which led to a relatively increased word count of expert feedback.

Specifically, the average number of items justifying scores was significantly higher in expert feedback for Criterion 1 ("The scenario refers explicitly to the STEAM subjects involved") (Table 4; Mann-Whitney $Z = -3.34$, $p < 0.001$), while the average number of changes proposed to peer assesses was significantly higher in expert feedback for Criteria 3 ("The scenario includes problem-solving activities with the GINOBOT robot") (Table 5; Mann-Whitney $Z = -4.12$, $p < 0.001$) and 4 ("The scenario seeks to engage girls as much as boys") (Table 5; Mann-Whitney $Z = -3.27$, $p < 0.001$). Another interesting finding was that word count in peer feedback tended to increase when peer assessors proposed changes to peer assessees related to female engagement (Criterion 4) (Spearman' *rho* correlation coefficient = 0.37, $p < 0.01$). We computed a crosstabulation and ran a relevant Chi-Square analysis to examine if participants' gender influenced the probability of proposing any changes to peer assessees for improving their scenarios in the criterion for female engagement (Criterion 4). We found that proposing changes for female engagement was neither associated with peer assessor gender nor with peer assessee gender.

Selection of pedagogical scenarios by peer assessees for developing a lesson plan in integrated STEAM education

After receiving expert and peer feedback, peer assessees worked in groups to select one pedagogical scenario among those that group members had already delivered for assessment and process it further to develop a lesson plan in integrated STEAM education. There were three groups with three pre-service teachers and another four groups with four. We employed tree modeling to investigate the effect of several parameters on this selection, including pre-service teachers' gender, whether they had been granted a scenario badge and/or an assessment badge, total expert assessor score in the first and second round of expert assessment, the difference in total scores between the first and second round of expert assessment, total peer assessor scores, and the absolute value of the difference in total scores between peer assessors.

Figure 2 presents the tree computed. At each split, the significant determinants of scenario selection are shown with the values which partitioned the sample at each branch (i.e., there is a left and a right branch in each split). The result of partitioning is depicted at nodes, where one can see the number of scenarios, which were selected or not (n), and the percentage of that number in the total sample. Partitioning is terminated at end nodes. Reading the tree from the top downwards, the first determinant in the first split is whether scenarios had been delivered by pre-service teachers who had been granted a scenario badge. If scenarios belonged to pre-service teachers who had not received such a badge, then these were most probably not selected for developing a lesson plan (first split, left branch, Node 1). Among scenarios delivered by pre-service teachers with a scenario badge (first split, right branch, Node 2), those selected were the ones with a clear improvement measured as difference in total expert assessor scores between the first and second round of expert assessment..

Discussion

The significant correlations computed as global measures of validity (correlations between total scores of expert and peer assessors) and reliability (correlations between total

scores of different peer assessors for the same pedagogical scenario) indicate that peer assessment can be employed in the case of pedagogical design of pre-service teachers in integrated STEAM education. Another strength of peer assessment in our study was that peer assessors were able to include in their feedback to peers at least one item for justifying their quantitative scores in each assessment criterion. The above findings corroborate the few studies available on peer assessment for pedagogical design, according to which, formative peer assessment can improve pedagogical design delivered by pre-service teachers (Fang et al., 2021; Lin, 2018; Ng, 2016; Tsai et al., 2002). There were, however, assessment criteria for which requirements for either validity (STEAM integration) or reliability (STEAM integration; female engagement) were not met. In the case of STEAM integration, there was also a significant difference in items for justifying scores between experts and peers, with the later presenting a lower average. It seems that peer assessors would need much more support and guidance in the training sessions preceding the enactment of peer assessment in order to secure the validity and reliability of their quantitative scores for STEAM integration. This should refer to a concrete anchoring of STEAM disciplines in current curricula as well as a thorough exemplification of possible synergies between STEAM disciplines within the frame of educational robotics involving, for instance, engineering design, programming, and mathematics. Another concern for pre-service teacher training for peer assessment should concentrate on the use of mathematics in integrated STEAM education. As we have seen from an additional qualitative analysis of the pedagogical designs delivered by participants in our study, mathematics were embedded in their designs as simple mathematical operations and not as comprehensive mathematical thinking processes. Analogous weaknesses have been reported in recent research in integrated STEAM education for primary school teachers (Roehrig et al., 2021).

With regard to female engagement, it was quite interesting that reliability for this assessment criterion was not satisfactory despite the fact that a substantial majority of

participants were women. This may imply that there was considerable heterogeneity among female participants in approaches on how to engage female students as well as in judging the effectiveness of these approaches. Female engagement seems to have been the criterion where participants confronted the most challenges in pedagogical design. This criterion had the lowest average expert score in both rounds of expert assessment, and presented the lowest score among criteria for peer assessors as well. Given these shortcomings of pedagogical design for female engagement, and given that there are urgent calls for addressing the gender gap in STEAM (Zacharia et al., 2020), much more attention should be paid for engaging girls as much as boys in pedagogical design for integrated STEAM education. Although several options have been suggested for initiating and sustaining girls' interest in STEAM, such as spatial tools (Moè et al., 2018) and role models (Barabino et al., 2020), not all of them are readily compatible with educational robotics. What is more, the selection of robotic kits for constructing artefacts, which will be the organizing principles of pedagogical design, seems to be quite crucial. A major concern here is that the motive structures, according to which female students operate, do not always overlap with male motivation, especially with regard to speed, power and competition (Johnson, 2003). Although there do not seem to be differences in learning outcomes between boys and girls in educational robotics (Zhong & Xiao, 2020), girls may be more committed to follow teacher instructions (Lindh & Holgersson, 2007; Shih et al., 2012), but for that to happen, girls should first be adequately motivated and engaged. More research will be needed in this direction to support female engagement in integrated STEAM education through pedagogical design.

Average scores for each assessment criterion provided by peers were higher than expert scores. Peer over-scoring is common in peer assessment in higher education (Panadero et al., 2013; Lu & Chiu, 2021). It may be enhanced in the case of female peers, who were the majority in our sample, and who may receive higher scores than male peers,

not due to gender bias, but because female peers may be assumed to perform better than males (Baker, 2008; Falchikov & Magin, 1997; May & Gueldenzoph, 2006; Tucker, 2014). This positive bias needs to be addressed in future training sessions, especially when implementing peer assessment in pedagogical design for integrated STEAM education, since it would detract from the opportunities for improvement, which peer assessment may introduce. Indeed, this was reflected in our study in the difference between expert and peer feedback in the number of changes suggested to peers for improving pedagogical scenarios. An option to address over-scoring may be anonymity of peer assessors and assessees, although it has not always delivered the expected outcomes (Yu & Sung, 2016). For pre-service teachers in primary education, the option of anonymity would probably not contribute to tackling the positive bias since females outnumber their male peers by a wide margin. Anonymity may result in more critical feedback including changes recommended to peers (Howard, 2010; Lin, 2018), but it may severely compromise genuine and constructive peer interaction (Rotsaert et al., 2018). Indeed, it has been found that peer collaboration, when combined with peer assessment, yielded better outcomes as compared to peer assessment alone (Fang et al., 2021). Moreover, training was found to counteract the negative effects of non-anonymous peer assessment (Li, 2017). An option could be to plan a transition from anonymous to non-anonymous peer assessment, which was reported to lead through iterations to equal feedback quality with anonymous peer assessment (Rotsaert et al., 2018). Furthermore, since the concentration on the implementation of specific assessment criteria has not been enough in our study, pre-service teacher training for peer assessment in pedagogical design needs to incorporate a stronger component of the interrelationship between the peer assessor and peer assessee role, e.g., what is expected from peer assessors and what is needed by peer assessees in peer feedback to improve their designs. Reflective focus group discussions among peers may foster this exchange.

The selection process by peer assessees after receiving peer feedback, where they collectively decided which pedagogical scenario to single out and fully develop into a lesson plan, was determined by recognition of excellence in pedagogical design (scenario badge) and improvement in pedagogical design between the two rounds of expert assessment. On the one hand, this finding would imply that pre-service teacher training may benefit from exploiting performance badges and letting pre-service teachers use these badges in their social media and networks. On the other hand, we need to highlight that no aspect of peer assessment was included among the determinants of the tree model, which may imply that peer scores and feedback may not be as valued as much as expert scores and feedback. Previous research showed that pre-service teachers, despite being familiarized through peer discussion and elaboration with peer assessment formats and assessment criteria, may be still dependent upon expert (teacher) advice for the use of assessment criteria (Ng, 2016) or they may still prefer instructor feedback over peer feedback (Seroussi et al., 2019). Such an attitude may have been exacerbated by the female majority of our sample, since female prospective teachers have been found to be more reluctant to give and receive peer feedback than their male peers (Evans & Waring, 2011; Peled et al., 2014). Overall, pre-service teachers may remain ambivalent as to how peer feedback could improve their pedagogical design as long as they lack confidence in their peers' abilities to act as competent assessors. Future research should focus on the potential contribution of peer assessment for empowering pre-service teachers in pedagogical design for integrated STEAM education. Consolidating pre-service teachers' peer assessment skills would support teacher collaboration in formal and informal teacher networks and communities of practice as well as promote distributed leadership.

References

Anker-Hansen, J., & Andrée, M. (2019). Using and rejecting peer feedback in the science classroom: A study of students' negotiations on how to use peer feedback when designing experiments. *Research in Science & Technological Education*, *37*, 346–365. https://doi.org/10.1080/02635143.2018.1557628.

Baker, D. F. (2008). Peer assessment in small groups: A comparison of methods. *Journal of Management Education*, *32*, 183–209. https://doi.org/10.1177/1052562907310489.

Barabino, G., Frize, M., Ibrahim, F., Kaldoudi, E., Lhotska, L., Marcu, L., Stoeva, M., Tsapaki, V., & Bezak, E. (2020). Solutions to gender balance in STEM fields through support, training, education and mentoring: Report of the International Women in Medical Physics and Biomedical Engineering Task Group. *Science and Engineering Ethics*, *26*, 275-292. https://doi.org/10.1007/s11948-019-00097-0.

Brown, R. E., & Bogiages, C. A. (2019). Professional development through STEM integration: How early career math and science teachers respond to experiencing integrated stem tasks. *International Journal of Science and Mathematics Education*, *17*, 111–128. https://doi.org/10.1007/s10763-017-9863-x.

Cestone, C. M., Levine, R. E., & Lane, D. R. (2008). Peer assessment and evaluation in team-based learning. *New Directions for Teaching and Learning*, *116*, 69–78. https://doi.org/10.1002/tl.334.

de Jong,T., Gillet, D., Rodríguez-Triana, M. J., Hovardas, T., Dikke, D., Doran, R., Dziabenko, O., Koslowsky, J., Korventausta, M., Law, E., Pedaste, M., Tasiopoulou, E., Vidal, G., & Zacharia, Z. C. (2021). Understanding teacher design practices for digital inquiry–based science learning: The case of Go-Lab. *Educational Technology Research & Development*, *69*, 417-444. https://doi.org/10.1007/s11423-020-09904-z.

Evans, C., & Waring, M. (2011). Student teacher assessment feedback preferences: The influence of cognitive styles and gender. *Learning and Individual Differences*, *21*, 271–280. https://doi.org/10.1016/j.lindif.2010.11.011.

Falchikov, N., & Magin, D. (1997). Detecting gender bias in peer marking of students' group process work. *Assessment & Evaluation in Higher Education*, *22*, 385–396. https://doi.org/10.1080/0260293970220403.

Fang, J.-W., Chang, S.-C., Hwang, G.-J., & Yang, G. (2021). An online collaborative peer-assessment approach to strengthening pre-service teachers' digital content development competence and higher-order thinking tendency. *Educational Technology Research and Development*, *69*, 1155–1181. https://doi.org/10.1007/s11423-021-09990-7.

Hovardas, T. (2016). A learning progression should address regression: Insights from developing non-linear reasoning in ecology. *Journal of Research in Science Teaching*, *53*, 1447–1470. https://doi.org/10.1002/tea.21330.

Hovardas, T., Tsivitanidou, O., & Zacharia, Z. C. (2014). Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students. *Computers & Education*, *71*, 133–152. http://dx.doi.org/10.1016/j.compedu.2013.09.019.

Hovardas, T., Pedaste, M., Zacharia, Z., & de Jong, T. (2018). Model-based science inquiry in computer-supported learning environments: The case of Go-Lab In: A. K. M. Azad, M. Auer, A. Edwards, & T. de Jong (Eds.), *Cyber-physical laboratories in engineering and science education* (pp. 241–268). Springer, Dordrecht. https://doi.org/10.1007/978-3-319-76935-6_10.

Hovardas, T., Xenofontos, N., Irakleous, M., Pavlou, Y., Kouti, G., & Zacharia, Z. C. (2020). Specifications for learning scenarios and instructional approaches. Deliverable D3.4.

GINOBOT project, Research Promotion Foundation Proposal Number INNOVATE/0719/0098.

Howard, C. D., Barrett, A. F., & Frick, T. W. (2010). Anonymity to promote peer feedback: Pre-service teachers' comments in asynchronous computer-mediated communication. *Journal of Educational Computing Research*, *43*, 89–112. https://doi.org/10.2190/EC.43.1.f.

Iacovou, K. (2021). Literature review on integrated STEAM education. Master Thesis, University of Cyprus, Department of Education.

Johnson, J. (2003). Children, robotics, and education. *Artificial Life and Robotics*, *7*, 16–21. https://doi.org/10.1007/BF02480880.

Kelley, T. R., & Knowles, J. G. (2016). A conceptual framework for integrated STEM education. *International Journal of STEM Education*, *3*, 11. https://doi.org/10.1186/s40594-016-0046-z.

Li, L. (2017). The role of anonymity in peer assessment. *Assessment & Evaluation in Higher Education*, *42*, 645–656. https://doi.org/10.1080/02602938.2016.1174766.

Lin, G.-Y. (2018). Anonymous versus identified peer assessment via a Facebook-based learning application: Effects on quality of peer feedback, perceived learning, perceived fairness, and attitude toward the system. *Computers & Education*, *116*, 81–92. http://dx.doi.org/10.1016/j.compedu.2017.08.010.

Lindh, J., & Holgersson, T. (2007). Does lego training stimulate pupils' ability to solve logical problems? *Computers & Education*, *49*, 1097–1111. https://doi.org/10.1016/j.compedu.2005.12.008.

Lu, M., & Chiu, M. M. (2021). Do teamwork guidelines improve peer assessment accuracy or attitudes during collaborative learning? *IEEE Transactions on Education*, https://doi.org/10.1109/TE.2021.3130242.

Margot, K. C., & Kettler, T. (2019). Teachers' perception of STEM integration and education: a systematic literature review. *International Journal of STEM Education*, *6*, 2. https://doi.org/10.1186/s40594-018-0151-2.

May, G. L., & Gueldenzoph, L. E. (2006). The effect of social style on peer evaluation ratings in project teams. *Journal of Business Communication*, *43*, 4–20. https://doi.org/10.1177/0021943605282368.

McGuire, L., Mulvey, K. L., Goff, E., Irvin, M. J., Winterbottom, M., Fields, G. E., Hartstone-Rose, A., & Rutland, A. (2020). STEM gender stereotypes from early childhood through adolescence at informal science centers. *Journal of Applied Developmental Psychology*, *67*, 101109. https://doi.org/10.1016/j.appdev.2020.101109.

Moè, A., Jansen, P., & Pietsch, S. (2018). Childhood preference for spatial toys. Gender differences and relationships with mental rotation in STEM and non-STEM students. *Learning and Individual Differences*, *68*, 108-115. https://doi.org/10.1016/j.lindif.2018.10.003.

Ng, E. M. V. (2016). Fostering pre-service teachers' self-regulated learning through self- and peer assessment of wiki projects. *Computers & Education*, *98*, 180–191. http://dx.doi.org/10.1016/j.compedu.2016.03.015.

O'Dea, R. E., Lagisz, M., Jennions, M. D., & Nakagawa, S. (2018). Gender differences in individual variation in academic grades fail to fit expected patterns for STEM. *Nature Communications, 9*, 1–8. https://doi.org/10.1038/s41467-018-06292-0.

Panadero, E., Romero, M., & Strijbos, J. (2013). The impact of a rubric and friendship on peer assessment: effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation*, *39*, 195-203. https://doi.org/10.1016/j.stueduc.2013.10.005.

Peled, Y., Bar-Shalom, O., & Sharon, R. (2014). Characterisation of pre-service teachers' attitude to feedback in a wiki-environment framework. *Interactive Learning Environments*, *22*, 578–593. https://doi.org/10.1080/10494820.2012.731002.

Roehrig, G., Dare, E., Ring-Whalen, E., & Wieselmann, J. (2021). Understanding coherence and integration in integrated STEM curriculum. *International Journal of STEM Education*, *8*, 2. https://doi.org/10.1186/s40594-020-00259-8.

Rotsaert, T., Panadero, E., Schellens, T. (2018). Anonymity as an instructional scaffold in peer assessment: Its effects on peer feedback quality and evolution in students' perceptions about peer assessment skills. *European Journal of Psychology of Education*, *33*, 75–99.

Sadler, P. M., Sonnert, G., Hazari, Z., & Tai, R. (2012). Stability and volatility of STEM career interest in high school: A gender study. *Science Education*, *96*, 411–427. https://doi.org/10.1002/sce.21007.

Selimbegović, L., Karabegović, M., Blažev, M., & Burušić, J. (2019). The independent contributions of gender stereotypes and gender identification in predicting primary school pupils' expectancies of success in STEM fields. *Psychology in the Schools*, *56*, 1614–1632. https://doi.org/10.1002/pits.22296.

Seroussi, D.-E., Sharon, R., Peled, Y., & Yaffe, Y. (2019). Reflections on peer feedback in disciplinary courses as a tool in pre-service teacher training. *Cambridge Journal of Education*, *49*, 655-671. https://doi.org/10.1080/0305764X.2019.1581134.

Shih, B. Y., Chang, C. J., Chen, Y. H., Chen, C. Y., & Liang, Y. D. (2012). LEGO NXT information on test dimensionality using Kolb's innovative learning cycle. *Natural Hazards*, *64*, 1527–1548. https://doi.org/10.1007/s11069-012-0318-y.

Siddiq, F., & Scherer, R. (2019). Is there a gender gap? A meta-analysis of the gender differences in students' ICT literacy. *Educational Research Review*, *27*, 205–217. https://doi.org/10.1016/j.edurev.2019.03.007.

Tasiopoulou, E., Myrtsioti, E., Niewint Gori, J., Xenofontos, N., Hovardas, T., Cinganotto, L., Anichini, G., Garista, P., & Gras-Velazquez, A. (2020). STE(A)M IT – An interdisciplinary STEM approach. Integrated STEM teaching State of Play. European Schoolnet, Brussels. http://steamit.eun.org/files/D2.1_STEAM_IT_State_of_play_final.pdf.

Thibaut, L., Knipprath, H., Dehaene, W., & Depaepe, F. (2018). How school context and personal factors relate to teachers' attitudes toward teaching integrated STEM. *International Journal of Technology and Design Education*, *28*, 631–651. https://doi.org/10.1007/s10798-017-9416-1.

Topping, K. J. (2021). Digital peer assessment in school teacher education and development: a systematic review. *Research Papers in Education*. https://doi.org/10.1080/02671522.2021.1961301.

Tsai, C.-C., Lin, S. S. J., & Yuan, S.-M. (2002). Developing science activities through a networked peer assessment system. *Computers & Education*, *38*, 241–252. https://doi.org/10.1016/S0360-1315(01)00069-0.

Tsivitanidou, O., Zacharia, Z. C., & Hovardas, T. (2011). Investigating secondary school students' unmediated peer assessment skills. *Learning and Instruction*, *21*, 506–519. https://doi.org/10.1016/j.learninstruc.2010.08.002.

Tucker, R. (2014). Sex does not matter: Gender bias and gender differences in peer assessments of contributions to group work. *Assessment & Evaluation in Higher Education*, *39*, 293–309. https://doi.org/10.1080/02602938.2013.830282.

van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2010). Peer assessment as a collaborative learning activity: The role of interpersonal variables and conceptions. *Learning and Instruction*, *20*, 280–290. https://doi.org/10.1016/j.learninstruc.2009.08.010.

van Zundert, M., Sluijsmans, D. M. A., & Van Merriënboer, J. J. G. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction*, 20, 270–279. https://doi.org/10.1016/j.learninstruc.2009.08.004.

Wang, M. T., & Degol, J. L. (2017). Gender gap in science, technology, engineering, and mathematics (STEM): Current knowledge, implications for practice, policy, and future directions. *Educational Psychology Review*, *29*, 119–140. https://doi.org/10.1007/s10648-015-9355-x.

Xiao, Y., & Lucking, R. (2008). The impact of two types of peer assessment on students' performance and satisfaction within a Wiki environment. *Internet and Higher Education*, *11*, 186–193. https://doi.org/10.1016/j.iheduc.2008.06.005.

Yu, F.-Y., & Sung, S. (2016). A mixed methods approach to the assessor's targeting behavior during online peer assessment: effects of anonymity and underlying reasons, *Interactive Learning Environments*, *24*, 1674-1691. https://doi.org/10.1080/10494820.2015.104140.

Zacharia, Z. C., Hovardas, T., Xenofontos, N., Pavlou, I., & Irakleous, M. (2020). Education and employment of women in science, technology and the digital economy, including AI and its influence on gender equality. Policy Department for Citizens' Rights and Constitutional Affairs, European Parliament (Report prepared at the request of the FEMM Committee, Policy Department for Citizens' Rights and Constitutional Affairs, Directorate-General for Internal Policies). https://www.europarl.europa.eu/RegData/etudes/STUD/2020/651042/IPOL_STU(2020)651042_EN.pdf.

Zhong, B., & Xia, L. (2020). A systematic review on exploring the potential of educational robotics in mathematics education. *International Journal of Science and Mathematics Education*, *18*, 79–101. https://doi.org/10.1007/s10763-018-09939-y.

Zhou, S. N., Zeng, H., Xu, S. R., Chen, L. C., & Xiao, H. (2019). Exploring changes in primary students' attitudes towards science, technology, engineering and mathematics (STEM) across genders and grade levels. *Journal of Baltic Science Education*, *18*, 466. https://doi.org/10.33225/jbse/19.18.466.

Table 1

Mean scores for pedagogical scenarios for each assessment criterion in the two rounds of expert assessment

	First round	Second round	Wilcoxon Signed Ranks Test
			Z
Criterion 1: The scenario refers explicitly to the STEAM subjects involved	1.68 (0.63)	2.04 (0.54)	-2.71**
Criterion 2: The scenario describes a real-world problem to be solved through thinking critically and creatively	1.28 (0.61)	1.84 (0.80)	-2.89**
Criterion 3: The scenario includes problem-solving activities with the GINOBOT robot	1.84 (0.80)	2.04 (0.68)	-1.51ns
Criterion 4: The scenario seeks to engage girls as much as boys	1.08 (0.28)	1.40 (0.50)	-2.83**

Note: Each criterion was scored by the expert assessor along a three-point Likert-scale (1 = not addressed at all; 2 = partially addressed; 3 = fully addressed); standard deviations are given in parentheses; ns = non-significant; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table 2

Spearman's rho correlation coefficients for the validity and reliability check of peer assessment for each assessment criterion

	Validity check (between expert assessor and peer assessors; *N* = 50)	Reliability check (between first and second peer assessor, *N* = 25)
Criterion 1: The scenario refers explicitly to the STEAM subjects involved	0.14^{ns}	0.33^{ns}
Criterion 2: The scenario describes a real-world problem to be solved through thinking critically and creatively	0.47^{***}	0.61^{**}
Criterion 3: The scenario includes problem-solving activities with the GINOBOT robot	0.42^{**}	0.87^{***}
Criterion 4: The scenario seeks to engage girls as much as boys	0.39^{**}	0.45^{ns}

Note: ns = non-significant; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table 3

Average scores for each assessment criterion in expert and peer feedback

	Average scores in expert feedback	Average scores in peer feedback	Mann-Whitney *Z*
Criterion 1: The scenario refers explicitly to the STEAM subjects involved	2.04 (0.54)	2.46 (0.65)	-2.84**
Criterion 2: The scenario describes a real-world problem to be solved through thinking critically and creatively	1.84 (0.80)	2.42 (0.74)	-2.90**
Criterion 3: The scenario includes problem-solving activities with the GINOBOT robot	2.04 (0.68)	2.46 (0.74)	-2.53ns
Criterion 4: The scenario seeks to engage girls as much as boys	1.40 (0.50)	2.40 (0.79)	-4.79***

Each criterion was scored by the expert assessor and peer assessors along a three-point Likert-scale (1 = not addressed at all; 2 = partially addressed; 3 = fully addressed); standard deviations are given in parentheses; ns = non-significant; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table 4

Average number of items justifying scores in expert and peer feedback

	Average number of items justifying scores in expert feedback	Average number of items justifying scores in peer feedback	Mann-Whitney *Z*
Criterion 1: The scenario refers explicitly to the STEAM subjects involved	1.28 (0.46)	1.02 (0.14)	-3.34***
Criterion 2: The scenario describes a real-world problem to be solved through thinking critically and creatively	1.20 (0.50)	1.13 (0.33)	-0.47ns
Criterion 3: The scenario includes problem-solving activities with the GINOBOT robot	1.16 (0.37)	1.06 (0.24)	-1.33ns
Criterion 4: The scenario seeks to engage girls as much as boys	1.20 (0.50)	1.04 (0.20)	-1.76ns

Note: Standard deviations are given in parentheses; ns = non-significant; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table 5

Average number of changes proposed in expert and peer feedback for improving pedagogical scenarios

	Average number of changes proposed in expert feedback	Average number of changes proposed in peer feedback	Mann-Whitney *Z*
Criterion 1: The scenario refers explicitly to the STEAM subjects involved	0.16 (0.47)	0.08 (0.28)	-0.55^{ns}
Criterion 2: The scenario describes a real-world problem to be solved through thinking critically and creatively	0.48 (0.59)	0.19 (0.39)	-2.34^{ns}
Criterion 3: The scenario includes problem-solving activities with the GINOBOT robot	0.48 (0.77)	0.00 (0.00)	-4.12^{***}
Criterion 4: The scenario seeks to engage girls as much as boys	0.44 (0.51)	0.10 (0.31)	-3.27^{***}

Note: Standard deviations are given in parentheses; ns = non-significant; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Figure 1

Procedure: After an introduction session to the study and a training session for peer assessment, pre-service teachers developed a first version of pedagogical scenarios in integrated STEAM education with a focus on educational robotics. A first round of expert assessment followed and pre-service teachers reworked their scenarios. This second version of pedagogical scenarios were subjected to a second round of expert assessment and peer assessment. Based on expert and peer scores, a selection of pre-service teachers were awarded performance badges (scenario badge; assessment badge). Pre-service teachers then formed groups and selected one scenario among the ones already delivered by group members to fully develop into a lesson plan.

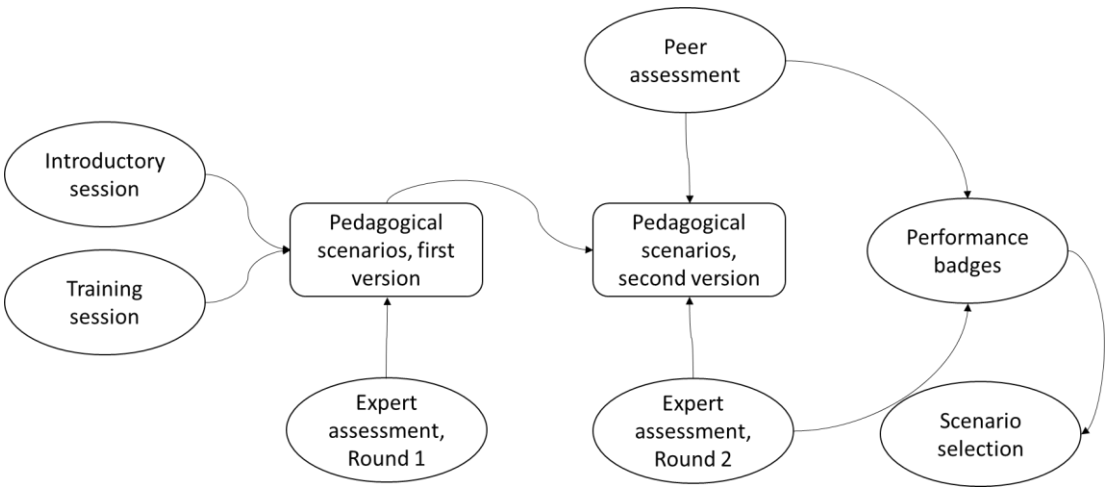

Figure 2

Tree model for selection of pedagogical scenarios by peer assessees to develop a lesson plan in integrated STEAM education. Significant determinants are shown at each split with thresholds for partitioning the sample at left and right branches. Each node depicts the number of scenarios selected or not (n) and their percentage in the total sample. Overall percentage of cases correctly classified = 92.0%.