International Journal on
"Technical and Physical Problems of Engineering"
(IJTPE)
Published by International Organization of IOTPE

ISSN 2077-3528
IJTPE Journal
www.iotpe.com
ijtpe@iotpe.com

December 2024 | Issue 61 | Volume 16 | Number 4 | Pages 253-265

ENHANCING VALIDITY AND RELIABILITY IN ONLINE PEER ASSESSMENTS: IMPACT OF TRAINING-REGULATION AND GENDER

Y. Rahmani B. Nachit M. Bassiri

Multidisciplinary Laboratory in Educational Sciences and Training Engineering (LMSEIF), Higher Normal School, Hassan II University, Casablanca, Morocco
yassine.rahmani@enscasa.ma, nachitbrahim@yahoo.fr, bassiri.mustapha@gmail.com

Abstract- The quality of student assessments is a concern, particularly in an environment where traditional assessment methods are prevalent. Despite the compelling evidence supporting the efficacy of online peer assessment, its widespread use for summative purposes in higher education has not kept pace. This study seeks to investigate the impact of training, regulation, and gender on the validity and reliability of online peer assessment. 139 first-year physical education and sports teaching bachelor students with varying levels of training were asked to evaluate their peers' projects In a Moodle platform, based on an evaluation grid. We analyzed the correlation between students and instructor assessments to report validity, and we used the intra-class correlation to quantify reliability among students' assessments. The experiment results showed that training significantly improved both the validity and reliability of student peer assessments. Gender was not found to significantly influence validity, underscoring the dominant role of training and regulation. Students in trained groups produced more consistent and valid assessments. However, the involvement of untrained evaluators negatively affected reliability. The results obtained highlight the importance of structured training and regulation for effective peer assessment. Teachers should prioritize hands-on practice and provide ongoing feedback to help students develop objective assessment skills and a common understanding of its criteria.

Keywords: Validity, Reliability, Online Peer Assessment Design, ICT.

1. INTRODUCTION

Online peer assessment (OPA) is an impactful pedagogical approach that targets the proliferation of many high-level transferable skills and prepares the learner for lifelong learning. By engaging in OPA, students not only master the course content, but also develop independent learning, the ability to evaluate their work, effective communication, and the understanding of multiple points of view. This seemingly simple process, complex in implementation and management, effectively improves various necessary skills for success in the 21st century. Studies have demonstrated that integrating digital platforms into assessment processes not only enhances learning outcomes but also fosters the development of critical skills like self-assessment and collaborative learning [1]. Several meta-analyses report that the majority of instructors opt for OPA for formative purposes only [2], [3], due to concerns about students' ability to evaluate their peers accurately and fairly [4].

One of the main hindrances to the adoption of OPA for summative purposes is the ability of students to evaluate peer work validly and reliably. This depends on several factors, namely, students' understanding of quality, standards, and attitudes, as well as their ability to make evaluative judgments [5]. It is important to note that the bias related to peer assessment also exists in assessments carried out by the teacher in a more or less explicit way. The literature that has addressed the issues of validity and reliability of OPAs provides some guidelines to improve the quality of these two parameters, such as the use of subject-specific analytical grids supplemented by examples and training [6], the involvement of a sufficient number of assessors in the assessment of an artifact, and the anonymity of assessments [7]. These measures can be effective when students have previous experiences with peer assessment, but in our case, where the education system adopts traditional teaching and assessment methods, simple exemplification or tutorials appear insufficient to lead students to conduct a valid and reliable assessment.

Nonetheless, the possible impact of demographic characteristics, including gender, on the quality of OPA has garnered limited consideration in the current literature [8]. Research examining the influence of gender on peer assessment is limited, creating a significant void in comprehending the extent to which gender disparities may affect the Quality of evaluations. In Morocco, traditional assessment governs the assessment practices of teachers in higher education, and the use of alternative or digitalized assessment practices in the classroom remains very rare [9].

Registered: 270709 | Article Info: 1921 | Accepted: 241127
Reserved: 241006 | Available: 241231

Therefore, achieving good validity and reliability of OPAs must be based on a rational typology and must consider the specificities of the context, yet adopting quality and measurement standards and ensuring adequate support and training for students during all phases of OPA.

The target group of our study was first-year students in the teaching of physical education and sports enrolled in the Information and Communication Technology in Teaching module 'TICE1'. All of these students had never had an OPA experience and 77% had never heard of this assessment strategy. To overcome this problematic situation and study the impact of training and practice on the validity and reliability of OPAs, we marked out the TICE1 module with three peer assessments. The first two were considered as a training phase and had a low weighting in the total grade of the continuous assessment, while the third had a considerable weighting in the summative assessment of the module and served as material for our study.

This paper attempts to add a study of the validity and reliability of peer assessment in a context where traditional assessment methods still govern, as well as to provide an in-depth description of the implementation of an OPA in light of the design elements contained in the literature, while paying special attention to training and practice in OPA. Additionally, this study investigates whether gender differences influence the validity of peer assessments and how these differences interact with training exposure.

Through this research we attempt to answer the following research questions:

- RQ1: To what extent does the repetition of training OPAs improve the validity of peer assessments?
  - H0: The validity of OPAs is not dependent on training and education.
  - H1: Training significantly impacts the validity and reliability of OPAs.
- RQ2: Does gender influence the validity of peer assessments in online learning contexts, and does this effect vary based on training exposure?
- RQ3: Does the reliability of assessments increase in groups that are subjected to training in peer assessments?

2. LITERATURE REVIEW

To theoretically frame our research, we attempted to define three main concepts: "Peer assessment", "Peer assessment validity" and "Peer assessment reliability".

2.1. Peer Assessment

The concept of "peer assessment" emerged from early discussions of collaborative learning that led to the exploration and expansion of its theories and applications, Piaget's work on cognitive construction inspired Vigotsky in developing his theories, emphasizing how learners can acquire knowledge and skills more effectively by interacting with each other. However, the term "peer assessment" did not emerge until the work of K. Topping [10], although Topping noted earlier examples of PA in higher education contexts. Most research now gives him credit for establishing the systematic foundations of peer assessment. Building on these theoretical foundations, recent studies have contributed to the development of new methodologies for the application of OPA supported by the development of e-learning and OPA authoring tools, and also by increasingly numerous empirical evidence that confirms its pedagogical contribution in various educational contexts.

Over the last two decades, educators in developed countries have widely adopted peer assessment, especially for formative purposes, in various contexts and disciplines. Recent studies have shown that the use of OPA as an assessment device throughout the course has been effective in helping learners improve their problem-solving skills, their effective communication skills [11], in addition to their motivation and their self-regulation skills [12].

While much of the research on OPA has focused on general educational contexts, its application in professional engineering education has garnered increasing attention. Studies in this domain highlight its potential to enhance not only technical problem-solving and analytical skills but also essential professional competencies such as teamwork, effective communication, and self-directed learning.

The majority of peer assessment activities in the engineering field focus on soft skills rather than technical skills. In addition, some published research in the fields of science and engineering focuses on social science tasks, such as writing reports or essays [13]. However, OPA of technical skills is equally critical in engineering education, as it allows students to assess and identify errors in the design of manual and computer-based documentation and calculations of engineering projects [14].

The introduction of OPAs has encouraged, for example, a project-based chemical engineering course including peer-assessment methodologies that demonstrated improved learning outcomes and instructional quality, as indicated by student survey data [15]. Similarly [16] established a peer assessment assignment within a training methodology for Finite Element Modelling (FEM). This study compared test scores from students in a traditional FEM course with those in an innovative, peer-assessment-based approach. Results indicated that the innovative methodology not only improved learning outcomes but also significantly increased student engagement compared to the traditional format.

Despite these benefits across various fields of education, students may feel an initial anxiety about this process, which requires support from teachers to further have them immersed in the process, as well as a set of procedures to increase the validity and reliability of OPA which remain the major problems related to this assessment method.

2.2. Peer Assessment Validity

An assessment is considered valid if the items that constitute it allow one to assess the degree of achievement of the objectives of a training [17]. In OPA, validity concerns both the assessment instrument and the score awarded by the peer assessor.

The instrument designed for an OPA should be subjected to an in-depth analysis of its validity before it can be used on a large scale. It is necessary to identify the psychometric characteristics of the assessment instruments used because the results of instruments with insufficient validity cannot subsequently lead to a correct comparison. To ensure the validity of the instrument De Katele proposes to carry out a comparative analysis of the items of the same test by at least three experts or teachers by adopting an internal empirical validation through the calculation of the homogeneity coefficient [18]. The validity construct of OPA can be determined by comparing peers' scores to the teachers. The following procedure is often used in the literature to quantify the validity of OPAs from the teacher's perspective. First, the teacher or the expert should provide a trustworthy rating or feedback called the "gold standard" in the literature. Second, peer rating validity is calculated as the correlation coefficient between student ratings and instructor ratings for the same work [19], this coefficient should not be influenced by distribution patterns.

Validity from the student's perspective SPv is expressed by the difference between the rating of a peer or a set of peers and that of the teacher, to know whether the peer rater under- or over-evaluates compared to the gold standard [5]. In a summative context, it is necessary to achieve a high level of validity to truly reflect the degree of achievement of the learning objectives.

2.3. Peer Assessment Reliability

In general, the reliability of assessment concerns the consistency of the results obtained from several assessments of the same product; different synonyms are used such as fidelity, reproducibility, or stability [20]. Reliability is well- understood if it is linked to the variability of measurements and sources of error. For example, the instructions for rating items in an assessment grid can be interpreted differently from one peer to another or from one peer to an expert, the smaller the gap between these interpretations, the more we can qualify this assessment as reliable, because it reproduces the same results each time regardless of the assessor. There are several types of reliability, the most commonly treated is intra-rater reliability which refers to the correlation of results obtained from two or more assessments carried out by the same person with the same measuring instruments of the same product at different times. We also find reliability or internal consistency which concerns the internal consistency and homogeneity of the items constituting the test, for example the different aspects of the development of a skill.

To measure inter-rater reliability, we propose that several evaluators evaluate the same performance or skill of the same student based on the same evaluation grid and then we check whether they produce identical or distinct measures. This can be done using various statistical methods.

The theory of generalizability proposes an in-depth theoretical development of reliability. It is defined as the proportion of the variance of an observed score that is not attributable to measurement errors [21]. This involves detailing and estimating the variance of the true score, the variance of the error score, and the components of the variance of the observed score, while allowing the calculation of coefficients based on these estimates. This approach proposes the adoption of the intraclass correlation coefficient (ICC) to determine the agreement between two or more measurements taken in a time interval [22], because it highlights systematic measurement biases and also verifies the temporal stability of the scores.

The reliability of peer assessment depends on several factors such as the context, the level of the course, the performance assessed, the clarity of the assessment criteria, and the training and support provided. Reliability also tends to improve at higher levels of study, as students at this stage are likely to be more cognitively advanced and potentially have higher-order thinking skills than those taking introductory courses.

Despite significant progress in improving the validity and reliability of peer assessments through strategies such as structured rubrics, rigorous training, and statistical measures, challenges still exist, particularly in large-scale online implementations. These include differences in evaluator judgment, inconsistencies in feedback quality, and challenges in assuring alignment with targeted learning goals. To address these limitations, the integration of artificial intelligence (AI) in OPA has emerged as a transformative solution, offering innovative tools to standardize evaluations, improve grading accuracy, identify biases, and support assessors in providing consistent, high-quality feedback.

2.4. AI-Assisted Peer Assessment

Research on the integration of AI to optimize the quality of OPAs is still in its early stages. Generally, researchers focus on using natural language processing (NLP) AI assistance to enhance the quality of feedback, ensure it aligns with the assessment task, and enhance the reliability of grades provided by peer assessor. Similarly, the utilization of ChatGPT has proven beneficial. Research demonstrates that the integration of AI enables the creation of more thorough and constructive feedback. It also enhances the reliability of peer-assigned notes by implementing inference mechanisms, outperforming traditional models that rely on the average or median of notes assigned to the same assignment [23]. The development of large language models (LLMs) adept at understanding and generating human language has opened up possibilities for effectively monitoring and evaluating the qualitative comments provided by student reviewers on peer essays. However, there is little discussion about using AI to validate student grades for validity; most experts in the field rely on computer-assisted calibration processes to verify students' ability to initiate a valid OPA [24].

3. MATERIAL AND METHODOLOGY

3.1. Design of the Study

This study was designed as a quasi-experimental investigation to explore the impact of training sessions on the validity and reliability of online peer assessments (OPA) in a higher education context, with gender analysis specifically applied to the study of validity. We sought to determine whether participation in training sessions improves the validity and reliability of student assessments when compared to a teacher's and peer's evaluations. The research framework involved multiple assessments conducted in a hybrid learning environment, combining face-to-face instructional components with online peer assessment activities.

Participants were first-year Bachelor of Education students specializing in physical and sports at the Normal High School of Tetuan in Morocco (n = 139). The students were enrolled in the TICE1 module, which focuses on developing ICT integration skills in teaching Physical and Sports Education. The population was divided into three distinct groups based on their participation in the training sessions, as detailed below.

At the start of the academic year, students were randomly assigned to one of three sections within the TICE1 module. This randomization ensured that no prior academic performance (baccalaureate grades or access interview results) influenced group assignment. These sections provided a foundation for later groupings in the study, ensuring that the participants reflected a diverse, unbiased sample of the cohort.

To host the online assessment system, a MOODLE platform was created for this purpose, three peer assessments were scheduled, the first two were worth 5% of the continuous assessment grade each and they are considered as training sessions. The third OPA was the subject of this study, it represented 25% of the continuous assessment grade. A space dedicated to assistance was created and students had the opportunity to contact the platform administration by chat to resolve any difficulties they might encounter during the various assessment phases. During this assessment, students had to solve a problem encountered in teaching physical education and sports using ICT.

Submitted projects required students to propose solutions to real-life teaching challenges using ICT tools. The project tasks were aligned with the learning objectives of the TICE1 module and assessed key competencies. Before the start of each OPA session, the assessment grids were presented and clarified in class and a simulation of the OPA phases in Moodle was made available on the platform. After the end of each assessment workshop, a regulation session was scheduled, and samples of assessed projects and their assessments were presented anonymously. The teacher mentions their strengths and weaknesses as well as the biases to avoid during the evaluation. During the third peer evaluation, each student was invited to submit their project on the platform and evaluate 4 peer projects. The projects were distributed randomly and anonymously. All submitted projects were re-evaluated by the teacher before the end of the evaluation phase; subsequently, an exhaustive list of the marks given to each project by the teacher and the peer evaluator was established.

Each evaluation was marked out of 100, 80 points for the project, and 20 points for the grading process. The mark reserved for the grading was automatically assigned by the platform. To ensure the reliability and validity of the evaluation grids used in the study, a multi-step process was implemented. Two teachers independently developed separate evaluation grids, each containing detailed items and associated scores. These grids were then consolidated into a single four-item grid through iterative refinement.

To validate the grid, the two teachers independently evaluated 10 sample projects from previous student cohorts using the consolidated grid. This step served two purposes:

1)Testing Consistency of Interpretation: by having the teachers apply the grid independently, the process highlighted potential ambiguities or inconsistencies in how the evaluation criteria were understood and applied.

2) Assessing Inter-Rater Reliability: The inter-rater reliability of the scores assigned to each item was assessed by calculating the intraclass correlation coefficient (ICC). A strong ICC indicated that the teachers interpreted and applied the criteria consistently, while a moderate ICC revealed items requiring further clarification or reformulation.

Three items demonstrated strong agreement between the teachers, confirming their reliability. However, the fourth item produced a moderate ICC, necessitating revision. This item concerned the integration of a video component within a presentation software. One teacher considered a hyperlink to a video sufficient to meet the criterion, while the other required the video to be embedded directly within the presentation. To address this discrepancy, the item was reformulated to eliminate subjective interpretations. The revised criterion explicitly stated that the video component would be considered present if included as either a hyperlink to an online video or an embedded video file within the presentation.

Following this revision, a second round of assessments was conducted with the updated grid, achieving strong ICC values for all items. This iterative process ensured that the grid's criteria were clearly defined, unambiguous, and consistently interpretable, thereby enhancing its reliability for use in the study.

3.2. Data Extraction and Analysis

To investigate the validity of peer assessments and to form the study groups, data were collected using the Moodle platform. Moodle Workshop activity does not natively support exporting detailed peer assessment data. To address this limitation, the grades assigned by each peer assessor to their assigned projects were manually recorded in an Excel file. This process involved carefully copying the data for each peer evaluation from the platform and ensuring the accurate pairing of assessors with the projects they evaluated. The groups were formed according to gender and the frequency of participation in the training assessments, so we formed 3 groups (Table1).

Table 1. Group characteristics and training participations

Croups	Students	Description	Gender Composition	
Group 1	68	Students who participated in both training sessions	24 Men	44 Women
Group 2	46	Students who participated in one training session	18 Men	28 Women
Group 3	25	Students who did not participate in any training sessions	15 Men	10 Women

To study the construct of validity, we proceeded in three steps.

1) To evaluate the validity of the assessments from the teacher's perspectives (TPv), the scores attributed by each student to the four assigned projects were compared to the scores attributed by the teacher to the same projects, using the Pearson correlation coefficient [4].

2) We compared the assessments of each group to those of the teacher using the Spearman- Rho correlation.

3) To determine the impact of training on the validity of the assessments of each group, we performed the nonparametric Kruskal-Wallis H test because the validity scores were ordinal and the data did not meet the assumptions of normality.

When significant differences were detected, the Dunn post hoc test was applied for pairwise comparisons to identify specific differences between the groups, allowing a detailed analysis of the impact of training sessions on assessment validity.

To examine the potential impact of demographic variables on the validity of online peer assessments, gender was included as an independent variable in this study. Gender was recorded for each participant and analyzed in combination with the number of training sessions attended. boxplots were constructed to visually compare the distributions of validity scores across gender and training groups. The Mann-Whitney U test, was performed to evaluate whether significant differences in validity existed between men and women within each training group. These analyses aimed to identify whether gender differences influenced the validity of peer assessments in the context of varied training exposure.

To determine and interpret the inter-rater reliability, we referred to the guideline of selecting and reporting intraclass correlation coefficients for reliability research [25]. We calculated the intraclass correlation coefficient by comparing the ratings given by a set of students to the same projects. Subsequently, the students were divided into groups according to their group memberships.

Table 2. Organization of groups for the reliability study

groups	cases	description
Booth from 1	4 cases	students involved in the correction of peer projects from group 1
Booth from 2	4 cases	students involved in the correction of peer projects from group 2
From 2 and 1	6 cases	students involved in the correction of peer projects from groups 2 and 1
From 2 and 3	2 cases	students involved in the correction of peer projects from groups 2 and 3
From 3 and 1	2 cases	students involved in the correction of peer projects from groups 3 and 1

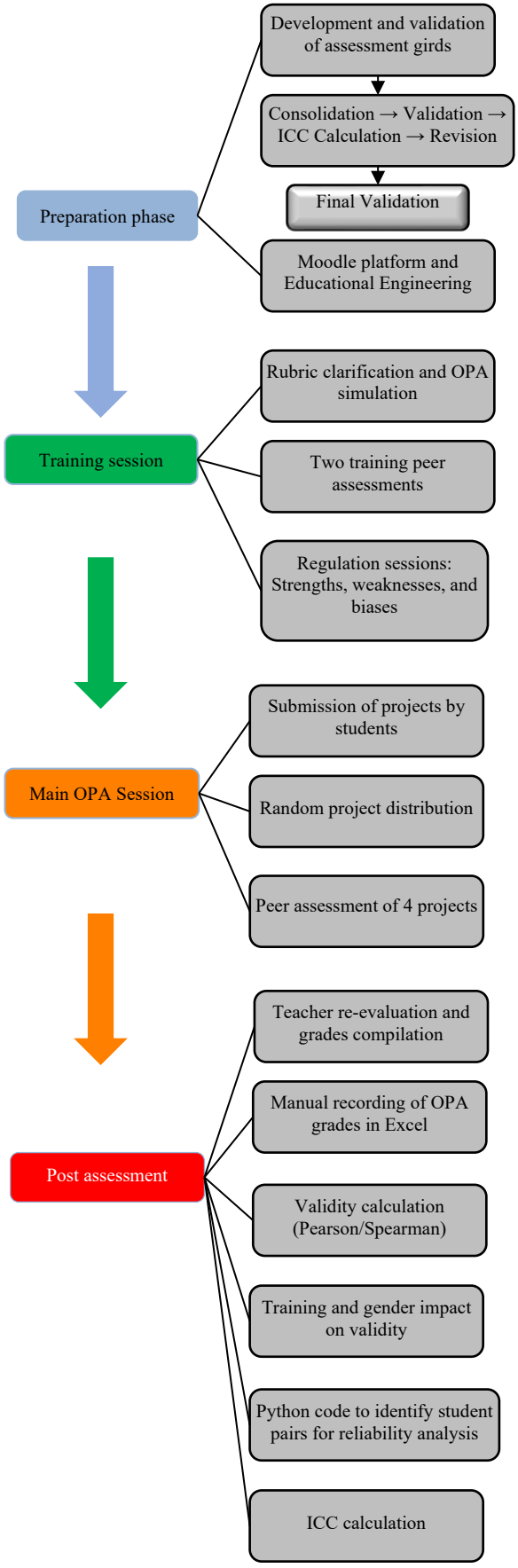

Figure 1. Study procedure flow

The projects assigned for evaluation were distributed randomly by the platform, so we developed a Python code (Appendices 1) to address the need for identifying and matching students who assessed the same projects, as the Moodle platform assigned projects randomly. The script automated the detection of evaluator; we were able to detect 36 matches distributed as Table 2.

Since the choice of peer evaluators was made in a randomized manner by the platform from the same population, we opted for the calculation procedure according to ICC guidelines: ICC estimates for each student were calculated using SPSS v 26 software based on an absolute agreement type mean score (k=2) and a two-factor mixed effect module. The overall methodology of this study, including group formation, training interventions, and the analysis process, is summarized in the flowchart in Figure 1.

4. RESULTS

The following three subsections address each research question respectively: the first subsection analyzes the construct of validity. The second examines whether gender differences influence the validity of peer assessments and how these differences interact with training exposure, while the third highlights the construct of inter-rater reliability.

3.3. Results of the Correlation Tests

First, we compared the evaluations of each student for the four assigned projects to those of the teacher Figure 2, the results showed a strong correlation among students in group $rgoup1$=0.91(SD=0.14), a moderate correlation was observed in group $rgoup2$=0.65(SD=0.21), the correlation value continued to decrease in the third group, we had a weak correlation in this group $rgoup3$=0.32(SD=0.23). the boxplot below showed a few outliers, particularly in the "trained twice" group. These outliers suggest that while most students achieved strong validity scores, individual differences may persist, potentially due to variations in prior knowledge, engagement with training sessions, or task complexity.

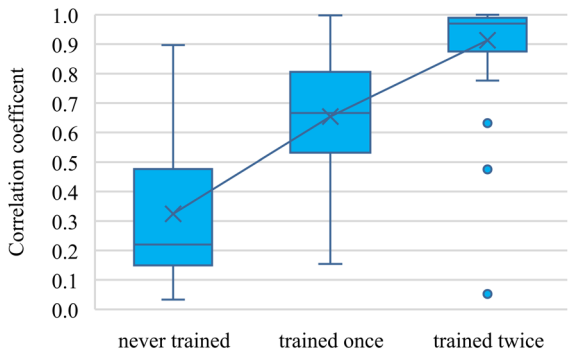

Figure 2. Correlation coefficient scores by training group: validity results

These correlation values highlight the role of training in clarifying assessment criteria and aligning student assessments with gold standards. Repeated training consolidates these gains, demonstrating its critical role in producing valid evaluations. Secondly, we compared the assessments produced by each group for each project to the teacher's evaluations, as shown in Table 3. The results presented below correspond to each project and each group.

Table 3. Spearman correlation between students and teacher ratings for each project

Groups	Projects correlation	Correlation coefficient	Sig. (2-tailed)
Trained twice	SAP1-TAP1	0.817**	0.000
	SAP2-TAP2	0.720**	0.000
	SAP3-TAP3	0.860**	0.000
	SAP4-TAP4	0.850**	0.000
Trained once	SAP1-TAP1	0.501**	0.000
	SAP2-TAP2	0.628**	0.000
	SAP3-TAP3	0.701**	0.000
	SAP4-TAP4	0.544**	0.000
Never trained	SAP1-TAP1	0.478	0.014
	SAP2-TAP2	0.373	0.066
	SAP3-TAP3	0.491	0.013
	SAP4-TAP4	0.120	0.568
** Correlation is significant at the 0.01 level (2-tailed). Legend: SAP: student assessment for the project TAP: teacher evaluation for the project			

For group 1, we observed strong correlations for the four projects, as well as the absence of significant differences between peer and teacher evaluations ($rp1$=0.817, p<0.001) ($rp2$=0.720, p<0.001) ($rp3$=0.860, p<0.001) ($rp4$=0.850, p<0.001). For the second group, the correlations were moderate in general with no significant difference between the scores attributed by the students and those of the teacher ($rp1$=0.501, p<0.001) ($rp2$=0.628, p<0.001) ($rp3$=0.701, p<0.001) ($rp4$=0.504, p<0.001).

In the third group, a weak correlation was observed, with statistically significant differences between the scores generated by the students of this group and the scores attributed by the teacher to each project ($rp1$ = 0.478, p = 0.014) ($rp2$ = 0.373, p = 0.066) ($rp3$ = 0.491, p = 0.013) ($rp4$ = 0.12, p = 0.568). These findings suggest that repeated training reinforces students' evaluative consistency across various projects. Significant differences between student and teacher grades in untrained group underscores the challenges of unguided peer assessment, while the higher consistency in trained groups demonstrates the efficacy of structured preparation.

4.1. Non-Parametric Analysis of Training Impact on Peer Assessment Validity

Finally, to detect the impact of training on the validity of peer assessments, we conducted the Kruskal-Wallis non-parametric test. This test revealed significant differences between the study groups in terms of the validity of the assessments and a large effect of training on the validity of the assessments $H(2) = 78.77$, $p < 0.001$, $\eta = 0.791$. These results led to the rejection of hypothesis H0 and the retention of hypothesis H1.

To explore differences between groups in terms of the validity of the assessments, a Dunn's post hoc test was performed. Pairwise comparisons of the "never trained", "trained once", and "trained twice" groups revealed significant differences. The results of the pairwise comparisons are presented in Table 4.

Table 4. Analysis results of Dunn's Post Hoc Test of assessments validity for three groups

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
never trained-trained once	-30.7	10.0	-3.07	0.002	0.006
never trained-trained twice	-76.8	9.4	-8.1	0.000	0.000
trained once-trained twice	-46.05	7.6	-5.9	0.000	0.000
Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. The significance level is 0.01. a. Significance values					

P values were adjusted using Bonferroni correction to account for multiple comparisons. Results indicated that there were significant differences between the validity of the assessments of each pair of groups (never trained-trained once p=0.006), (never trained-trained twice p<0.001), (trained once-trained twice p<0.001).

These results emphasize that even a single training session leads to measurable improvements in validity, while repeated training maximizes it, ensuring assessments are both valid and consistent. The large effect size (η=0.791) underscores the critical and transformative role of structured and iterative training in fostering high-quality peer assessments, particularly in academic settings requiring precise evaluation skills.

4.2. Impact of Gender on the Validity of OPA

First, we compared the validity scores for different training groups and genders (Figure 3). Men have a slightly higher median validity score than women in the "Never Trained" group, although there is substantial variety for both genders. In the "Trained Once" group, women have higher median validity than men, with less variability than the prior group.

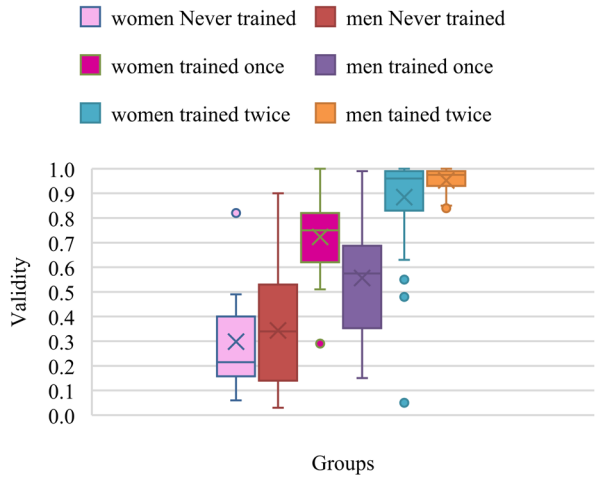

Figure 3. Validity of peer assessments across training groups and gender

To statistically evaluate the potential impact of gender, a Mann-Whitney U test was conducted (Table 5).

Finally, in the "Trained Twice" group, the median scores for both genders converge, and variability is minimal, indicating greater consistency in assessments after repeated training.

Table 5. Mann-Whitney U test comparing validity between men and women assessors

Test Statistics[a]	
	Validity
Mann-Whitney U	1958.000
Wilcoxon W	3611.000
Z	-1.623
Asymp. Sig. (2-tailed)	0.105
a. Grouping Variable: Gender	

For the impact of gender on the validity of OPA, we found that there were no significant differences between groups U=1958.00, Z=−1.623, p =0.105. The p value (p >0.05), suggests that the differences in median correlation scores are not due to an inherent gender effect. The negative Z value supports the conclusion that gender has no substantial impact on the validity of OPA. The analysis of gender revealed that training benefits both men and women equally. We observed slight variations in median scores, with men outperforming women in the "never trained" group and women surpassing men in the "trained once" group;

these differences were negligible and diminished with increased training. In the "trained twice" group, the scores converged, demonstrating the equalizing effect of repeated training. These results suggest that training reduces variability and ensures consistency across genders, emphasizing the inclusivity of structured peer assessment programs. Gender does not appear to influence the validity of assessments, highlighting the effectiveness of training as an overriding factor.

4.3. Inter-Rater Reliability Test Results

When both raters belong to group 1, we were able to achieve good reliability and low variability between raters ICC = 0.84 SD = 0.082. When raters belong to group 2, we also witnessed good reliability and a slight increase in variability ICC = 0.77 SD = 0.10. Reliability continues to decrease when raters are from group 1 and 2, we had moderate reliability ICC = 0.54 SD = 0.16. When a student from group 3 is involved in the assessment a drastic drop in reliability was observed. An outlier was found when one student belongs to group 2 and the other to group 3, we found an ICC=0.79, inspecting the validity of the evaluations of the two students we found that they both had r= 0.55, which partially explains the high reliability observed in this case. This convergence in evaluation patterns likely contributed to the observed high reliability, as reliability measures focus on consistency between assessors rather than alignment with teacher evaluations.

The boxplot (Figure 4) provides the ICC variability across the different assessor pairings. Groups with students from the same group (bothfrom1 and bothfrom2) exhibit lower variability, as evidenced by the smaller interquartile range (IQR) and whiskers. However, group pairings involving assessors from different groups (from 2 and 1, from 2 and 3, and from 3 and 1) show increased variability, particularly in from2and3, where the IQR is notably larger, indicating inconsistent reliability across assessments.

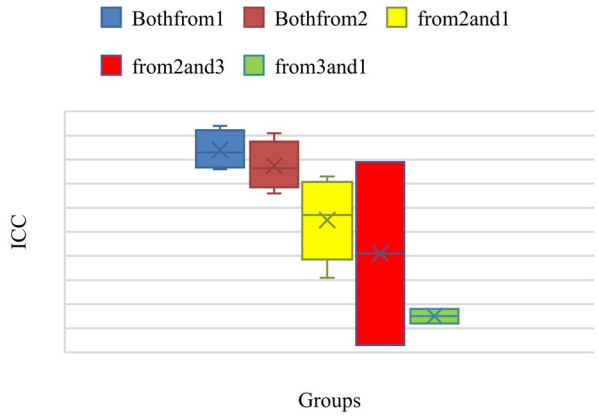

Figure 4. ICC across rater group pairings

The validity and reliability results agree to demonstrate the impact of training on the quality of peer evaluations. Students in group 1, having received two training sessions, show both high validity and reliability in their evaluations. On the other hand, those in group 3, without training, present less reliable and less valid evaluations.

5. DISCUSSION

This study examined the validity and reliability of OPAs in an educational context where traditional assessment practices are predominant. We provided a detailed description of the implementation of OPAs in our context and then explored the two key aspects, the validity and reliability of student assessments. We first explored the impact of training and instruction on the validity of OPAs by testing two hypotheses. The results of the correlation tests demonstrated that the validity construct from the teacher's perspective progressively improved with the repetition of training. The students who participated in this study evaluated all the projects using the same assessment grid in the same digital environment.

They all attended the regulation sessions and had access to the simulations of the OPA phases, however, the validity of the assessments was not the same in the study groups. The validity construct measured either individually or in groups improves with the repetition of the training assessments. Comparing the validity of assessments between groups of students we observed a clear impact of training sessions on the validity of peer assessments, leading to the retention of hypothesis H1. Secondly, we examined the effect of gender on the validity of OPAs alongside the impact of training. The findings revealed no statistically significant differences in validity scores between men and women assessors, indicating that gender does not play a discernible role in improving validity in this context. These results underscore that training, assessment design, and rubric clarity are far more critical factors in ensuring the quality of OPA than demographic variables. The observed gender differences in validity diminished significantly after multiple training sessions. This highlights that well-structured training interventions can effectively standardize assessment practices and mitigate any initial variability in validity, regardless of gender.

Finally, we explored the reliability of OPAs in relation to participation in training assessments. The ICCs aligned with the validity results, highlighting the positive impact of training on this aspect of OPA. Students from the trained groups produced more consistent assessments, however, the involvement of an element belonging to group 3 in the assessment of a project affected the reliability, leading to assessments of low reliability.

It is important to note that ICC calculations are sensitive to group dynamics and composition. Variability within groups, as well as the inclusion of less trained or untrained assessors, can introduce biases that influence reliability estimates. The stability of ICC values may be impacted by smaller sample sizes in subgroup analysis, this study employed a two-way mixed-effects model, ensuring absolute agreement while accounting for rater-specific tendencies. Even in the presence of potential ICC biases, our findings underscore the critical role of training in reducing variability and enhancing consistency in OPA.

Studies conducted in developed countries have reported that the use of detailed and well-developed evaluation rubrics alongside exemplification can improve the validity of peer assessments [26], [27], the latter study reported good validity without resorting to training, as the students involved were already accustomed to online peer assessment. However, the results of our study conducted in a less developed educational system, where students have less experience in ICT and have never had a peer assessment experience, demonstrated that the untrained group produced assessments of low validity, with significant differences between the study groups.

Several studies have reported the presence of training sessions and/or tutoring in the design of their OPA methods. This was often provided in the form of exemplification by the teacher [28] or discussion and mutual development of the evaluation grids by the students [29]. The correlation coefficient reported in both studies was (r=0.7). However, the repetition of the training sessions and the on-hands approach adopted in our study allowed the achievement of significant validity r=0.91 in group 1. These findings have crucial implications for educational contexts where students lack familiarity with ICT tools and OPA practices.

Another study conducted by Formanek, et al., on astronomy MOOC , in which a group of students followed a single round of training (assessment of a set of assignments and regulation sessions) gave a moderate correlation between the students' and the teacher's grades r=0.65 [30], which is consistent with the correlation coefficient observed in group 2 (those who participated in a single training assessment, r=0.65). The iterative assessment cycles and teacher feedback in our approach were decisive in achieving higher validity levels (r=0.91) in the "trained twice" group. In teaching contexts where students lack prior exposure to OPAs or ICT tools, exemplification and evaluation grids alone are insufficient to guarantee valid assessments. Repeated training remains indispensable for enhancing peer assessment quality and ensuring accurate, reliable evaluations.

In addition to these findings, comparisons with studies in engineering education provide further insights into the validity of OPAs, particularly in contexts involving diverse and complex assessment tasks. A study was conducted with first-year undergraduate engineering students in a professional skills course, in which they were asked to assess their peers' work based on clear and structured evaluation criteria [31]. Students evaluated tasks of varying complexity, ranging from formal email writing to multimedia development. This study investigated the validity of peer grades by comparing them with lecturer grades. The resulting Pearson correlation coefficients (ranging from 0.564 to 0.824) indicated moderate to high validity, suggesting that peers could provide evaluations that closely aligned with those of instructors, even when assessing tasks of differing complexity. These results align with the validity levels found in our study, especially for the "trained twice" and "trained once" groups, who demonstrated moderate to high validity in their assessments. However, the authors of this study did not specify gender of participants, whether they received prior training in OPA or were already accustomed to the process.

The literature rarely addresses how gender affects the validity of peer assessment. There is currently little and conflicting information about gender bias and gender differences in peer assessment. For example [32], explore the effects of gender and training on peer scoring, in this study, students were trained to correct their peers' essays in psychology on the Eduflow platform. authors focused on the validity from the student's perspective, i.e., whether the student significantly overperformed or underperformed the essays compared to the teacher's assessment. The researchers found that gender had no effect on peer scoring accuracy and that training is the decisive element in optimizing the validity of OPAs, which is consistent with the findings of the present research.

The results of previous studies on the reliability of online peer assessments are mixed, some studies mentioned that students who have already practiced OPA several times become more reliable, because they acquire the ability to evaluate based on the evaluation criteria provided and they finally manage to distinguish between high- and low-quality work, to finally become evaluators close to the professional [33]. However, other studies have mentioned that training did not have a significant impact on reliability, for example Zakharov and al adopted a calibration during which students were asked to evaluate examples of essays already marked by the teacher and reach a certain level of correlation to begin peer assessment [34]. This procedure did not subsequently result in making student assessments more reliable, the authors mentioned a passable reliability despite the presence of training.

The answer to the mixed results presented in empirical studies may lie in our research, students must have the same assessment skills and should have the closest possible understanding of the assessment criteria to have a coherent interpretation. An important finding to highlight is that the involvement of a student from group 3 in the assessment, drastically reduces reliability, which can lead to erroneous assessments and lead the assessed peers to overestimate or underestimate their work. One of the strengths of the OPA lies in the readjustment of learning carried out in light of the feedback received, however when an assessment provides unfair indicators to the person being assessed, the desired learning readjustments will not be done correctly.

Finally, we should note some limitations in this study. First, the sample size for reliability analysis, particularly in subgroup comparisons, was relatively small. This may influence the stability of intraclass correlation coefficients and introduce variability into the results. Furthermore, despite measures taken to ensure a balanced design, potential biases in the random assignment of projects to students may have affected validity and reliability results. OPAs may exhibit inconsistencies due to variations in task complexity or familiarity with the project topics.

To address these gaps, future research should test the proposed methodology in larger, more diverse educational settings. Expanding the sample size would enhance the statistical robustness of reliability analyses and provide more generalizable insights. Moreover, implementing the methodology across various disciplines and educational systems, including those with differing levels of ICT integration, would help validate its applicability.

6. IMPLICATIONS FOR TEACHERS USING OPA

This study demonstrates the importance of training associated with practice in optimizing the validity and reliability of OPAs. Teachers who work in an educational context where the integration of new technologies remains limited and traditional assessment practices dominate, should allocate special attention to training and supporting students before launching OPA workshops.

Training should include:

- practical Exercises when Students engage in iterative assessment tasks using real examples, receiving feedback from teachers to refine their evaluation practices.
- Regulation sessions that include guided discussions on the assessment criteria and its results, to promote a common understanding and reduce variability. During this phase, it is necessary to ensure that the examples presented are anonymous, as this is essential to avoid any potential anxiety among students.
-Teacher Feedback Integration: Alternating training with regulation sessions allows students to adjust their evaluation practices in light of teacher guidance, improving the validity and consistency of peer assessments.

Instructors should not be satisfied with tutorials, exemplification and explanation of assessment grids; the adoption of a hands-on approach remains crucial. Training helps students to:

1) become accustomed to producing objective assessments based only on the assessment criteria provided, which leads to improved OPA validity;

2) develop a common understanding of both the procedure and the interpretation of the evaluation criteria, which helps to optimize the reliability of the OPAs;
3) alternating training and regulation sessions allows the student to readjust their evaluation practices in light of the teacher's guidance.

These approaches are especially beneficial in contexts where students lack prior experience with peer assessments or familiarity with ICT tools. Repeated training ensures that students not only adapt to the digital platforms used for assessments but also align their evaluations with professional standards.

the workload for the teacher to ensure the quality of the OPAs appears a little heavy, however the integration of AI can provide valuable support. For example, the workshop activity in MOODLE assigns a grade to the evaluation procedure. This grade is based on the degree of consensus between the evaluators involved in the evaluation of the same work, which implicitly provides an index of reliability, including AI technology into the OPA platforms could greatly improve this procedure. AI has the potential to identify trends in evaluator behavior, identify biases or discrepancies in real time, and dynamically modify grade computations to more accurately represent the quality of peer assessments. Additionally, AI could help educators by recognizing and displaying assessments that demonstrate poor reliability, allowing for prompt corrections to fix discrepancies and strengthen the assessment process as a whole.

Pending the incorporation of AI into the OPA process, teachers should inspect the low marks of the assessment procedure. To facilitate this process, the Python code we have developed offers a practical solution by enabling quick matching of assessors who evaluated the same work. This tool facilitates the eventual calculation of reliability indices through statistical software, providing teachers with immediate insight into the consistency of peer assessments. Our study has shown that low reliability is associated with low validity of one of the assessors, this inspection will allow teachers to rectify the marks attributed to the student assessed and provide personalized support to students who have not yet developed a reliable assessor attitude.

It is important to note that the implementation of OPAs in resource-limited educational contexts requires the training of teachers in digital literacy and assessment practices. Sometimes the lack of adequate training in ICT and digital tools can significantly hamper the effective use of digital learning environments and related assessment devices [35].

The results of our study can have practical implications for training design in massive open online courses (MOOCs), where maintaining assessment quality at scale is a challenge. MOOC developers should integrate structured training and calibration methodologies such as pre-assessment training modules with interactive tutorials and examples to prepare students for the OPA process. AI-assisted feedback mechanisms can provide immediate and real-time feedback on student assessments, effectively correcting errors. For scalability, adaptive learning algorithms or peer-led training sessions can personalize the training experience to meet the diverse needs of learners. By integrating these strategies, MOOC platforms can ensure that OPAs remain effective learning tools, promoting reflective practices and improving assessment validity across diverse learner populations.

Finally, OPA is a relevant teaching and learning strategy that allows students to receive immediate and abundant feedback from peers and to engage in reflective feedback on their achievements and performances. These benefits are dependent on the quality of peer assessments. However, when these assessments are erroneous, reflective feedback is not done correctly.

APPENDICES

Appendix 1. The Python Code Used to Detect Similarities in Assigned Assessments

```
import pandas as pd
import os

# Get the directory of the current script
current_dir = os.path.dirname(os.path.abspath(__file__))

# Construct the path to the input file in the parent
directory
input_file = os.path.join(current_dir, '..', 'input.xlsx')
output_file = os.path.join(current_dir, '..', 'output.xlsx')


# Function to identify rows with common students and
group them
def find_and_group_common_students(df):
    correspondences = []
    evaluators = df.iloc[:, 0]  # Get the names of the
evaluators

    for i in range(len(df)):
        students_i = set(df.iloc[i, 1:].dropna())  # Get
students for evaluator i
        for j in range(i + 1, len(df)):
            students_j = set(df.iloc[j, 1:].dropna())  # Get
students for evaluator j
            common_students = students_i & students_j
            if len(common_students) >= 2:
                correspondences.append((i+2, evaluators[i],
j+2, evaluators[j], common_students))  # Add i+2 and j+2
to compensate for index 0 and header

    grouped_by_count = {}
    for line_number1, eval1, line_number2, eval2,
common_students in correspondences:
        key = f"{len(common_students)} correspondences"
        if key not in grouped_by_count:
            grouped_by_count[key] = {}
        common_students_key =
str(sorted(list(common_students)))
        if common_students_key not in
grouped_by_count[key]:
```

```
            grouped_by_count[key][common_students_key]
= []

grouped_by_count[key][common_students_key].append(
(line_number1, eval1))

grouped_by_count[key][common_students_key].append(
(line_number2, eval2))

    # Convert lists to sets to remove duplicates, then sort
them back into lists
    for count_key in grouped_by_count:
        for common_students_key in
grouped_by_count[count_key]:

grouped_by_count[count_key][common_students_key] =
sorted(set(grouped_by_count[count_key][common_stude
nts_key]))

    return grouped_by_count

# Read the input Excel file
df = pd.read_excel(input_file)

# Call the function to find and group correspondences
grouped_correspondences =
find_and_group_common_students(df)

# Create the output Excel file using pd.ExcelWriter
with pd.ExcelWriter(output_file) as writer:
    # Write the initial content in the first sheet
    df.to_excel(writer, sheet_name='Input Data',
index=False)

    # Add sheets for each group of correspondences with
the names of the sheets
    for count_key in
sorted(grouped_correspondences.keys(), key=lambda x:
int(x.split()[0]), reverse=True):
        # Create an empty DataFrame for the group
        group_data = pd.DataFrame(columns=['Label'])  #
Initialize empty DataFrame with a 'Label' column

        # Add "Correspondence" as a static title
        correspondence_row =
pd.DataFrame([["Correspondence"]], columns=['Label'])
        group_data = pd.concat([group_data,
correspondence_row], ignore_index=True)

        # Prepare the text to write in the sheet
        for common_students_key in
grouped_correspondences[count_key]:
            common_students_list =
eval(common_students_key)
            new_row = pd.DataFrame([["" + ',
'.join(common_students_list)]]).rename(columns={0:
'Label'})  # Adjusted to match the structure

            group_data = pd.concat([group_data, new_row],
ignore_index=True)

        for line_number, _ in
grouped_correspondences[count_key][common_students
_key]:
            line_data = df.iloc[line_number -
2:line_number - 1].reset_index(drop=True)  # Reset
index to avoid conflicts
            group_data = pd.concat([group_data,
line_data], ignore_index=True)
        empty_row = pd.DataFrame([[""] *
len(df.columns)])  # Add an empty row for spacing
        group_data = pd.concat([group_data, empty_row],
ignore_index=True)

    # Write the group data to the new sheet
    group_data.to_excel(writer, sheet_name=count_key,
index=False, header=False)
```

REFERENCES

[1] R. Hamzaoui, R. El Ayachi, B. Jabir, M. El Mohadab, B. Bouikhalene, "Digital Assessment in Training of Future Teachers: State of Use and Impact on Trainee Performance", International Journal on Technical and Physical Problems of Engineering (IJTPE), Vol. 16, No. 59, pp. 288-294, Jun 2024.

[2] H. Li, Y. Xiong, X. Zang, M. L. Kornhaber, Y. Lyu, K. Chung, H.K. Suen, "Peer Assessment in the Digital Age: a Meta-Analysis Comparing Peer and Teacher Ratings", Assessment and Evaluation in Higher Education, Vol. 41, No. 2, pp. 245-264, February 2016.

[3] K.J. Topping, "Digital Peer Assessment in School Teacher Education and Development: A Systematic Review", Research Papers in Education, Vol. 38, No. 3, pp. 472-498, May 2023.

[4] E. Panadero, M. Romero, J.W. Strijbos, "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, Vol. 39, No. 4, pp. 195-203, December 2013.

[5] D. Nicol, "Resituating Feedback from the Reactive to the Proactive", Feedback in Higher and Professional Education, pp. 34-49, January 2013.

[6] M. Kobayashi, "Does Anonymity Matter? Examining Quality of Online Peer Assessment and Students Attitudes", Australasian Journal of Educational Technology, Vol. 36, No. 1, Art, January 2020.

[8] E. Panadero, M. Alqassab, "An Empirical Review of Anonymity Effects in Peer Assessment, Peer Feedback, Peer Review, Peer Evaluation and Peer Grading", Assessment and Evaluation in Higher Education, Vol. 44, No. 8, pp. 1253-1278, November 2019.

[9] B. Abderrahmane, "Alternative Assessment and English Language Teaching and Learning in Morocco: High School Teachers Perceptions and Favourite Methods and Techniques", Education and Training Paths, Vol. 2, No. 2, Art, March 2019.

[10] K. Topping, "Peer Assessment Between Students in Colleges and Universities", Review of Educational Research, Vol. 68, No. 3, pp. 249-276, 1998.

[13] Y. Demiraslan Cevik, "Assessor or Assesses? Investigating the Differential Effects of Online Peer

Assessment Roles in the Development of Students' Problem-Solving Skills", Computers in Human Behavior, Vol. 52, pp. 250-258, November 2015.
[11] R.L. Hulsman, J.F. Peters, M. Fabriek, "Peer-Assessment of Medical Communication Skills: The Impact of Students Personality, Academic and Social Reputation on Behavioural Assessment", Patient Educ Couns, Vol. 92, No. 3, pp. 346-354, September 2013.
[12] H. Li, Y. Xiong, C.V. Hunter, X. Guo, R. Tywoniw, "Does Peer Assessment Promote Student Learning? A Meta-Analysis", Assessment and Evaluation in Higher Education, Vol. 45, No. 2, pp. 193-211, February 2020.
[13] G. Naveh, D. Bykhovsky, "Online Peer Assessment in Undergraduate Electrical Engineering Course", IEEE Transactions on Education, Vol. 64, No. 1, pp. 58-65, February 2021.
[14] F. Javidan, "An Online Peer Assessment Method in Computational-Based Engineering Courses: Combining Theoretical and Computer Tools", REES AAEE 2021 Conference: Engineering Education Research Capability Development, pp. 811-818, Melbourne, Australia, March 2022.
[15] E. Cifrian, A. Andres, B. Galan, J.R. Viguri, "Integration of Different Assessment Approaches: Application to a Project-Based Learning Engineering Course", Education for Chemical Engineers, Vol. 31, pp. 62-75, April 2020.
[16] P.L. Bishay, "Teaching the Finite Element Method Fundamentals to Undergraduate Students Through Truss Builder and Truss Analyzer Computational Tools and Student-Generated Assignments Mini-Projects", Computer Applications in Engineering Education, Vol. 28, No. 4, pp. 1007-1027, 2020.
[17] L. Bouzidi, A. Jaillet, "Can Online Peer Assessment be Trusted?", Journal of Educational Technology and Society, Vol. 12, No. 4, pp. 257-268, 2009.
[18] J.M. De Ketele, F.M. Gerard, "Validation of Assessment Tests Using the Skills-Based Approach", MEE, Vol. 28, No. 3, pp. 1-26, 2005.
[19] H. Luo, A.C. Robinson, J.Y. Park, "Peer Grading in a MOOC: Reliability, Validity, and Perceived Effects", Online Learning, Vol. 18, No. 2, Art., Jun 2014.
[20] M. Bertrand Leiser, N. Kuhne, "Chapter 6 - Standardized Measuring Instruments and Their Metrological Qualities", Practical Guide to Rehabilitation Research, in Intervention Methods, Techniques and Tools, Louvain-la-Neuve: De Boeck Superior, pp. 115-131, 2014.
[21] J. Correa Rojas, "Intraclass Correlation Coefficient: Applications to Estimate the Temporal Stability of a Measurement Instrument", Psychological Sciences, Vol. 15, No. 2, January 2024.
[22] P.E. Shrout, J.L. Fleiss, "Intraclass Correlations: Uses in Assessing Rater Reliability", Psychol Bull, Vol. 86, No. 2, pp. 420-428, March 1979.
[23] A. Darvishi, H. Khosravi, S. Sadiq, D. Gasevic, "Incorporating AI and Learning Analytics to Build Trustworthy Peer Assessment Systems", British Journal of Educational Technology, Vol. 53, No. 4, pp. 844-875, May2022.
[24] L. Brkic, I. Mekterovic, M. Fertalj, and D. Mekterovic, "Peer Assessment Methodology of Open-Ended Assignments: Insights from a Two-Year Case Study Within a University Course Using Novel Open-Source System", Computers and Education, Vol. 213, Article 105001, May 2024.
[25] T.K. Koo, M.Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research", J Chiropr Med, Vol. 15, No. 2, pp. 155-163, Jun 2016.
[26] M. Alqassab, J.W. Strijbos, E. Panadero, J. Fernandez Ruiz, M. Warrens, J. To, "A Systematic Review of Peer Assessment Design Elements", Educational Psychology Review, Vol. 35, February 2023.
[27] F. Zhang, C. Schunn, W. Li, M. Long, "Changes in the Reliability and Validity of Peer Assessment Across the College Years", Assessment and Evaluation in Higher Education, Vol. 45, No. 8, pp. 1073-1087, November 2020.
[28] C. Guler, "Use of WhatsApp in Higher Education: What's Up with Assessing Peers Anonymously?", Journal of Educational Computing Research, Vol. 55, No. 2, pp. 272-289, April 2017.
[29] A. Raes, E. Vanderhoven, T. Schellens, "Increasing Anonymity in Peer Assessment by Using Classroom Response Technology Within Face-To-Face Higher Education", Studies in Higher Education, January 2015.
[30] M. Formanek, M.C. Wenger, S.R. Buxner, C.D. Impey, T. Sonam, "Insights About Large-Scale Online Peer Assessment from an Analysis of an Astronomy MOOC", Computers and Education, Vol. 113, pp. 243-262, October 2017.
[31] J. Petrovic, P. Pale, "Exploring Usage of Summative Peer Assessments in Engineering Education", The Towards a New Future in Engineering Education, New Scenarios That European Alliances of Tech Universities Open Up, pp. 2146-2150, Universitas Polytechnical de Catalunya, Spain, September 2022.
[32] J.C.G. Ocampo, E. Panadero, F. Diez, "Are Men and Women Really Different? The Effects of Gender and Training on Peer Scoring and Perceptions of Peer Assessment", Assessment and Evaluation in Higher Education, Vol. 48, No. 6, pp. 760-776, August 2023.
[33] M.M. Patchan, C.D. Schunn, R.J. Clark, "Accountability in Peer Assessment: Examining the Effects of Reviewing Grades on Peer Ratings and Peer Feedback", Studies in Higher Education, December 2018.
[34] W. Zakharov, H. Li, M. Fosmire, P.E. Pascuzzi, J. Harbor, "A Mixed Method Study of Self- and Peer-Assessment: Implications of Grading Online Writing Assignments on Scientific News Literacy", College and Undergraduate Libraries, Vol. 28, No. 1, pp. 67-84, January 2021.
[35] N. Zaibout, M. Laafou, M. Madrane, "Assessments Practices of Learning Outcomes in Digital Learning Environments", International Journal on Technical and Physical Problems of Engineering (IJTPE), Issue 59, Vol. 16, No. 2, pp. 82-89, June 2024.

BIOGRAPHIES

Name: **Yassine**
Surname: **Rahani**
Birthday: 20.11.1986
Birthplace: Taza, Morocco
Bachelor: Earth and Universe Sciences, Multidisciplinary Faculty of Taza (MFT), Sidi Mohamed Ben Abdellah University, Taza, Morocco, 2011
Master: Educational Technology, Higher Normal School (ENS), Abdelmalek Essaadi University, Tetouan, Morocco, 2019
Doctorate: Student, Multidisciplinary Laboratory in Education Sciences and Training Engineering, Higher Normal School (ENS), Hassan II University, Casablanca, Morocco, Since 2022
The Last Scientific Position: Trainer, Higher Normal School (ENS), Tetouan, Morocco, Since 2024
Research Interests: Online Assessment, Educational Sciences, Training Engineering, Hybrid Teaching

Name: **Brahim**
Surname: **Nachit**
Birthday: 23.02.1972
Birthplace: Casablanca, Morocco
Bachelor: Applied Mathematics, Faculty of Sciences Ben M'sick, Hassan II University, Casablanca, Morocco, 2004
Master: Training Engineering and Didactics of Sciences, Faculty of Sciences Ben M'sick, Hassan II University, Casablanca, Morocco, 2017
Doctorate: Didactics of Mathematics, Hassan II University, Casablanca, Morocco, 2014
The Last Scientific Position: Prof., Didactics of Mathematics, Higher Normal School (ENS), Hassan II University, Casablanca, Morocco, Since 2019
Research Interests: Didactics Sciences, Training Engineering, Pedagogy
Scientific Publications: 40 Papers

Name: **Mustapha**
Surname: **Bassiri**
Birthday: 01.01.1965
Birthplace: Casablanca, Morocco
Bachelor: Animal Biology, Faculty of Sciences Ben Mesick, Hassan II University, Casablanca, Morocco, 1991
Diploma: Teacher of Secondary Cycle Professorship in Physical Education and Sport (EPS), Higher Normal School (ENS), Hassan II University, Casablanca, Morocco, 1998
Aggregation in EPS: Higher Normal School (ENS), Hassan II University, Casablanca, Morocco, 2006
Master: Training and Supervision Profession, Higher Normal School (ENS), Hassan II University, Casablanca, Morocco, 2013
Doctorate: Training Engineering and Didactics of Sciences and Technology, Hassan II University, Casablanca, Morocco, 2018
The Last Scientific Position: Assoc. Prof., Researcher of Training Engineering, and Head of the Sector, Higher Normal School (ENS), Hassan II University, Casablanca, Morocco, Since 2019
Research Interests: Engineering Training, Didactics of Sciences and Technology
Scientific Publications: 73 Papers