

# An exploration on effectiveness of anonymous peer assessment strategy in online formative assessments

Krishan Kumar

School of Computing, Information and  
Mathematical Sciences  
The University of the South Pacific  
Lautoka, Fiji  
[krishan.kumar@usp.ac.fj](mailto:krishan.kumar@usp.ac.fj)

Bibhya Sharma

School of Computing, Information and  
Mathematical Sciences  
The University of the South Pacific  
Suva, Fiji  
[bibhya.sharma@usp.ac.fj](mailto:bibhya.sharma@usp.ac.fj)

Gavin Jahir Khan

School of Computing, Information and  
Mathematical Sciences  
The University of the South Pacific  
Suva, Fiji  
[gavin.khan@usp.ac.fj](mailto:gavin.khan@usp.ac.fj)

Salsabil Nusair

School of Computing, Information and  
Mathematical Sciences  
The University of the South Pacific  
Suva, Fiji  
[salsabil.nusair@usp.ac.fj](mailto:salsabil.nusair@usp.ac.fj)

Krishna Raghuwaiya

School of Education  
The University of the South Pacific  
Suva, Fiji  
[krishna.raghuwaiya@usp.ac.fj](mailto:krishna.raghuwaiya@usp.ac.fj)

**Abstract**—This Research in Practice WIP paper presents findings to quantify the effectiveness of using online peer assessment strategy through Moodle workshop tool in a Communications and Information Technology online and blended course. The study focused on the reliability and validity of grades given by graders using Moodle workshop tool. All assessments were also manually graded by facilitators as independent variable for comparison with final Moodle grade allocated by Moodle workshop tool. The quasi-experimental study was conducted in a compulsory undergraduate course across 14 campuses on formative procedural knowledge assessment. Assessments from five graders were considered for analysis where students were also allowed to self-assess their work as one of the graders. The methods used for data analysis were descriptive statistics, correlation, interrater reliability tests and validity tests applied at 95% confidence level. The findings for peer assessment activity were reliable and valid within 0.8 inter-rater agreement with self-assessment and without self-assessment. Furthermore, the final peer assessment grade allocated by Moodle workshop tool was consistent with the facilitator grade. However, Mann Whitney-U test revealed a significant difference between Moodle grade and facilitator grade for online and blended modes mostly due to: the understanding of the online students on the assessment itself, assess to ICT tools and internet connection facilities while grading peers work and the marking variance of the facilitators.

**Keywords**—peer assessment, self-assessment, online course, blended course

## I. INTRODUCTION

The growing demand in recent years from employers and other stakeholders that graduates should have the experience in transferable skills before joining the work sector has triggered Higher Education Institutions (HEI) around the world to rethink, remodel and redesign assessments which complements the teaching and learning of 21st century learners. The academics from HEI are moving towards designing assessments which enhance student experience by including transferable skills comprising of communication, critical thinking, collaboration/teamwork and self-management skills [1] [2]. In many HEIs, these transferable skills are reflected in their graduate attributes and normally captured in the strategic plans [3] [4] [5] [6] [7]. Learning is a lifelong and life-wide process where in all works of life people learn from making judgments about one another. In HEI, student learning are no different as students learn from one another by sharing ideas and working

collaboratively in team projects while at the same time providing constructive feedback on others work. Peer Assessment (PA) can be a useful pedagogy and an effective assessment approach for student learning at HEI. PA comes with a range of benefits but also presents some challenges [8] [9]. Students become assessors within the context of participation in practice; that is, the kinds of highly contextualized learning faced in real-life and work [10]. According to literature, self-assessment and PA have been defined as a pedagogical plan that enables students to have self-paced learning, identify their own strengths and weakness, target areas of remediation and develop meta-cognitive skills [9] [10]. For the purpose of this study, we will use the definitions stipulated in [1] with appropriate modifications.

- Self-assessment: students grade, judge and make decisions on their own work against particular criteria provided by facilitators.

- Peer-assessment: students grade and give feedback on the work of their peers against particular criteria provided by facilitators.

This study aims to explore the efficacy of using anonymous online PA strategy through Moodle workshop tool in a course offered through online and blended modes. Studies have been carried out using PA in various disciplines [11] [12] [13] [14] [15]. However, there is very limited literature on the use of online PA in the Pacific context which creates a potential research opportunity. Therefore, exploring PA in online and blended mode course will add great insights to current literature on PA from HEIs in the Pacific.

Given the above aim, the following two research questions were derived:

RQ1. How reliable and valid are the student grades, when anonymous PA strategy is used in online formative assessments?

RQ2. What is the effect of using anonymous PA strategy:

RQ 2a. On student grades against facilitator grades in online and blended mode of study?

RQ 2b. On student grades of male and female students against the facilitator grade?

In remaining sections of this paper, section II outlines a literature review on PA, section III provides details of research methodology, followed by section IV with results

and discussion and section V outlining conclusions. Towards the end, limitations and future are also presented.

## II. LITERATURE REVIEW

### A. Learning Management System (LMS)

According to an educational point of view, LMS is a web-based software system used for the administration, documentation, tracking, reporting and delivery of educational courses to students on all modes of study [16]. Student interaction is through the means of online forums, and various online assessments including quizzes, assignments, lessons and peer assessments can be delivered via an LMS [7] [17] [18] [19] [20] [21] [22]. One of the popular LMS used by HEI is Modular Object Oriented Dynamic Learning Environment (Moodle) [18].

### B. Peer Assessment

PA has many benefits and has been used by the education system around the world in various subject domains including technology [11], educational design [12] [13] [23], english [24], presentation Skills [14], science [15], and business [25]. In addition, PA is an online activity which has inherited the social influence requiring the trust between the assessor and assessee [26]. Due to this, researchers have argued that the quality of PA can be improved if it is administered under anonymous setting where the anonymity of assessor and assessee can be either unidirectional or bidirectional [11] [12] [13] [23] [24] [14] [15] [25]. Under unidirectional settings either the assessor or the assessee are anonymous compared to bidirectional setting where, both the assessor and the assessee are anonymous [27]. According to [27], careful orchestration is need when considering the use of PA for online formative assessments. This includes PA moderating variables comprising of PA aids (i.e. marking rubrics and guides), PA grading (i.e. assessors are allowed to grade peers work and give feedback) and anonymity type (i.e. unidirectional or bidirectional). Apart from these, studies have very rarely reported on a standard number of assessments needed to be graded under peer grading moderating variable. Two studies [12] [14], have reported high number of assessments to be allocated per assessor. Furthermore, the reliability and validity of PA is mostly affected by grader, rubric and student socio-cultural background [25]. As such the need for verifying the reliability and validity of administering PA in the South Pacific context is fairly important for generalization. Tests such as inter-rater agreement, Cronbach alpha test, and parametric or non-parametric help to verify reliability and validity of data [25]. Mixed results were reported by studies in review article [27], where the sample size was not suitable for generalization of results and the effect of anonymity on PA results. Peer grading accuracy under anonymity setting also needs further investigation as only two studies have found participants giving lower grades under anonymous settings [23] [24]. A large scale study will serve as the catalyst for future studies in Pacific context and also contribute to overall literature on PA. Finally, provided student assessment is carried out fairly under anonymous setting, PA will reduce the workload of facilitators in courses with large enrollments. The following hypotheses were formulated to answer the derived research questions for this study:

H1: Student PA grades will be valid at the 0.80 interrater agreement level.

H2: There is no significance difference in student and facilitator grades between online and blended modes of study.

H3: There is no significance difference in student and facilitator grades across both gender.

## III. METHODOLOGY

The quasi-experimental research design utilizes quantitative search methods to explore the effectiveness of using PA strategy in a course offered in online and blended modes of delivery. The study was designed to collect students' performance data and use quantitative techniques to test selected hypotheses [28].

### A. Background

The University of the South Pacific is a regional multi-campus, higher education institution in South Pacific which was established in 1968. The university is collectively owned by its 12 member countries (see Fig. 1) where each member country houses at least one campus with its main campus based in Fiji [6]. The main campus coordinates and facilitates most of the courses in online, blended, face-to-face and print modes, and more recently ventured in to cohort-based programmes [7] and emergency remote teaching.

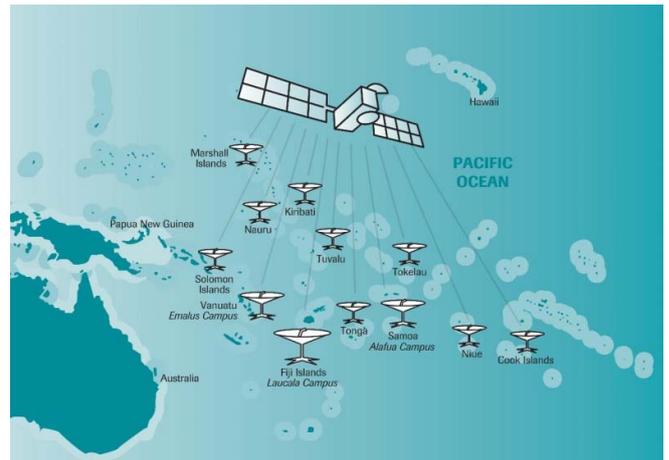


Fig. 1. Member Countries of USP connected through USPNet [6].

Setup phase	Submission phase	Assessment phase	Grading evaluation phase	Closed
<ul style="list-style-type: none"> <li>Set the workshop description</li> <li>Provide instructions for submission</li> <li>Edit assessment form</li> <li>Prepare example submissions</li> <li>Switch to the next phase</li> </ul>	<ul style="list-style-type: none"> <li>Provide instructions for assessment</li> <li>Allocate submissions</li> <li>Submissions deadline: Tuesday, 7 August 2018, 11:35 PM (5:18 days ago)</li> <li>Time restrictions do not apply to you</li> </ul>	<ul style="list-style-type: none"> <li>Open for assessment from: Friday, 10 August 2018, 6:00 AM (5:16 days ago)</li> <li>Assessment deadline: Friday, 24 August 2018, 11:55 PM (6:01 days ago)</li> <li>Time restrictions do not apply to you</li> </ul>	<ul style="list-style-type: none"> <li>Calculate submission marks</li> <li>Calculate assessment marks</li> <li>Provide a conclusion of the activity</li> </ul>	<ul style="list-style-type: none"> <li>Close workshop</li> </ul>

Fig. 2. Moodle Workshop setup for selected assessment

### B. Settings

The course chosen for this study from USP is UU100 – Communications and Information literacy. It is a generic 14-week first year course offered in blended and online modes. Students enrolled in the blended study mode were required to attend mandatory weekly laboratory classes, whereas students in the online mode were provided support with non-mandatory virtual support sessions. The aim of this course is to ensure all incoming students develop knowledge and competence in Information and Communications Technology

(ICT) skills and Research Skills. UU100 course comprises of an array of assessments including weekly topic assessments, online review quizzes, assignment and e-portfolio activities. Majority of the assessments are submitted online through the dropbox in Moodle with turnitin similarity check service enabled.

### C. Data collection and analysis

The study was focussed on performance data of students' who participated in online PA activity through Moodle workshop tool. The quantitative data were extracted from Moodle backend database by writing Structured Query Language (SQL) queries where the participants were made aware of the data use process and their identification will be kept strictly confidential. The project received ethics approval from the university's ethics committee for the use of student data.

### D. Selected Assessment description and PA setup variables

The assessment selected for the study was an information technology component of UU100 course titled simple text-based reports (Topic 3), where the students had to complete a task using Microsoft Word 2013 and above. The assessment was purely skills based and had a predefined answer. Online and blended mode students used one single Moodle page where grouping feature was used to manage the students and both groups of students were given the same assessment (Topic 3).

The PA aids used to administer the online formative assessment were instructional video and pdf guide for peer grading, criteria and rubric, training where students have to mark the samples before engaging in actual grading process, self-assessment was enabled where students were allowed to assess their own work and privacy option where anonymity of assessor and assessee with teacher presence.

The online PA tool used for the study was Moodle workshop tool which includes a four phase process (see Fig. 2). Setup phase is the initial phase where the facilitator organizes the settings of workshop module with required instructions for participants. In the submission phase, the participants submit their work according to the restricted timeframe. The submission start and end dates can be specified for participant reference. The instructor has to control in monitoring submissions report. In the assessment phase, instructors have the option to configure a mandatory training option where the participants have to assess two instructor marked assessments and once satisfied with their marking then participants engage in actual assessment of their allocated submission and provide feedback to their peers work according to specified rubric criteria and instructions given by facilitator. In grading phase, the tool calculates the final marks for submissions and for assessments with feedbacks to authors and reviewers. This study is based on student submission marks only and not the assessment marks. Furthermore, facilitators can manually override any calculated grade and also have option of publishing selected student submissions for student reference. Finally, closed phase releases all submission marks and assessment marks to students and closes the assessment [29]. The PA dataset used for analysis is outlined in Table 1. The extracted data from Moodle backend was organized in

tabulated format where G1 refers to self-assessment grade of participants own submission, G2 to G5 are peer graders mark, F is facilitator grade marked by an expert manually and M is grade allocated by Moodle workshop tool which is the average of G1 to G5 and refers to student grades in research question. For the purpose of this study only assessments with five grades were selected for analysis (N=326) out of 1512 submissions. The primary reason for this is the set criteria where students were allocated five assessments to grade including their own assessment. The total marks Topic 3 submission was out of 16.

TABLE I. PA DATASET FOR THE SIMPLE TEXT-BASED ASSESSMENT

Assessment	G1	G2	G3	G4	G5	F	M
A1	13.9	11.2	14.4	13.7	12.4	13.5	13.1
A2	15.4	15.1	14.9	15.1	14.6	12	15
A3	7.6	12.2	16	6.7	11.8	10	10.9
A4	15.7	13.1	13.6	14.3	13.1	14	14

### E. Demographics

The demographic distribution of N = 326 participants is presented in Fig. 3 and 4 by gender, mode of study and campus respectively. Majority of the participants were female (66.30%), compared to male (33.70%). Moreover, 52.80% of the participants were online mode students compared to 47.20% blended mode students. A dominant 55.80% of students were from main campus (Fiji - Laucala campus), followed by Vanuatu (13.20%), Solomon Islands (11.70%), while participants from other campuses were less than 5%.

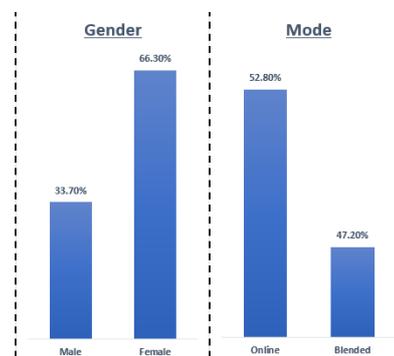


Fig. 3. Provides demographic data by gender and study mode of PA graders

## IV. RESULTS AND DISCUSSION

The data analysis was carried out with SPSS-25 using an array of statistical analysis consisting of descriptive statistics, normality test, reliability analysis, inter-item and inter-class correlation and Mann-Whitney U tests. The reliability analysis Cronbach's alpha value of 0.863, which is greater than the benchmark of 0.70, indicating a very good internal consistency.

To test the normality of performance data, Shapiro-Wilk test was carried out as shown in Table III. Since p-value < 0.05 for all variables G1 to M, the data does not follow normal distribution and further analysis followed non-parametric tests. The descriptive statistics outlined in Table III, shows mean and standard deviation of grades. The self-assessment =13.979 with SD = 2.3021, while other grades G2, G3, G4, G5, F, M have consistent mean and standard deviation. The validity and reliability of PA of the study under moderating variables PA aids, peer grading and bidirectional

TABLE II. MANN-WHITNEY U TEST

	Online and blended study mode comparison		Gender comparison	
	F	M	F	M
Mann-Whitney U	12965.500	11107.000	11418.000	11200.000
Wilcoxon W	27843.500	25985.000	34854.000	34636.000
Z	-.329	-2.516	-.576	-.845
Asymp. Sig. (2-tailed)	.742	.012	.565	.398

anonymity are shown in Fig. 5. Since the participants of this study were from various social-cultural background and study mode. It became equally important to prove, there are no disagreements between peer graders. Table IV, shows there as positive correlation between all five (G1:G5) grades, facilitator (F) and Moodle (M) grade.

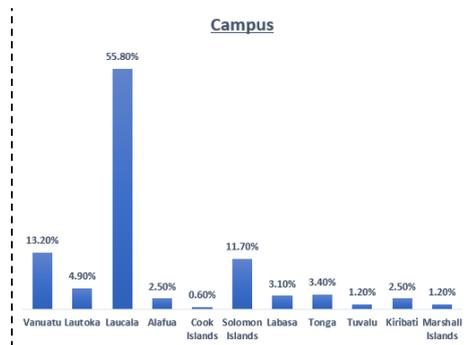


Fig. 4. Provides demographic data by campus of PA graders

TABLE III. TESTS OF NORMALITY & ITEM ANALYSIS

N = 326	Shapiro-Wilk			Descriptive Analysis	
	Statistic	df	Sig	Mean	SD
G1	.805	326	.000	13.979	2.3021
G2	.857	326	.000	12.726	3.3831
G3	.875	326	.000	12.656	3.2405
G4	.853	326	.000	12.661	3.4493
G5	.849	326	.000	12.600	3.4571
F	.815	326	.000	12.695	2.9768
M	.873	326	.000	12.926	2.2904

TABLE IV. INTER-ITEM CORRELATION MATRIX

	G1	G2	G3	G4	G5	F	M
G1	1.000						
G2	.417	1.000					
G3	.391	.424	1.000				
G4	.397	.465	.451	1.000			
G5	.239	.369	.393	.405	1.000		
F	.345	.468	.487	.530	.464	1.000	
M	.628	.752	.742	.770	.692	.646	1.000

Along with Cronbach’s alpha = 0.863 > 0.70 and intra-class correlation coefficient of 0.857 which is comparable to the Kappa inter-rater [30] [31] shown in Fig.5,

	Intraclass Correlation <sup>a</sup>	95% Confidence Interval		Value	F Test with True Value 0		
		Lower Bound	Upper Bound		df1	df2	Sig
Single Measures	.462 <sup>a</sup>	.413	.512	7.277	325	1950	.000
Average Measures	.857 <sup>a</sup>	.831	.880	7.277	325	1950	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.  
a. The estimator is the same, whether the interaction effect is present or not.  
b. Type A intraclass correlation coefficients using an absolute agreement definition.  
c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Fig. 5. Intraclass correlation coefficient

with supporting descriptive statistics confirms H1 can be accepted, where student PA grades are at 0.80 interrater agreement.

A Mann-Whitney U test carried out to test H2 is shown in Table II. Since the p-value for F = 0.742 > .05, we accept the null hypothesis that there is no significant difference in facilitator grade of both online and blended study modes compared to student grades the null hypothesis is rejected for student grades (M) with p-value = 0.012 < 0.05. The finding reveals that there is notable difference between student grades in online and blended mode of study. The primary reason for this is the self-assessment (G1) wherein some students might have given themselves higher grades.

Furthermore, to test H3 a Mann-Whitney U test was carried out as shown in Table II. Since the p-value for F = 0.565 > .05 and M = 0.398 > 0.05, we accept the null hypothesis for H3 and confirm that there is no significance in student grades in both genders compared to facilitator grade.

## V. CONCLUSION

The overall aim of this study was to investigate the reliability and validity of student grades under anonymous PA strategy in online formative assessments. Three hypotheses were developed to prove the reliability and validity of grades in online and blended modes of study. Notable findings revealed that the student grades were reliable and valid of peer graders at 0.8 inter-rater agreement. However, when student grade (M) is compared with facilitator grade (F) in online and blended modes, there was a significant difference in student grade (M). This could be due to the fact that the self-assessment was allowed.

In addition, there was no significant difference between student grades and facilitator grade across both genders. The study was only limited to assessments with five graders and the performance data was non-parametric.

Future work will consider analyzing the whole dataset and consider finding the minimum number of peer graders required to administer anonymous PA online formatives assessments with standard answers.

## ACKNOWLEDGMENT

The authors would like to thank The University of the South Pacific, Center for Flexible learning for funding the project.

## REFERENCES

- [1] C. Adachi, J. H.-M. Tai and P. Dawson, "Academics' perceptions of benefits and challenges of self and peer assessment in higher education," *Assessment & Evaluation in Higher Education*, pp. 294-306, 2018.
- [2] S. Bennett, P. Dawson, M. Bearman, E. Molloy and D. Boud, "How technology shapes assessment design: Findings from a study of university teachers," *British Journal of Educational Technology*, pp. 672-682, 2017.
- [3] Adelaide university, "gradattributes," 2 January 2020. [Online]. Available: <https://www.adelaide.edu.au/learning/strategy/gradattributes/>.
- [4] Stellenbosch University, "Graduate\_attributes," [Online]. Available: [https://www.sun.ac.za/english/faculty/healthsciences/chpe/Pages/Graduate\\_attributes.aspx](https://www.sun.ac.za/english/faculty/healthsciences/chpe/Pages/Graduate_attributes.aspx). [Accessed 25 March 2020].
- [5] Glasgow University, "graduate attributes," 2 November 2016. [Online]. Available: [https://www.gla.ac.uk/media/media\\_220663\\_en.docx](https://www.gla.ac.uk/media/media_220663_en.docx). [Accessed 25 March 2020].
- [6] The University of the South Pacific, "strategicplan2013-2018," 28 November 2019. [Online]. Available: [www.usp.ac.fj/strategicplan2013-2018](http://www.usp.ac.fj/strategicplan2013-2018).
- [7] B. Sharma, F. J. Lauano, S. Narayan, A. Anzeg, B. Kumar and J. Raj, "Science teachers accelerated programme model: a joint partnership in the Pacific region," *Asia-Pacific Journal of Teacher Education*, vol. 46, no. 1, pp. 38-60, 2018.
- [8] D. Boud and N. Falchikov, "(2006) Aligning assessment with long Term learning," *Assessment & Evaluation in Higher Education*, 31(4), pp. 399-413, 2006.
- [9] B. Carnell, "Aiming for Autonomy: Formative Peer Assessment in Final-year Undergraduate Course," *Assessment & Evaluation in Higher Education*, pp. 1269-1283, 2016.
- [10] N. Falchikov, "Peer Feedback Marking: Developing Peer Assessment," *Innovations in Education and Training International*, pp. 175-187, 1995.
- [11] L. Li, "The role of anonymity in peer assessment," *Assessment & Evaluation in Higher Education*, pp. 645-656, 2017.
- [12] A. Raes, E. Vanderhoven and T. Schellens, "Increasing anonymity in peer assessment by using classroom response technology within face-to-face higher education," *Studies at higher Education*, pp. 178-193, 2013.
- [13] T. Rotsaert, E. Panadero and T. Schellens, "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*, vol. 33, no. 1, pp. 75-99, 2018.
- [14] E. Vanderhoven, A. Raes, H. Montrieux, T. Rotsaert and T. Schellens, "What if pupils can assess their peers anonymously? A quasi-experimental study," *Computers & Education*, pp. 123-132, 2015.
- [15] F.-Y. Yu and S. Sung, "A mixed methods approach to the assessor's targeting behavior during online peer assessment: effects of anonymity and underlying reasons," *Journal of Interactive learning Environments*, vol. 24, no. 7, pp. 1674-1691, 2015.
- [16] Kumar, K., Sharma, B. N., Nusair, S., & Khan, G. J. (2019). Anonymous online peer assessment in an undergraduate course: An analysis of Students' perceptions and attitudes in the South Pacific. 2019 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE). Yogyakarta - Indonesia: IEEE.
- [17] B. Sharma, R. Nand, M. Naseem and E. V. Reddy, "Effectiveness of online presence in a blended," *Studies in Higher Education*, pp. 1-20, 2019.
- [18] M. Dougiamas and P. Taylor, "Moodle: Using Learning Communities to Create an Open Source Course," in *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, Chesapeake, VA, USA.
- [19] M. L. Wen and C.-C. Tsai, "University Students' Perceptions of and Attitudes Toward (Online) Peer Assessment," *Higher Education*, pp. 27-44, 2006.
- [20] D. Pereira, M. A. Flores and L. Niklasson, "Peer and self-assessment applied to oral presentations from a multidisciplinary perspective," *Assessment & Evaluation in Higher Education*, pp. 1008-1032, 2016.
- [21] M. M. Ashenafi, "Peer-assessment in higher education – twenty-first century practices, challenges and the way forward," *Assessment & Evaluation in Higher Education*, pp. 226-251, 2017.
- [22] M. A. Flores, A. M. V. Simao, A. Barros and D. Pereira, "Perceptions of effectiveness, fairness and feedback of assessment methods: a study in higher education," *Studies in Higher Education*, pp. 1523-1534, 2015.
- [23] M. J. Wilson, M. M. Diao and L. Huang, "'I'm not here to learn how to mark someone else's stuff': an investigation of an online peer-to-peer review workshop tool," *Assessment & Evaluation in Higher Education*, pp. 15-32, 2015.
- [24] C. H. Peterson and N. A. Peterson, "Impact of Peer Evaluation Confidentiality on Student Marks," *International Journal of the Scholarship of Teaching and Learning*, 2011.
- [25] R. Lu and L. Bol, "A comparison of anonymous versus identifiable e-peer review on college student writing performance and the extent of critical feedback," *Journal of Interactive Online Learning*, 2007.
- [26] K. D. Strang, "Effectiveness of peer assessment in a professionalism course using an online workshop," *Journal of Information Technology Education: Innovations in Practice*, vol. 14, no. 1, pp. 1-16, 2015.
- [27] N. A. van Gennip, M. S. Segers and H. H. Tillema, "Peer assessment for learning from a social perspective: The influence of interpersonal variables and structural features," *Educational Research Review*, vol. 4, no. 1, pp. 41-54, 2009.
- [28] E. Panadero and M. Alqassab, "An empirical review of anonymity effects in peer assessment, peer feedback, peer review, peer evaluation and peer grading," *Assessment & Evaluation in Higher Education*, pp. 1-26, 2019.
- [29] J. W. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, NY: Sage, 2009.
- [30] Moodle, "Using workshop," 2 September 2019. [Online]. Available: [https://docs.moodle.org/38/en/Using\\_Workshop#Setup\\_phase](https://docs.moodle.org/38/en/Using_Workshop#Setup_phase). [Accessed 15 April 2020].
- [31] J. L. Fleiss and J. Cohen, "The Equivalence of Weighted Kappa and the Intraclass Coefficient as measures of reliability," *Education Psychological Measurement*, pp. 613-619, 1973.
- [32] J. N. Mandrekar, "Measures of Interrater Agreement," *Journal of Thoracic Oncology*, pp. 6-7, 2011.