THE ROLE OF SELF AND PEER ASSESSMENT IN HIGHER EDUCATION[1]

Iglesias Pérez, M.C.[a], Vidal-Puga, J.[a], Pino Juste, M.[b]

[a] Department of Statistics and Operations Research. Universidade de Vigo.
[b] Department of Didactics, School Organization and Research Methods. Universidade de Vigo.

ABSTRACT

Self-assessment and peer assignment have clear advantages for the training of responsible, critical, and reflective professionals. In recent years, self and peer evaluation have also been shown to be even more effective than lecturer evaluation when we assure anonymity through online platforms learning tools. Therefore, self and peer assessments are to become a core aspect of student-centred evaluation processes in Higher Education. In the present work, we compare the formative evaluation from the lecturer with the self and peer assessments through a virtual learning environment. The subject of study if formed by assessments prepared by students in a first-year course in a Social Sciences degree at the Universidade de Vigo, Spain. We find a strong concordance between peer assessment and lecturer assignment, and a moderate agreement between self-assessment and lecturer assignment. These results show that students perform well as peer evaluators, with peer assignment being a procedure with high validity and reliability.

Keywords: Self-assignment, Peer assignment, Continuous assignment, Higher Education, concordance Analysis.

INTRODUCTION

The adaptation of degree courses to the European Higher Education Area (EHEA) has generated a series of methodological changes. These changes imply a significant increase in the time that lecturers devote to the feedback of the activities proposed during the teaching-learning process in order for students to consolidate the competences described in the different curricula. Even with the help of a

[1] Juan Vidal-Puga acknowledges financial support by the Spanish Ministerio de Economía, Industria y Competitividad through grant ECO2017-82241-R, and Xunta de Galicia through grant ED431B 2019/34. María del Carmen Iglesias-Pérez acknowledges financial support by Xunta de Galicia through grant GRC ED431C 2016/040.

learning management system, the average time to correct activities usually ranges between 7 and 10 days per activity (Cantabella et al., 2016).

Likewise, the growing demand for lifelong training and more responsible, critical, and thoughtful professionals has favoured new approaches to the relationship between learning and its evaluation. It has also greatly influenced, since the beginning of the new century, the development of new forms of evaluation, such as self and peer assignment (Dochy et al., 1999).

Given this background, in recent years, there has been a growing concern about the validity of these types of evaluation, whose paradigms have remained unchanged for decades (Bukowski et al., 2017). An increasing number of studies have addressed the effectiveness of the different types of evaluation to determine the student's academic performance or to provide feedback on it. Even so, this increase has not gone hand in hand with the evaluation of practices with students (Asikainen et al., 2014) nor much less the effect of the use of virtual learning environments.

In different studies, it has become clear that self-assessment and peer evaluation are very useful learning tools (González de Sande and Godino-Llorente, 2014, and references therein). They bring significant benefits for student-learning processes when implemented from principles of evaluation for learning (formative evaluation) (Panadero and Brown, 2017). They are even valid as a summative evaluation technique (Deeley, 2014). Tutor feedback during the teaching-learning process is not as useful as peer reviews to improve student progression (McConlogue, 2015).

Therefore, self-assessment and peer evaluation are becoming core aspects of student-centred evaluation processes in the field of higher education (Wanner and Palmer, 2018). Both forms of evaluation are useful for developing critical skills in students, such as taking responsibility for their learning, developing a better understanding of the subject's content, evaluation criteria and their values and judgments, and developing critical reflection skills. Moreover, peer assessment allows interaction in the group and cooperative work, makes students a critical subject and helps them to issue a qualification of their peers' work, favouring the acquisition of critical ability, making students more autonomous and responsible, not only of one's work but also that of their colleagues.

Hence the need for a paradigm shift in terms of evaluation that takes advantage of recent developments in technology and statistical techniques for both formative and summative evaluation of student academic performance that uses these other more efficient strategies (Bukowski et al., 2017).

Cheng et al. (2015) have documented the effectiveness of peer evaluation using a digital tool. They have shown that cognitive feedback (for example, direct correction) is more useful for students' academic achievement than affective feedback (for example, praising comments) and meta-cognitive feedback (for example, reflecting comments).

Peer evaluation requires students to judge the work of their peers based on the evaluation criteria usually offered in a rubric (Jones and Alcock, 2014). Moreover, electronic rubrics can be used to control anonymity (Martín-Monje et al., 2014). The rubric facilitates the issuance of an evaluative judgment understood as the ability to make decisions about the quality of work of oneself and others (Tai et al., 2018).

In many cases, this evaluative judgment can be influenced by undesirable social effects such as peer pressure and favouritism or fear of disapproval, especially when students need to evaluate their peers in a face-to-face environment (Raes et al., 2015; Vanderhoven et al., 2015). Therefore, as Cartney (2014) points out, it is essential to take into account the emotional aspects, as well as the cognitive aspects of peer learning.

Hence the importance of using strategies that allow anonymity in peer evaluation to counteract these undesirable social effects. We believe that greater anonymity will induce a reduced perception of peer pressure, a higher feeling of comfort, and more positive attitudes towards peer evaluation. However, most teachers do not use anonymous forms of peer evaluation (Panadero and Brown, 2017).

We must take into account the implementation of these programs that the perceptions of university students on peer review before participating in an activity of this type is usually very positive. Students held high expectations of both the process and the competence of their peers as reviewers. However, after peer review, positive perceptions generally change downward (Mulder et al., 2014). Both teachers and students agree on the low use of participatory evaluation modalities

in universities. They highlight the need to establish training processes, both for teachers and for students, that affect the knowledge and implementation of these modalities in order to improve students in autonomous and strategic learning (Ibarra Saiz and Rodríguez Gómez, 2014).

Based on these premises, the objective of this study is to verify the reliability and validity of peer evaluation and self-assessment through the use of a virtual learning environment, based on their agreement with the lecturer's evaluation (gold standard). The existence of a high concordance will allow teachers to have a reliable anonymous evaluation procedure, especially useful when the number of students in the classroom is high.

METHODOLOGY

Research context

The experience was carried out in a first-year course named Introduction to Administrative Statistics, in the degree of Public Direction and Management of the University of Vigo, Spain, through the use of a collaborative learning tool (Moodle 2.5) which allows self and peer assignment. The course is taught in two teaching modalities: classroom and blended learning. In the blended learning modality, the students' work evaluated by means of self and peer assignment supposes 60% of the final qualification, whereas those of the classroom modality supposes 35% of the final qualification.

Students were instructed orally during the first classroom session of the importance of their evaluation work, given its impact on the qualification of classmates. The evaluation rubric for the correction of each activity was also clearly explained to the students (thirteen works, plus an additional one in the blended learning modality that we discard in this study). Finally, the operation of the "workshop" of Moodle was explained, whose most important characteristics are described below.

The Moodle "workshop" tool used in the course was programmed to randomise three jobs for the peer assessment, as well as the self-assessment, to each of the submitted works. Students get two grades (with weights of 80% and 20% by default): one for the submitted work (qualification per assignment) and another for

their peer evaluations (qualification by evaluations). The final grade per submission is a weighted mean of the ratings assigned by all reviewers of that submission. The evaluation grade estimates the quality of the peer assignment done by the participant. This quality estimation depends on the distance between their evaluation and the "best evaluation," considered to be the one made by the lecturer, or the median of all the evaluations, in case the lecturer does not evaluate that assignment. The final grade per assignment is the weighted average of the grades, with the lecturer's grade, if done, weighting 16 times those of the students'.

Research Design

The research was carried out in the 2016-17 academic year. During the period of the lecture of the subject (2nd semester), students sent assignments and made corrections to peers and to themselves, using the Moodle "workshop". Once the semester was completed, a collection of the submissions and corrections corresponding to the thirteen works common to the two modalities (classroom and blended learning) was made, resulting in a total of 225 submissions, 597 corrections of students (144 self-assessments and 453 peer assessments), and 225 evaluations made by the lecturer. As for the peer evaluations of the same work, the platform assigned an average of the coefficients received, resulting in 2.1 peer qualifications per submission and six works with no peer evaluation.

We conducted two studies to compare the performance of the peer assessment and self-assessment with the lecturer's evaluation. In the first one, we study the concordance between the lecturer's qualification and that of peers, as well as between the lecturer's qualification and the self-evaluation, for the work submitted, on a scale of 0 to 80 points provided by the platform. In the second one, for each student, we consider the final grade (arithmetic mean of all submissions), which was rescaled from 0 to 10 points and studied the concordance between the final grade awarded by the lecturer and the one corresponding to the peers. The sample included all the students who made a submission, 31 in total: 20 of the classroom modality and 11 of the blended modality; 19 men and 12 women.

As concordance measures, we use the intraclass correlation coefficient, ICC (definition and interpretation in Fleiss, 1986), the concordance correlation

coefficient of Lin, CCC (Lin et al., 2002), and the graph of Bland and Altman (Bland and Altman, 1986). We also study the linear correlation of the evaluation methods, which are compared using the t-test for related samples.

We establish the validity of the peer assignment based on the presence of high concordance between peer assignment and lecturer assignment.

We establish peer assignment reliability by randomly selecting two pairs of corrections for each submission and studying the agreement between the two groups of corrections formed (ICC).

To compare the influence of gender and modality on the evaluation of students, we use the t-test of comparison of independent samples and the non-parametric Mann-Whitney U test.

Results

Submissions Evaluation

Table 1 shows the results of the correlation and difference of means between the evaluation methods. The correlation between the grades assigned by the lecturer and the peers for the papers submitted ($n = 219$) is significant ($p < 0.001$) with a high value $r = 0.801$, while considering only the self-assessments ($n = 144$) we obtain a lower correlation, $r = 0.754$ ($p < 0.001$). We do not find significant differences between the lecturer evaluation and the average peer grade ($t = 0.291$, $p = 0.771$), but there is a significant difference between the lecturer evaluation and the self-evaluations ($t = 5.526$, $p < 0.001$). Students tend to self-rate above the lecturer evaluation, with an average difference of 5.58 points out of 80 (or 0.697 out of 10).

Table 1. Basic statistics of the difference in methods (mean and standard deviation), t-test of comparison of paired samples and correlation coefficient r between the lecturer evaluation and peer evaluation, and between lecturer evaluation and self-evaluation.

	Mean	Standard deviation	t	df	p-value	r	p-value

Peer evaluation - lecturer evaluation	0.233	11.850	0.291	218	0.771	0,801***	0.000
Self evaluation - lecturer evaluation	5.583	12.125	5.526***	143	0.000	0.754***	0.000

df = degree of freedom, *** p-value <0.001.

Figure 1 shows the relationship between student and lecturer qualifications separated into two groups: peers and self-assessments. The bisector indicates the perfect match points. A majority of self-assessments can be seen above the bisector, indicating a tendency to self-rate above the lecturer's grade.

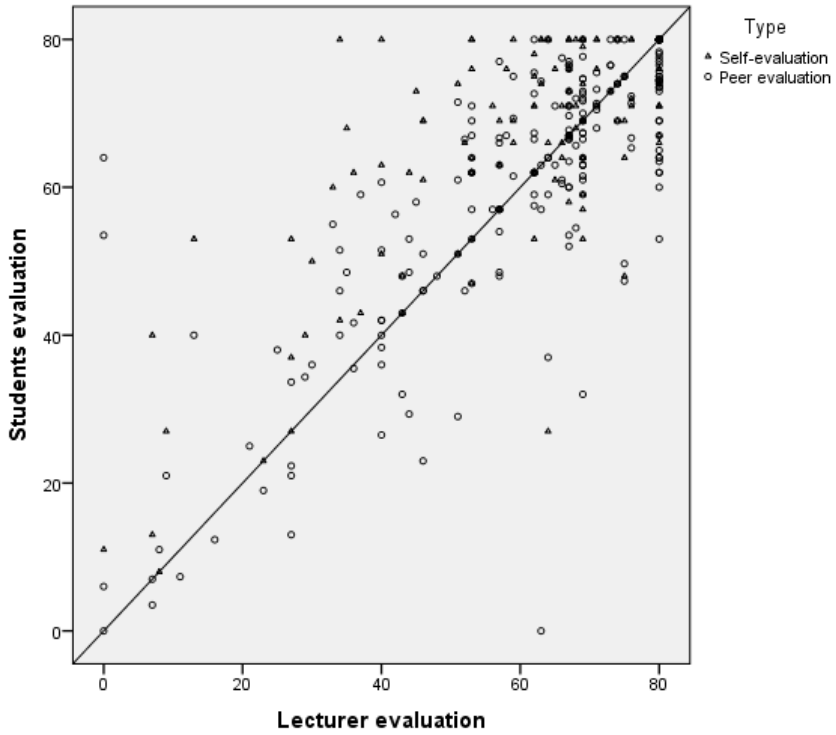

Figure 1: Relationship between student and lecturer qualifications by groups:

peers and self-evaluations.

In order to explore a possible gender effect and the modality in the qualification granted by the students, we make a comparison by groups of the variable difference between self-evaluation and reader evaluation. The results are presented in Table 2. They show significant differences by gender ($p < 0.01$) but not by modality ($p > 0.05$). The self-assessment by men exceeds the lecturer's average by 10.15 points

out of 80 (approx. 1.3 out of 10), while in women, the average is 3.07 points out of 80 (approx. 0.4 out of 10).

Table 2. Basic statistics (mean and standard deviation), t-test comparison of means and Mann-Whitney U test of the difference between self-evaluation and qualification of the lecturer, depending on gender and teaching modality.

Self evaluation - lecturer evaluation		n	Mean	SD	t	df	p-value	U	p-value
Gender	male	51	10.156	12.017	3.480**	142	0.001	1,514.5**	0.000
	female	93	3.075	11.493					
Teaching modality	classroom	61	7.131	12.382	1.317	142	0.190	2,298.5	0.338
	Blended learning	83	4.445	11.879					

SD = standard deviation; ** p-value <0.01

We present the results of the concordance measures for the evaluation methods in Table 3. The coefficients of concordance between the self-assessments and the lecturer's qualifications are CCI = 0.704 and CCC =0.702, which indicate a moderate concordance. Concordance between peer and lecturer evaluations take ICC values = 0.801 and CCC = 0.800, that is, high concordance, implying the validity of the peer assessment method.

Table 3. Concordance measures of the qualification methods for the submissions.

	Intraclass correlation coefficient							C.Lin
	CCI	95% Interval confidence		F	df1	df2	p-value	CCC
Peer vs lecturer evaluation	0.801***	0.748	0.844	9.016	218	218	0.000	0.800
Self vs lecturer evaluation	0.704***	0.540	0.804	6.724	143	143	0.000	0.702

ICC = Intraclass Correlation Coefficient; CCC = Concordance Correlation Coefficient; ***p-value < 0.001.

We establish peer assignment reliability by randomly selecting two pairs of corrections for each submission and calculating its concordance. The intraclass

correlation coefficient was 0.727, indicating reasonable reliability.

Final students grade

To know to what extent the workshop tool can free the lecturer from the task of correcting all the students' work, we carry out a concordance study between the final grade achieved by each student calculating the average of the grades assigned by the lecturer and the final grade resulting from the average of the submissions corrected by peers. We present the results in Table 4. We have considered the average based on the submitted works and, also, the average based on the thirteen proposed activities. We scale these ratings from 0 to 10 points.

Table 4. Basic statistics of the difference in methods (mean and standard deviation), t-test of comparison of paired samples, correlation and concordance coefficients, r, ICC and CCC, between the lecturer evaluation and peer evaluation.

	Mean	SD	t	df	p-value	r	ICC	CCC
APE - ALE (on submitted works)	0.0568	0.6986	0.445	29	0.660	0,898***	0.884***	0.880
APE - ALE (on 13 works)	0.0621	0.3713	0.916	29	0.367	0.992***	0.992***	0.992

APE=Average peer evaluation; ALE=Average lecturer evaluation; SD=standard deviation; df = degrees of freedom; *** p-value <0.001.

We find a high concordance between the students' final grades for both procedures (APE and ALE) based on the submitted works: ICC = 0.884 and CCC = 0.880, a strong correlation, r = 0.898, and the difference in means between both methods was not significant: 0.0567 ± 0.6986 (t = 0.445 and p = 0.660). Similar results but with a higher concordance (ICC=CCC=0.992) were found when comparing the two procedures based on the averages over the 13 mandatory works. In addition, the Bland and Altman graph represented in Figure 2 shows that 95% of the differences between the final grade of the peers and the lecturer are between mean ± 1.96*SD, that is, between -1.312 and 1.426 points for the averages over the submitted works, and between -0.665 and 0.789 for the averages over the mandatory works. In the latter case, less than one point of difference, although it can reach more than half

a point. If we accept that a difference of up to half a point can be considered admissible, in Figure 2 (right), we see that only 17% of grades (5 of 30) exceed this half a point difference.

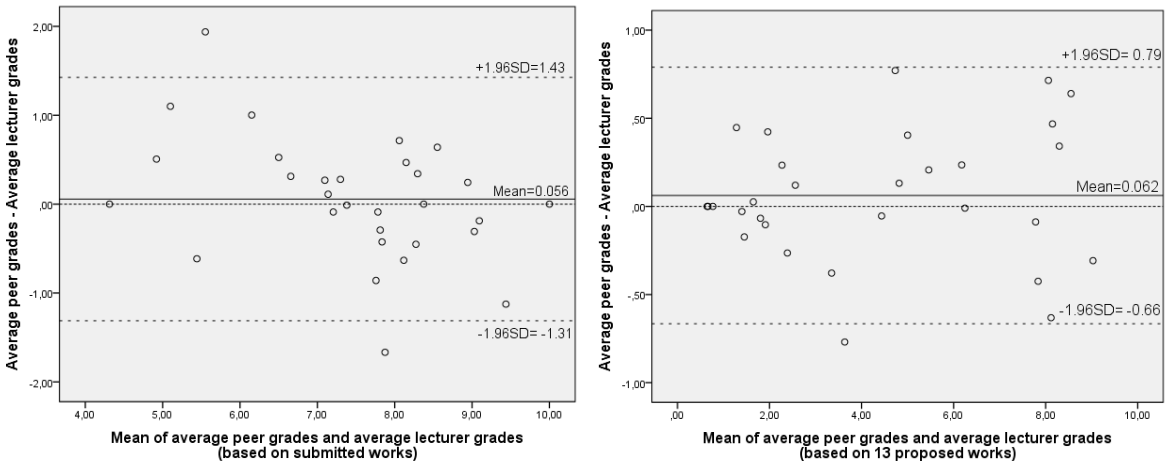

Figure 2: Bland and Altman graph: distribution of the difference in students' final grades over the mean of the grading methods (peer and lecturer evaluation) based on the submitted works (left) and based on the thirteen proposed works (right)

DISCUSSION AND CONCLUSIONS

In this study, we compare lecturer evaluation, self-assessment, and peer assessment.

The agreement between the qualifications of the lecturer and the self-assessments is moderate, and we also find a significant effect of gender. The difference between the grade of the self-assessment and the grade of the lecturer is greater in men than in women. This may be due to a higher self-estimate of intelligence by men (von Stumm et al. , 2009) or to the existence of a cultural gender difference. In this second sense, it might be important, as Cartney (2014) points out, to take into account cultural differences in the design of peer evaluation programs to make them more effective.

However, the agreement between the qualification of lecturer and peers is high for each of the works submitted and very high for the final qualification of the students. In fact, the difference between the final qualification of lecturer and peers for each student is always less than 1 point, and in 83% of cases less than half a point.

The same conclusion is reached by González de Sande and Godino-Llorente (2014)

comparing the instructor's formative evaluation and feedback, self-evaluation (SA), and peer evaluation (PA) in a study on the evaluation of engineering problems (without the use of online platforms). Their results suggest that PA is a more effective learning tool than SA, and both are more effective than the formative assessment of the instructor.

Jones and Alcock (2014) also reach similar conclusions, finding high validity and reliability among evaluators, which suggests that students performed well as peer evaluators, even though no evaluation rubrics were used in their study.

Therefore, it seems that a peer correction system is a good tool in the formative and final evaluation of students.

The results of this research on co-evaluation using a web-based learning platform with evaluation rubrics may encourage university professors to integrate peer evaluation into their formative and summative assessments in a more effective way.

Peer evaluation is useful because both the active participation in learning activities and the review of evaluation activity facilitate learning for students involved in these processes (Hodgson et al., 2014). Along the same line, Wanner and Palmer (2018) point out that students tend to consider formative self-evaluation and peer evaluation beneficial to obtain more information about the evaluation process and to improve their work.

Besides, Tai et al. (2018) consider that these pedagogical self-assessment practices facilitate the acquisition of skills that students require both inside and outside of higher education settings.

Limitations

Although the number of works evaluated was very high, and the sample of students is representative of the usual enrolment in the degree studied, the study is reduced to one academic year.

Some longitudinal aspects, such as the possible improvement of the evaluation throughout the course, have not been taken into account.

REFERENCES

Asikainen, H., Virtanen, V., Postareff, L., Heino, P. (2014) The validity and students' experiences of peer assessment in a large introductory class of gene technology. *Studies in Educational Evaluation*, 43, 197-205. DOI: 10.1016/j.stueduc.2014.07.002

Bland J.M., Altman, D.G. (1986) Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *The Lancet*, 327, 307-310. DOI: 10.1016/S0140-6736(86)90837-8

Bukowski, W.M., Castellanos, M., Persram, R.J. (2017) The Current Status of Peer Assessment Techniques and Sociometric Methods. *New Directions for Child and Adolescent Development*, 157, 75-82. DOI: 10.1002/cad.20209

Cantabella, M., López-Ayuso, B., Muñoz, A. and Caballero, A. (2016) A tool for monitoring lecturers' interactions with Learning Management Systems (in Spanish). *Revista española de Documentación Científica*, 39, e153. DOI: 10.3989/redc.2016.4.1354

Cartney, P. (2014) Exploring the use of peer assessment as a vehicle for closing the gap between feedback given and feedback used. *Assessment & Evaluation in Higher Education*, 35, 551-564. DOI: 10.1080/ 02602931003632381

Cheng, K-H., Liang, J-C., Tsai, C-C. (2015) Examining the role of feedback messages in undergraduate students' writing performance during an online peer assessment activity. *The Internet and Higher Education*, 25, 78-84. DOI: 10.1016/j.iheduc.2015.02.001

Deeley, S.J. (2014) Summative co-assessment: A deep learning approach to enhancing employability skills and attributes. *Active Learning in Higher Education*, 15, 39-51. DOI: 10.1177/1469787413514649

Dochy, F. J. R. C., Segers, M., Sluijsmans, D. (1999) The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher education*, 24, 331-350. DOI: 10.1080/03075079912331379935

Fleiss, J.L. (1999) *The design and analysis of clinical experiments*. John Wiley & Sons, Inc.

González de Sande, J.C., Godino-Llorente, J.I. (2014). Peer assessment and self-assessment: Effective learning tools in higher education. *International Journal of Engineering Education*, 30, 711-721.

Hodgson, P., Chan, K., Liu, J. (2014) Outcomes of synergetic peer assessment: First-year experience. *Assessment and Evaluation in Higher Education*, 39, 168-178. DOI: 10.1080/02602938.2013.803027

Ibarra Saiz, M.S., Rodríguez Gómez, G. (2014) Participatory assessment methods: an analysis of the perception of university students and teaching staff (in Spanish). *Journal of Educational Research,* 32, 339-362. DOI: 10.6018/rie.32.2.172941

Jones, I., Alcock, L. (2014) Peer assessment without assessment criteria. *Studies in Higher Education,* 39, 1774-1787. DOI: 10.1080/03075079.2013.821974

Lin, L., Hedayat, A.S., Sinha, B., Yang, M. (2002) Statistical Methods in Assessing Agreement. *Journal of the American Statistical Association*, 97, 257-270. DOI: 10.1198/016214502753479392

Martín-Monje, E., Vázquez-Cano, E., Fernández, M. (2014). Peer assessment of language learning resources in virtual learning environments with e-rubrics. *International Journal of Technology Enhanced Learning*, 6, 321-342. DOI: 10.1504/IJTEL.2014.069018

McConlogue, T. (2015) Making judgements: investigating the process of composing and receiving peer feedback. *Studies in Higher Education*, 40, 1495-1506. DOI: 10.1080/03075079.2013.868878

Mulder, R.A., Pearce, J.M., Baik, C. (2014) Peer review in higher education: Student perceptions before and after participation. *Active Learning in Higher Education*, 15, 157-171. DOI: 10.1177/1469787414527391

Panadero, E., Brown, G.T.L. (2017) Teachers' reasons for using peer assessment: positive experience predicts use. *European Journal of Psychology of Education*, 32, 133-156. DOI: 10.1007/s10212-015-0282-5

Raes, A., Vanderhoven, E., Schellens, T. (2015) Increasing anonymity in peer assessment by using classroom response technology within face-to-face higher

education. *Studies in Higher Education*, 40, 178-193. DOI: 10.1080/03075079.2013.823930

Tai, J., Ajjawi, R., Boud, D., Dawson, P., Panadero, E. (2018) Developing evaluative judgement: enabling students to make decisions about the quality of work. *Higher Education*, 76, 467-481. DOI: 10.1007/s10734-017-0220-3

Vanderhoven, E., Raes, A., Montrieux, H., Rotsaert, T., Schellens, T. (2015) What if pupils can assess their peers anonymously? A quasi-experimental study. *Computers and Education*, 81, 123-132. DOI: 10.1016/j.compedu.2014.10.001

Von Stumm, S., Chamorro-Premuzic, T., Furnham, A. (2009). Decomposing self-estimates of intelligence: structure and sex differences across 12 nations. *British Journal of Psychology*, 100, 429–442, DOI: 10.1348/000712608X357876

Wanner, T., Palmer, E. (2018) Formative self-and peer assessment for improved student learning: the crucial factors of design, teacher participation and feedback. *Assessment and Evaluation in Higher Education*, 43, 1032-1047. DOI: 10.1080/02602938.2018.1427698