International Journal of Innovative Teaching and Learning in Higher Education
Volume 1 • Issue 1 • January-March 2020

Evaluation of Multi-Peer and Self-Assessment in Higher Education:
A Brunei Case Study

David Hassell, University of Bath, Bath, UK
Kok Yueh Lee, Universiti Teknologi Brunei, Mukim Gadong A, Brunei, Darussalam
https://orcid.org/0000-0001-8836-6135

ABSTRACT

This article presents an evaluation of the use of peer and self-assessment as part of the learning process in a public speaking assessment coursework, with students from two departments taking part. Students were assessed by themselves, their peers and the lecturer using an online platform, Google forms, utilizing a set of rubrics. The marks were compared between markers to identify similarities and differences. After the process, student feedback on the experience was obtained using a questionnaire utilizing the Likert seven point scale to rate different questions. Analysis of the marks awarded found that whilst there might be correlations between different markers (i.e. peer – self) for marks on certain subsections of the work, there was no overall correlation between marks. Student perceptions to the exercise indicated that the use of rubrics was well received; students considered it a fair assessment method and it provided information on how to perform well in the assessment.

KEYWORDS

Civil Engineering, Computer Science, Higher Education, Peer-Assessment, Public Speaking Rubrics, Self-Assessment, Student Experience

INTRODUCTION

The use of self and peer assessment is increasingly being used within higher education for both formative and summative assessment, with the former providing timely feedback to students on their performance during a particular exercise. This assessment can take many forms, from informal verbal feedback based on student experience to student evaluation using model answers or assessment rubrics. In the latter case, self or peer assessment and the use of rubrics has been shown to improve performance if implanted effectively (Arendt, Trego, & Allred, 2016). Some of the benefits of using rubrics are that students are able to understand the tutor's expectations, understand the specific intended learning outcomes of the assignment or task and the assessment criteria, and the provision of feedback to students informing their achievement and performance skills (Andrade, 2005; Andrade & Du, 2005; Reddy & Andrade, 2010). Asikainen, Virtanen, Postareff and Heino (2014, p. 202)

DOI: 10.4018/IJITLHE.2020010104

suggested that "long-term pedagogical training is not the only way to develop the university teaching and learning", implying that the use of rubrics and peer assessment can be an effective teaching approach to improve student learning. However, initial work indicated that there was resistance to this shift from lecturer assessment to peer assessment by both staff and students (Liu and Carless, 2006) and it has subsequently been proposed that a number of different approaches are required to mitigate student reluctance (Sendziuk, 2010).

Work on the use of rubrics has been reported over a wide range of disciplines and academic levels (Andrade, Du, & Wang, 2008; Cho, Schunn, & Wilson, 2006; Moni & Moni, 2008; Tierney & Simon, 2004), but their use poses challenges for the lecturer including rubric reliability and validity (Andrade, 2005, Moskal & Leydens, 2000). Analysis of published work within Engineering education (Davey, 2011, Davey & Palmer, 2012) identifies considerable scatter between assessor and assesse, with later work (Davey, 2015) reporting that students undertaking self-assessment of predominantly quantitative calculations marked their work on average 16% higher than the tutor did. Rater reliability in this case is likely a consequence of the type of question posed as well as any training raters receive, with the greatest deviation observed in the more open-ended questions. Other studies report similar over-marking for more qualitative activities such as self-assessment in report writing (Bringula and Moraga, 2017) and peer-assessment in oral presentations (Langan et al., 2005). One approach to increasing accuracy between assessor and assessed is the use of multiple assessors (Cho, Schunn & Wilson, 2006), however the need for multiple assessors could increase the workload and resource burden on the academic staff. These above studies indicate that success of the approach is dependent on numerous factors including the method of implementation, the student cohort and the manner of the exercise assessed.

Collecting and processing numerous marking rubrics would introduce an additional burden on teaching staff when undertaking assessment, whilst increasing the possibility of inaccurate data entry and hence errors when compiling marks. Technology is being increasingly utilised within education (Palenque, 2016) and it is possible that this can also be used to simplify the peer assessment process, provide a simple approach to engage multiple assessors for an individual work, and increase reliability and reduce the lecturer assessment burden. This paper presents an evaluation of the use technology, specifically Google forms, for self and peer-assessment using rubrics as part of the learning process for public speaking in a module studied by two different student cohorts, Civil Engineering and Computer Science. Google forms is a simple online platform that allows for the quick and easy capture of participant responses for subsequent data analysis and this work addresses the following two questions:

Is Google forms a useful platform for the self and peer-assessment of student presentations?
Does the use of multiple peer assessors increase the reliability of peer review when compared to the lecturer?

Using a uniform pedagogical approach, it provides a fair basis for evaluating the effectiveness of the self, and peer to peer marking in accurately capturing the final student mark for students with different academic backgrounds within a South East Asian institution. Possible trends in differences are analysed through interrogation of the marking, and student experiences and perceived learning are evaluated using a tailored questionnaire.

METHODOLOGY

Research Context

Public Speaking is one of the topics covered over a period of two weeks in the module, Effective Communication. This module was studied by two cohorts of students, Civil Engineering and Computer

Science, undertook as part of their respective degrees of study. The students were required to prepare and deliver a five-minute speech on any pre-approved chosen topic to the class. Some of the public speech topics chosen include general aspects of bioengineering, volunteerism, cosmetic surgery, sports, or motivational talks. Assessment of the public speaking was undertaken using a specifically developed rubric that was classified into two main competencies, delivery and content. It consisted of nine criteria: five delivery-related criteria (appearance, body language, eye contact, language and voice, and pacing and timing) and four criteria on content-related (Introduction, Conclusion, Central Idea and Purpose, and Content and Originality). Students rated the quality of the public speaking skills according to the criteria based on a scale. The rating scale or weightage for each criterion varied and would receive a score of Exemplary [4], Good [3], Adequate [2], Ineffective [1] and Poor [0], with an overall total of 100%. The rubrics is provided in Appendix A.

Participants

Two cohorts of first year undergraduate students; 21 from the Civil Engineering Programme Area, and 88 from the School of Computing and Informatics, took part in the self and peer assessment exercise on public speaking. First year students were chosen as they were new to University education, allowing for the evaluation of the technique with students new to higher education and whether they would be willing to experience the technique through successive years of their University education.

Only 20 students from Civil Engineering and 62 students from the School of Computing completed both the self and peer assessments, with a total of 6486 peer assessments and 82 self-assessments being collected. Evaluation of the responses using a small sample modified Cochran Q Test showed that the subsequent results presented represent the student cohorts for a 5% margin of error with 96% confidence for Civil Engineering and 85% confidence for School of Computing and Informatics. Both cohorts were predominantly nationals of Brunei Darussalam, with just one or two international students from either Malaysia or the Emirate of Dubai.

Procedure

The speeches were marked by both the lecturer, the student themselves (self-assessment), and also other members of the class (peer-assessment). Previous research has indicated that student use of a rubric must include an element of training for the student to understand its implementation (Jonsson & Svingby, 2007; Reddy & Andrade, 2010) and prior to the assessment, the descriptors and indicators of the rubrics were presented to the students to guide them through the assessment exercise. The nature and the use of rubric and its criteria were explained, with focus on how to perform or score well in the public speaking assessment. To facilitate self and peer-assessment during the presentations, the rubric was replicated on Google forms and students were asked to assess their own public speaking skills and their peers immediately after each speaker had completed the assessment. The self-assessment mark recorded came from the individual student whilst the peer-assessment mark was the average peer mark across the cohort.

Subsequent to the assessment, students' opinions were obtained through the completion of an anonymised questionnaire with open and closed-ended questions in relation to the self and peer-assessment. The questionnaire used questions based on a Likert seven point scale (Likert, 1932) and 71.6% of students completed the questionnaire (66.7% from Civil Engineering, and 72.7% from the School of Computing and Informatics).

Data Analysis

Basic descriptive data analysis such as mean and standard deviation was calculated using Microsoft Excel, sample similarity was evaluated using Analysis of Variance (ANOVA) tests with a significance level (α) of 0.05, and correlation analysis was carried out using the Spearman's correlation coefficient to identify qualitative trends within the data. A value of 1 illustrated a perfect correlation between both variables, meaning that an increase in one was found to indicate an increase in the other, whilst

a value of -1 indicated a perfect anti-correlation between the variables, indicates that as one variable increased, the corresponding response for the second variable decreased. As the significance of the coefficient varies with sample size, results were taken to be statistically significant based on the data provided in Zar (1984) using $\alpha \leq 0.025$ unless otherwise stated in the text.

RESULTS AND DISCUSSION

Use of Google Forms

Whilst the use of Google forms made the collection of marks simple for subsequent analysis, all analysis of results such as averaging etc was undertaken using Microsoft Excel. During the exercise students added in their name and that of the student they were assessing, and one issue encountered during the post-processing of data was that in a minority of cases students used different names to describe themselves, such as a nickname, which required an element of manual interpretation when formulating the results. The main challenge encountered during the real-time assessment were poor network connectivity in the classroom. An additional challenge was network congestion for the larger cohort as every student tried to access the form simultaneously (n=88 for the Computer Science students). The implementation of Google forms was simple for both lecturer and students alike, and was a useful platform for an institution without direct access to an online learning platform.

Correlation Between Lecturer, Self and Peer-Assessment

The average results for each cohort is provided in Table 1 for peer, self and lecturer-assessment. For the Civil Engineering cohort, all but one student participated in both assessments, however for the Computer Science cohort, 26 students (30%) did not complete the online self-assessment. As a result, the data for both cohorts is presented with a reduced sample size compared to the class size.

As can be seen from Table 1, the averages awarded to each cohort by the individual assessment methods (peer, self and lecturer) are consistent with little deviation in average. In comparing the cohort average between Civil Engineering and Computer Science, the marking using both peer and self-assessment are both considerably higher than the lecturer assessment, with peer and self-assessment averages being within 10% for both assessment methods. These findings are similar to other studies in other research contexts and disciplines, with a number of authors reporting higher marks awarded through self-assessment in comparison to the lecturer or tutor (Bringula & Moraga, 2017). When comparing the marks awarded across the two student cohorts for each of the assessors, statistical comparison using an ANOVA one-way test found that whilst the results are statistically similar for both self and lecturer marking between the two cohorts, they are not similar for peer assessment.

Both the lecturer and self-assessment viewed the performance of each cohort as statistically similar, however this is not the same for peer assessment, this is likely due to the presence of statistical outliers in the peer marks for the School of Computing, not present for the Civil Engineering students. Evaluation of the third moment (skewness) of the peer assessment results highlights this, with significantly different values of -1.59 for School of Computing and 0.36 for Civil Engineering. This indicates that even with relatively high class sizes of 62 outliers can skew the data and impact on comparative data analysis. Table 2 shows the results of analysis of variance

Table 1. Average marks for each cohort based on peer, self and lecturer assessment

Course	N	Peer		Self		Lecturer	
		M	SD	M	SD	M	SD
Civil Engineering	20	77	5	73	16	56	11
Computer Science	62	73	6	71	14	57	11

Table 2. Results of Analysis of Variance (ANOVA) one-way test performed with α (two tailed) = 0.05 for similarity between each cohort for peer, self and lecturer assessment

Course	F	P-Value	F*crit*	Statistically Similar (?)
Peer assessment	5.58	0.02	3.96	No
Self-assessment	0.44	0.51	3.96	Yes
Lecturer	0.12	0.73	3.96	Yes

(ANOVA) one-way test performed with α (two tailed) = 0.05 for similarity between each cohort for peer, self and lecturer assessment.

These results indicate that whilst there seems to be consistency in marking between the two cohorts based on the average marks, there is a wide discrepancy between self and peer assessment and the marks awarded by the lecturer. This illustrates poor rater reliability, similar to other work (Andrade, 2005; Moskal & Leydens, 2000) and the work presented here calls into question the assertion that when using peer-assessment, only assessment marking with substantially different marks given by the assessors should require reassessment by the teacher to ensure reliability (Asikainen et al., 2014). Indeed, it was thought that through multiple peer assessment there would be an averaging of any bias in peer marking such that it converged towards that of the lecturer mark. Even with a rubric to guide students, or model answers, students seem to be on average more generous than the lecturer independent of degree discipline. A more detailed description of the marking is shown in Figure 1, which provides a direct comparison between student marks and those of the lecturer for both peer and self-assessment in both cohorts. There is no match between the lecturer and either cohort, with the average difference summarised in Table 3. Whilst peer assessment for an individual student is always found to provide marks above the lecturer, self-assessment shows both over marking and under marking. The differences between each approach to marking is statistically similar between each cohort, as shown in Table 4 This indicates that the discrepancies are statistically repeatable with student background not having an impact on the success or otherwise of the assessment reliability.

Figure 1. Comparison of self and peer assessment against lecturer marks

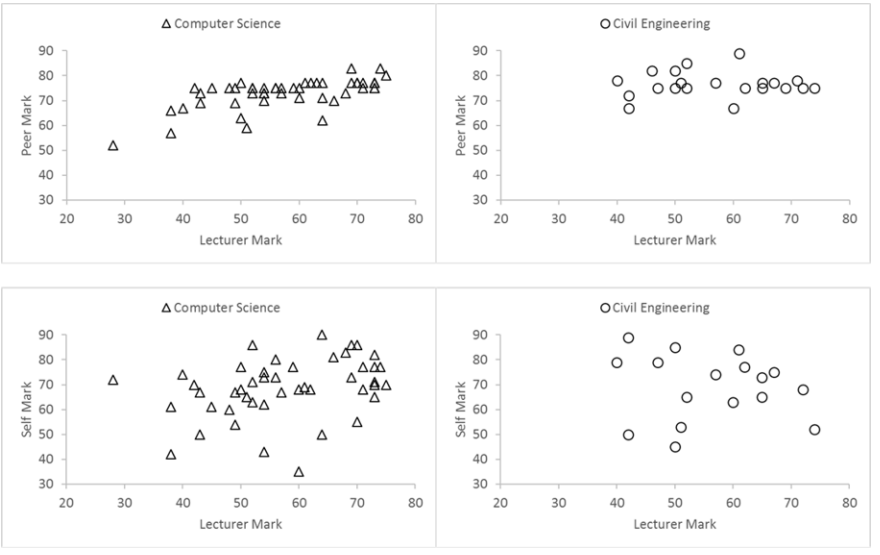


Table 3. Average difference in marks between assessors for each cohort

Course	Self / Lecturer		Peer / Lecturer		Self / Peer	
	M	SD	M	SD	M	SD
Civil Engineering (n = 20)	20	12	20	12	12	10
Computer Science (n = 62)	16	11	16	9	10	8

Table 4. Results of Analysis of Variance (ANOVA) one way test performed with α (two-tailed) = 0.05 for similarity in the difference between assessor marks for each cohort

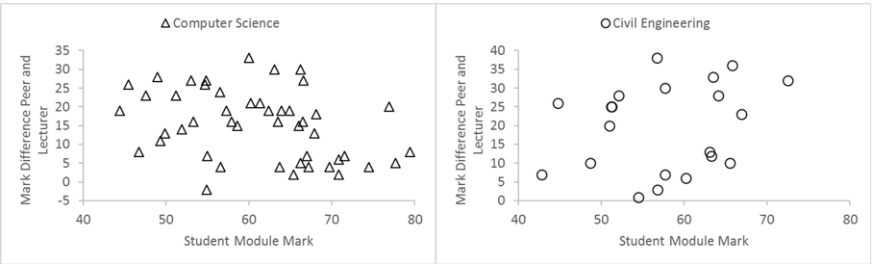
Course	F	P-Value	F*crit*	Statistically Similar (?)
Peer and lecturer	3.63	0.06	3.96	Yes
Self and lecturer	0.74	0.39	3.96	Yes
Peer and self	0.11	0.74	3.96	Yes

As the students were unable to quantitatively capture their own performance or that of their peers in comparison to the lecturer, there was no effort to calculate the Pearson coefficient which evaluates interval data. Instead, the ability of students to differentiate qualitatively the performance within the two cohorts was undertaken using an ordinal Spearman correlation approach. For Computer Science students, the Spearman correlation for the data is significant for both peer (0.64) and self (0.38). This indicates that for the Computer Science cohort, whilst the marking between assessors was different, the differentiation of abilities within the cohort was partially captured. This qualitative differentiation of ability was more accurately captured in the peer assessment with multiple assessors than for self-assessment with only a single assessor. No similar trend was found for Civil Engineering, and it is unclear whether this is due to the differences between student cohort or the smaller sample size. As a result, for the case of the Computer Science cohort the peer assessment captured qualitatively the ability of students within the cohort whilst failing to quantify ability through appropriate marking.

It is worth noting that self and peer assessments can be skewed, since assessing peers could be regarded as having to make negative remarks on peers. Some studies have reported that weaker students tend to over evaluate themselves and higher performing students underestimate their ability (Kwan & Leung, 1996; Sadler & Good, 2006) although this wasn't observed elsewhere for computer science students (Bringula and Moraga, 2017). Analysis of the results obtained in this work found no statistically significant evidence of positive or negative trends in elevated marking for either cohort between self-assessment and student ability, based on student performance across the module, supporting the results of the latter study. However, this was not the case for peer assessment, and Figure 2 presents results for peer assessment overmarking compared to student performance across the module for both cohorts. Whilst for Civil Engineering, the calculated Spearman rank coefficient was not statistically significant, for Computer Science there was a statistically significant negative correlation of -0.39. This indicates that for peer evaluation, students over evaluated the weaker students when compared to the higher performing students in a manner similar to that observed elsewhere for self-assessment. As this correlation was only observed for one cohort, as with the analysis of results provided in Figure 1 and presented previously, there can be no confirmation that this trend is independent of student cohort.

It is possible that the failure to accurately capture performance in a result of one or two specific subsections within the rubric, and study by other researchers (De Grez, Valcke, & Berings, 2010) observed that students found content-related criteria more difficult to assess as compared to delivery

Figure 2. Comparison of difference between peer and lecturer mark with student performance throughout the module



related criteria. In this work content-related is defined as the following sections; Introduction; Conclusion; Central Idea and Purpose; Content and Originality whilst delivery related content would be assessed by; appearance; body language; eye contact; language and voice; pacing and timing. The results in Figure 3 show the agreement between lecturer to peer assessment, lecturer to self-assessment and peer to self-assessment for each category within the rubrics. Results in general indicate a higher agreement between lecturer and both peer and self-assessment for both cohorts when comparing delivery to content criteria, supporting this previous study. However, the trend is less evident for the Civil Engineering students, and for self-assessment further evaluation of the marking indicating significant differences in the assessment of appearance, eye contact, and pacing and timing. Evaluation of the agreement between assessors within plus or minus one grade indicates a higher level of similarity with content related criteria. These results indicate that there are significant discrepancies in marking within all sections, and whilst one cohort indicates a bias towards increased agreement with the lecturer in delivery related criteria supporting the findings of previous work, this is not the case for the other cohort of students. These results indicate the difficulty in evaluating an assessment which has no defined correct answer but instead is partially open to the interpretation and prior experience of the assessor. With both cohorts there was a pronounced disagreement between the lecturer and peer and self-assessment for pacing and timing. One possible explanation for this could be the more relaxed attitude taken in South East Asia to the concept of time keeping and punctuality. This indicates a potential bias in marking based on cultural differences that might not be present in work evaluating the approach in a Western context. Overall, there is very poor agreement between lecturer and peer or self-assessment for all categories.

The poor correlation between different markers observed here highlights the challenge in the development of a thorough methodology for the implementation of robust and repeatable student self and peer-assessment. To adequately capture the development of student learning, rubric validity is important, with minimal student mark sensitivity based on assessor. Whilst it was thought that the application of multiple peer assessment might reduce the discrepancies between lecturer and peer assessment, this has not proven to be the case. The use of multiple peer evaluations was seen to smooth out individual subjectivity within the marking of specific students, however students were consistently found to over mark work by at least 10% when compared to the lecturer. This over marking identifies a key weakness of the technique for application in a summative setting, with the issue being independent of the subject of study. It is likely that the difference in marking is the result of a different perspective between the lecturer and students. How the technique extrapolates to other less technically minded courses such as business is uncertain, and will be the subject of further study.

Student Experience of Learning

A detailed summary of the results from the questionnaire is presented in Appendix B, Table 6, and Figure 4 presents the findings in relation to three main themes:

1. Students' opinions of the rubrics and peer and self-assessment;
2. Students' understanding of assessment;
3. Students' experiences on the overall peer and self-assessments.

The students were all non-native English speakers (L2) and the vast majority of them (at least 80%) had previous experience of peer and self-assessment. Application of ANOVA tests indicated that for all questionnaire questions, the responses of both cohorts were statistically similar for α (two tailed) of 0.05.

Students' Opinions of the Rubrics and Peer and Self-Assessment

The students felt that the rubric was written in a clear manner that was easy to understand, and that they took a serious and critical attitude towards its implementation during assessment. Computer Science students were found to be slightly less receptive towards the assessment compared to Civil Engineering students. One of the main reasons could be due to the number of peer-assessments each student had to do (up to 80 individual assessments for the School of Computing and Informatics) and the procedure could be perceived as wearying for the students. Whilst this would also indicate a reason for the observed poor mark correlation between students and lecturer, there is no reason why the students' "fatigue" should be any greater than that of the lecturer.

Whilst students thought that self-assessment was worthwhile, they did not think they had enough knowledge to evaluate themselves which is similar to the findings elsewhere (Cheng and Warren, 2005) and is in concordance with the discrepancy between marking highlighted previously. One contradiction is that the students did not think they had enough knowledge to evaluate themselves, however they then felt that their peers did, which could be the result of poor self-confidence.

Generally, students valued their peer feedback as expressed in the questionnaire open-ended responses and believed the rubrics helped with the students' progress. Both cohorts of students expressed interest in the use of rubrics in the assessment and for its consistency and fairness. They found the descriptors informative, clear and understandable, and students were able to evaluate their performance as well as that of their peers. There were however some uncertainties towards the awarding of mark as students were unable to decide the criteria's expectations between the marks range, as illustrated in the excerpt below:

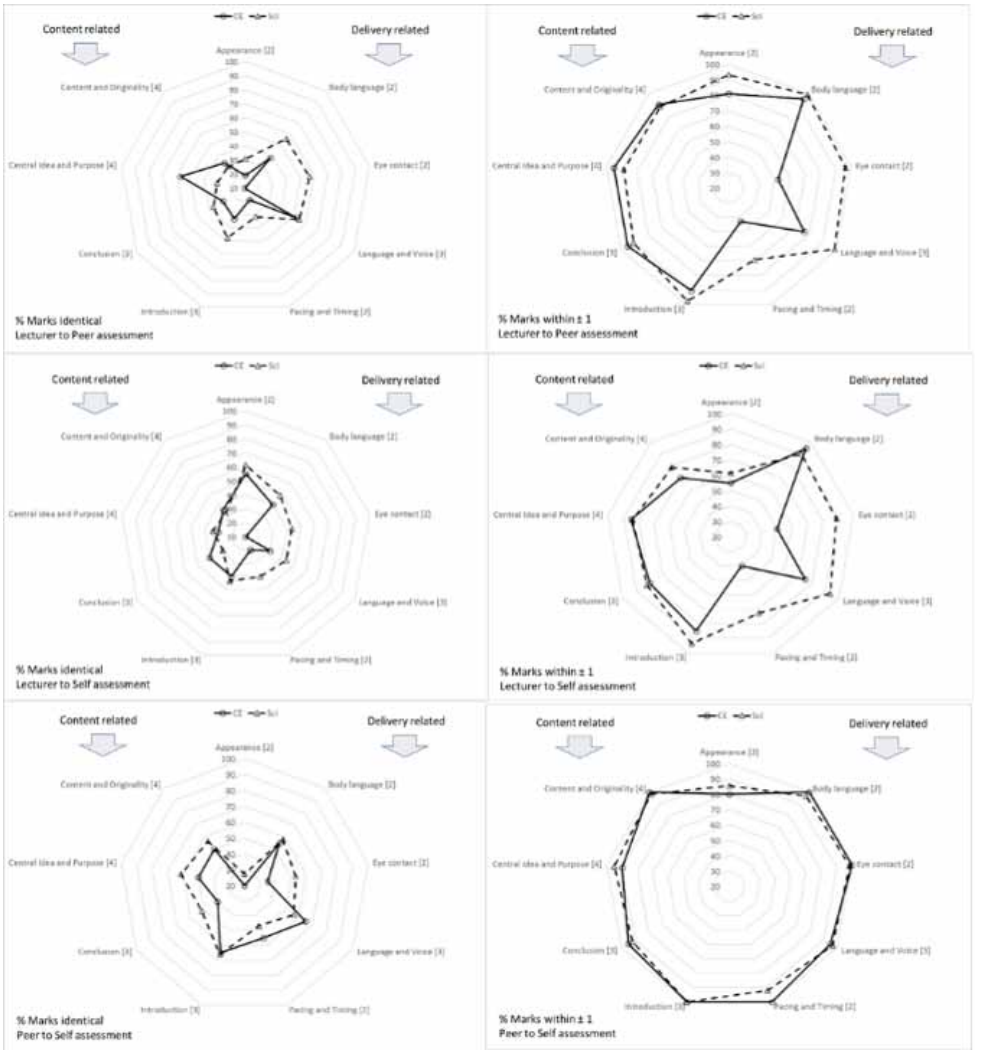
The rubric describes the difference between mark levels quite nicely, though sometimes it can be a little bit confusing because the mark I'm giving might fall between two levels and I cannot choose which one fairly.

This correlates well with the difference in marks observed between assessors, and especially those presented in Figure 3. These opinions strengthen the view that whilst not suitable for summative assessments, the use of peer-assessment is of value for student learning in a formative environment. There is however no evidence to suggest that this approach led to an increased retention of knowledge and understanding.

Students' Understanding of Assessment

In general, students felt that the peer-assessment was a worthwhile activity and both cohorts found the use of peer assessment to be a useful activity that aided their learning and enhanced their understanding of the module content. Despite this positive impact, both Civil Engineering and Computer Science cohorts indicated they were indifferent to their participation in the exercise and only weakly positive about the implementation of the exercise in subsequent modules. The questionnaire response on future implementation is quantitatively similar to that presented by Davey (2015) for students at an Australian University who reported using a similar Likert scale that students were keen to have

Figure 3. Percentage similarity in marking within each category for each cohort for (left) marks being exact and (right) marks being within ± 1 based on the rubrics

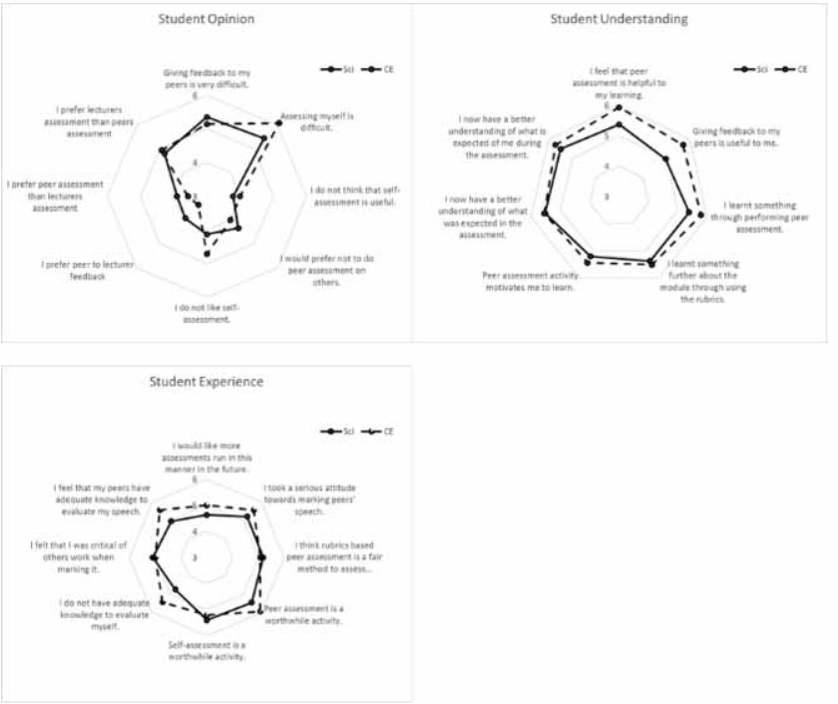


self-assessment in other courses. Whilst both cohorts found the process of peer-assessment useful, they still expressed a preference to lecturer assessment rather than peer-assessment. What is not so obvious however is whether the students realised that through applying the rubrics in this manner they were obtaining much more detailed indirect feedback from the lecturer on both their and peers work. The evaluation of this self-awareness would be a useful indicator of student understanding and self-reflection, and could be the focus of a future longitudinal study. This would then provide insight into the effect on student's long-term development and ability to reflect themselves on work rather than relying on others.

Students' Experiences on the Overall Peer and Self-Assessments

Students found the process of both self and peer-assessments difficult, yet both cohorts were positive about the learning benefits of the exercise with Civil Engineering students more positive about the learning impact than Computer Science students. These results support the view that a careful and

Figure 4. Average student cohort response to questionnaire questions



well-designed rubric can promote students' learning, enhancing the teaching and learning process whilst stimulating thinking processes (Andrade, 2000; Jonsson & Svingby, 2007). Civil Engineering students found self-assessment more difficult than peer assessment, and in general Civil Engineering students were more negative about self-assessment and more positive about peer assessment, yet they also reported a better perceived impact on their learning.

Some students also expressed concern towards the lack of anonymity in assessing their peers. They expressed hesitations towards peer assessment because of biasness or friendship, a similar problem highlighted by Langan et al., (2005), and which could explain the prevalence of over marking during the peer assessment.

CONCLUSION

The use of a uniform pedagogical approach has shown that self and multi peer-assessment does not closely mimic that of the lecturer for two student cohorts studying public speaking. The implementation of the multi peer-assessment was found to be easy using Google forms, with only a few issues relating to internet connectivity and spend with large numbers of users. The use of this approach did not reduce the elevated marks awarded through peer-assessment reported in previous studies, but it did smooth out individual subjectively in peer evaluation of a students work. This over marking was found independent of student cohort and marking for both self and peer-assessment was closer to that of the lecturer for delivery related criterion compared with content-related criteria. There were however notable exceptions, with self-assessment failing to accurately capture grading for appearance and both techniques failing to accurately capture pacing and timing. The latter is likely a result of the regional culture. This reinforces a growing body of literature indicating that peer and self-assessment in a range of contexts fails to accurately capture student performance, with students consistently over marking

by over 10%. Students found the experience of participating in self and peer-assessment beneficial, however they provided varying responses to perception of their ability and the level of difficulty in undertaking the exercise depending on cohort, and approach taken. Given the positive impact that the approach had on students and their experience, coupled with the poor agreement between lecturer mark and those obtained by multi peer and self-assessment, it would not be reliable to use either technique for summative assessment of students work. However, as a learning tool and comparative technique it could be used effectively in formative assessment to promote student learning. Further work will expand on this study to evaluate the marking reliability and learning experience with students from other less technical backgrounds, as well as to explore methods to increase student understanding of the rubric and mitigate student's tendency to over mark themselves and their peers.

REFERENCES

Andrade, H. G. (2000). Using rubrics to promote thinking and learning. *Educational Leadership*, *57*(5), 13–19. Retrieved from http://www.ascd.org/publications/educational-leadership/feb00/vol57/num05/Using-Rubrics-to-Promote-Thinking-and-Learning.aspx

Andrade, H. G. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College Teaching*, *53*(1), 27–31. doi:10.3200/CTCH.53.1.27-31

Andrade, H. G., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research & Evaluation*, *10*(3), 1–11. Retrieved from https://pareonline.net/pdf/v10n3.pdf

Andrade, H. G., Du, Y., & Wang, X. (2008). Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practice*, *27*(2), 3–13. doi:10.1111/j.1745-3992.2008.00118.x

Arendt, A., Trego, A., & Allred, J. (2016). Students reach beyond expectations with cafeteria style grading. *Journal of Applied Research in Higher Education*, *8*(1), 2–17. doi:10.1108/JARHE-03-2014-0048

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

Bringula, R., & Moraga, S. (2017). Utilization of Rubrics for Self-assessment of Computer Science Students Enrolled in a Research Writing Course. *Proceedings of the 22nd Western Canadian Conference on Computing Education* (p. 2). ACM. doi:10.1145/3085585.3085587

Cheng, W., & Warren, M. (2005). Peer assessment of language proficiency. *Language Testing*, *22*(1), 93–121. doi:10.1191/0265532205lt298oa

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, *98*(4), 891–901. doi:10.1037/0022-0663.98.4.891

Davey, K. R. (2011). Student peer assessment: Research findings from a case study in a master of chemical engineering coursework-program. *Education for Chemical Engineers*, *6*(4), e122–e131. doi:10.1016/j.ece.2011.08.004

Davey, K. R. (2015). Student self-assessment: Results from a research study in a level IV elective course in an accredited bachelor of chemical engineering. *Education for Chemical Engineers*, *10*, 20–32. doi:10.1016/j.ece.2014.10.001

Davey, K. R., & Palmer, E. (2012). Student Peer Assessment: A research study in a level III core course of the bachelor chemical engineering program. *Education for Chemical Engineers*, *7*(3), e85–e104. doi:10.1016/j.ece.2012.04.001

De Grez, L., Valcke, M., & Berings, D. (2010). Student response system and learning oral presentation skills. In *Procedia Social and Behavioral Sciences* (Vol. 2, pp. 1786–1789). Elsevier BV; doi:10.1016/j.sbspro.2010.03.985

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144. doi:10.1016/j.edurev.2007.05.002

Kwan, K., & Leung, R. W. (1996). Tutor versus peer group assessment of student performance in a simulation training exercise. *Assessment & Evaluation in Higher Education*, *21*(3), 205–214. doi:10.1080/0260293960210301

Langan, A. M., Wheater, C. P., Shaw, E. M., Haines, B. J., Cullen, W. R., Boyle, J. C., & Lockey, L. et al. (2005). Peer assessment of oral presentations: Effects of student gender, university affiliation and participation in the development of assessment criteria. *Assessment & Evaluation in Higher Education*, *30*(1), 21–34. doi:10.1080/0260293042003243878

Likert, R. (1932). A technique for the measurement of attitudes. *Archives de Psychologie*, *22*, 140. Retrieved from https://legacy.voteview.com/pdf/Likert_1932.pdf

Liu, N. F., & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education*, *11*(3), 279–290. doi:10.1080/13562510600680582

Moni, R. W., & Moni, K. B. (2008). Student perceptions and use of an assessment rubric for a group concept map in physiology. *Advances in Physiology Education*, *32*(1), 47–54. doi:10.1152/advan.00030.2007 PMID:18334568

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation*, *7*(10), 71–81. Retrieved from http://PAREonline.net/getvn.asp?v=7&n=10

Palenque, S. M. (2016). The Power of Podcasting: Perspectives on Pedagogy. *Journal of Institutional Research*, *5*(1), 4–7. doi:10.9743/JIR.2016.1

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, *35*(4), 435–448. doi:10.1080/02602930902862859

Sadler, P. M., & Good, E. (2006). The impact of self-and peer-grading on student learning. *Educational Assessment*, *11*(1), 1–31. doi:10.1207/s15326977ea1101_1

Sendziuk, P. (2010). Sink or Swim? Improving Student Learning through Feedback and Self-Assessment. *International Journal on Teaching and Learning in Higher Education*, *22*(3), 320–330. Retrieved from https://files.eric.ed.gov/fulltext/EJ938567.pdf

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation*, *9*(2), 1–10. Retrieved from http://PAREonline.net/getvn.asp?v=9&n=2

Zar, J. H. (1984). *Biostatistical analysis* (2nd ed.). Eaglewood Cliffs, NJ: Prentice-Hall, Inc.

APPENDIX A

Table 5. Public speaking rubric

Delivery	Exemplary [4]	Good [3]	Adequate [2]	Ineffective [1]	Poor [0]
Appearance [2]	• Relaxed • Display absolute confidence. • Appropriately dressed and relevant to topic	N/A	• . A little nervous • Quick recovery from minor mistakes. • Appropriately dressed within UTB dress code for assessment and examinations.	N/A	• Obviously nervous and is distracting the delivery. • Inappropriately dressed and against UTB dress code.
Body language [2]	• Good posture throughout the entire speech. • Body movements and gestures are natural, appropriate and relaxed.	• Good posture for most part of the speech. • Body movements are mostly natural, appropriate and relaxed. • Nervous movements do not interfere the speech.	• Good posture for some parts of the speech. • Body movements are inappropriate. • Nervous gestures are noticeable but do not obstruct the deliverance of speech.	• Poor posture and distracting and inappropriate body movements and gestures. • Unnatural and some awkward body movements.	• Bad posture and body position throughout the speech. • Body movements distracts from the overall speech. • No use of gestures to reinforce ideas or points.
Eye contact [2]	• Direct and consistent eye contact with all parts of the audience throughout the entire of the speech.	• Fairly consistent eye contact with audience. • Any lack of eye contact is only momentary.	• Eye contact with audience is somewhat inconsistent and distracting.	• Eye contact with audience is significantly lacking and inconsistent.	• No eye contact with audience.
Language & Voice [3]	• No grammatical errors. • Natural variation in volume, pitch and tone. • Maintains audience attention. • Speaker sounds genuinely interested in the topic. • Speaks very clearly and can be heard well throughout the speech.	• Very few grammatical errors. • Appropriate volume, pitch, inflection. • Speaks clearly but voice is too soft in a few parts of speech.	• Noticeable errors in sentence structure and grammar. • Some mispronunciation, inappropriate volume, pitch and inflection is noticeable but do not seriously hinder the delivery. • Voice is too soft in some parts of the speech.	• Distracting faults in sentence structure and language use. • Frequent mispronunciations and inappropriate, volume, pitch and inflection seriously hindering the delivery. • Voice is too soft in most parts of the speech.	• Inaudible and incomprehensible delivery with serious and frequent grammatical errors. • Delivery lacks any necessary emphasis, pitch or inflection. Monotone. • Voice is too soft throughout the entire speech.
Pacing & Timing [2]	• Pace is excellent throughout. • Dramatic pauses used well. • No hesitation. • Length matches allotted time. Well-timed.	• Pace is generally maintained. • Slight hesitations, very short, minimal pauses. • Delivery was appropriately paced.	• Pacing is somewhat inconsistent. • Some hesitations, pauses are apparent in some parts of the speech. • Speaker did not manage to cover all points within the allotted time.	• Pace is inconsistent; either too slow or fast. • Some hesitations, a lot of pauses in most part of the speech. • Speaker did not manage to cover most of the topic within the time frame.	• Either too fast or too slow; • A lot of hesitations, stumbles or staggers a lot. • Did not adhere to the time limit.

continued on following page

Table 5. Continued

Content	Exemplary [4]	Good [3]	Adequate [2]	Ineffective [1]	Poor [0]
Introduction [3]	• Full, complete introduction to topic. • Correctly reveals the purpose and major points of the speech and motivate the audience to listen.	• Clear brief introduction. • Attempts to reveal the purpose and major points of the speech and motivate the audience to listen.	• Somewhat short introduction. • Opening comments seem somewhat artificial, weak or unimaginative.	• Very short introduction statement. • Opening comments are inappropriate to the speech.	• No introduction. • No appropriate opening comments, neglecting audience or did not stimulate interest in speech.
Conclusion [3]	• Appropriate concluding statement. • Speaker summarises the main purpose and major points of the speech	• Appropriately related to the purpose and major points of the speech, but not very strong, convincing or emphatic.	• Includes unnecessary concluding statements or redundant information.	• Speech concludes abruptly with no concluding statements or end with inappropriate remarks.	• No conclusion.
Central Idea & Purpose [4]	• Maintain clear focus on central idea or topic. • Appropriate sources are used.	• Conveys a central idea or topic. • Focuses and orders the information to convey a generally unified point.	• Attempts to focus on a central idea or topic. • Provides some focus or order but the speech purpose is somewhat unclear.	• Provides very little or very minimal focus on central idea or topic. • Insufficient information or details.	• No focus or central idea or topic. • Unrelated information or details.
Content & Originality [4]	• Provides novel information on original topic. • Elaborate details to support central idea. • Informs and persuades effectively.	• Provides sufficient information or details. • Elaborate some details to support central idea.	• Provides marginal information on original topic. • Some related details but did not provide elaborations.	• Scarcely informative. • Information can be somewhat irrelevant.	• No apparent purpose. • Insufficient or unrelated details. • Lacks novelty or originality.

APPENDIX B

Table 6. Summary of number of student questionnaire responses and category of response for each survey statement. Likert 7 = strongly agree, 1 = strongly disagree, and; 4 = no opinion or neutral response.

Statement	Cohort	Survey Statement							≥ 5	Ave	SD
		1	2	3	4	5	6	7			
I have undertaken self-assessment previously	CE	0	0	1	1	1	6	5	12	5.93	1.21
	SCI	4	1	5	5	15	26	15	56	5.31	1.59
I have undertaken peer-assessment previously	CE	0	1	0	0	1	7	5	13	6.00	1.30
	SCI	5	1	1	7	15	25	17	57	5.38	1.62
I think that the rubric was written in a clear manner which allowed me to accurately mark my peers	CE	0	0	2	1	6	3	2	11	5.14	1.23
	SCI	1	1	2	8	27	22	10	59	5.32	1.17
The rubric is easy to understand	CE	0	0	1	2	4	5	2	11	5.36	1.15
	SCI	1	1	3	11	25	24	6	55	5.17	1.16
I feel that my peers have adequate knowledge to evaluate my speech (A)	CE	0	0	1	1	5	3	4	12	5.57	1.22
	SCI	2	1	2	18	22	22	4	48	4.96	1.22
I felt that I was critical of others work when marking it (A)	CE	1	0	1	3	2	5	2	9	5.00	1.66
	SCI	1	1	2	16	21	26	4	51	5.10	1.14
I would like more assessments run in this manner in the future (A)	CE	0	0	0	6	4	2	2	8	5.00	1.11
	SCI	4	2	4	21	21	13	6	40	4.63	1.45
I do not have adequate knowledge to evaluate myself (A)	CE	0	0	0	3	5	3	3	11	5.43	1.09
	SCI	3	2	5	24	15	11	11	37	4.73	1.51
I took a serious attitude towards marking peers' speech (A)	CE	0	0	0	3	3	5	3	11	5.57	1.09
	SCI	1	0	1	9	23	28	9	60	5.44	1.07
I think rubrics-based peer assessment is a fair method to assess student's performance (A)	CE	1	0	1	2	3	5	2	10	5.07	1.64
	SCI	0	3	1	12	26	22	7	55	5.18	1.14
Peer assessment in a worthwhile activity (A)	CE	0	0	0	0	3	9	2	14	5.93	0.62
	SCI	1	1	0	10	20	29	10	59	5.45	1.13
Self-assessment is a worthwhile activity (A)	CE	0	0	0	3	5	6	0	11	5.21	0.80
	SCI	1	1	1	10	20	28	10	58	5.41	1.17
Having used the rubrics, I now have a better understanding of what was expected of me in the assessment (B)	CE	0	0	1	1	3	0	4	7	5.56	1.51
	SCI	1	1	0	7	24	25	13	62	5.52	1.13
Having marked the public speaking using a rubric, I now have a better understanding of what is expected of me during the assessment (B)	CE	0	0	0	2	3	6	3	12	5.71	0.99
	SCI	1	0	1	6	25	29	9	63	5.49	1.03
Peer assessment activity motivates me to learn (B)	CE	0	0	0	5	1	5	3	9	5.43	1.22
	SCI	2	2	3	8	20	30	6	56	5.20	1.32

David Hassell is currently a Senior Teaching Fellow in the department of Chemical Engineering at the University of Bath.

Kok Yueh Lee is a lecturer at the Centre for Communication, Teaching and Learning at Universiti Teknologi Brunei. Her research interests include teaching and learning pedagogies, assessment in higher education, discourse analysis, genre analysis, and corpus analysis.