UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Journal of Peer Learning

Volume 6 Article 2

2013

Student acceptance and application of peer assessment in a final year genetics undergraduate oral presentation

Heather Verkade
Monash University, Heather.Verkade@monash.edu.au

Robert J. Bryson-Richardson
Monash University

Recommended Citation
Verkade, Heather and Bryson-Richardson, Robert J., Student acceptance and application of peer assessment in a final year genetics undergraduate oral presentation, *Journal of Peer Learning*, 6, 2013.
Available at:http://ro.uow.edu.au/ajpl/vol6/iss1/2

Journal of Peer Learning (2013-14) 6: 1-18

Student acceptance and application of peer assessment in a final year genetics undergraduate oral presentation

Heather Verkade and Robert J. Bryson-Richardson

ABSTRACT

Undergraduate students benefit from observation of each other's oral presentations through both exposure to content and observation of presentation style. In order to increase the engagement and reflection of final year students in an oral presentation task, a peer assessment component was introduced using a rubric that emphasised scientific skills over presentation quality. This study investigated the effect of peer assessment on students' reported motivation and reflection, and their level of acceptance of peer evaluation of an oral presentation. As a result of peer assessment, students reported being more engaged, feeling a sense of responsibility, and many felt that they reflected more on their own talk. Students considered presentation style over scientific quality and demonstrated a strong reticence to award low marks. The impact on the final marks was mitigated by using a 20% weighting on the peer assessment, a level that the majority of students considered acceptable. This analysis suggests that peer assessment can achieve the intended learning outcomes. This paper provides a suggested process for using peer assessment in oral presentations with a strong science component and discusses approaches to examine and mitigate the observed student reticence to award low marks.

INTRODUCTION

Developing skills in oral presentations is an objective of many undergraduate courses (Higher Education Council, 1992). Oral presentations to peers not only benefit the speaker but also the audience, who have the opportunity to learn from the material presented and from observation of other approaches to the task. In our experience, however, students do not necessarily pay attention to each other's presentations unless given a specific requirement or responsibility. Peer assessment is a well-accepted assessment tool in higher education (Boud & Falchikov, 2007; Falchikov, 2005) that can increase the engagement of students with the assessment process and the students' sense of responsibility in their own learning by emphasising student-centered learning (Goldschmid & Goldschmid, 1976; MacAlpine, 1999; Napan & Mamula-Stojnic, 2005; Orsmond, 2011; Patton, 2012; Stanier, 1997; Vickerman, 2009). This study examines student levels of acceptance of peer assessment of oral presentations and whether it causes them to change their approach to their own presentation and to watching other students' presentations.

Published by the University of Wollongong
ISSN 2200-2359 (online)

Peer assessment is simply a situation in which students judge the work of their peers (Falchikov & Goldfinch, 2000). Peer assessment of oral presentations is particularly valued when teaching for professions that require excellent communication skills, including business, law, science, and education (Falchikov & Goldfinch, 2000). Assessment of oral presentations in final year science courses prioritises scientific skills above general presentation skills of voice control, body language, eye contact, and use of slides, which are often more strongly emphasized in the assessment of oral presentations in other settings (Campbell, Mothersbaugh, Brammer, & Taylor, 2001; MacAlpine, 1999; MacPherson, 1999). The scientific skills that we and other scientists value most are: a) researching the literature to identify the relevant experiments, b) explaining the experiments, their conclusions, and their context, and c) answering pertinent questions posed by the audience (Orsmond, 2011). To encourage application of these skills, we developed our rubric to reflect the importance of these criteria. Here we have examined the application of a rubric, particularly in light of the criteria that the students apply when using the rubric to assess their peers.

In this study, we have evaluated peer assessment of an oral presentation assessment task focusing on human genetic disorders. We had four intersecting goals for introducing a peer component into this assessment task, the first two of which we have addressed directly in this study, and the second two of which have been considered by asking students for their opinions about their behaviour due to the presence of peer assessment.

- To increase student motivation in their talk preparation or presentation
- To cause the students to reflect more on their own talk as a result of close observation of how other students give their presentations
- To encourage the students to pay closer attention to the other students' talks so that they may learn more about human genetic disorders
- To induce students to consult and consider marking criteria, and thus learn to more carefully use the criteria in future assessment tasks

We had several concerns with introducing peer assessment, and these guided our four research questions.

1. Are the students accepting of peer assessment as implemented in this unit?
2. Do students consider that they change their approach to their presentation, or reflect more on their presentation, due to peer assessment?
3. Does the inclusion of peer marks alter the final assessment mark?
4. How do students use and apply the rubric?

We aimed to understand how accepting the students are in participating in the peer assessment process, as any major concerns they had about the peer marks may negate our anticipated learning outcomes. In particular, the students need to be comfortable with the level of mark for which they are responsible and with the ability of their peers to assess them. This can be a particular concern in a final year science subject/unit in which students are expected to provide significant scientific content and reasoning in their oral presentations. Concern that students may not view their fellow students as suitably experienced judges of these components guided our analysis. Two of our major goals for introducing peer learning were that the students would become more reflective on their own presentation and become more attentive and engaged with the other student's presentations. In particular, students need to feel a sense of responsibility in the activity before they will commit the effort to assess their peers well (Dochy, Segers, & Sluijsmans, 1999).

Validity is a common concern with the introduction of peer assessment (Topping, 1998), and we shared this concern. We would define validity as being a marking approach that provides a mark that truly represents the quality of the presentation, as defined in the meta-analysis of peer assessment of Falchikov and Goldfinch (2000). The validity of the assessment depends strongly on the design and implementation of the assessment criteria or rubric (Falchikov and Goldfinch, 2000), and the rubric used in this study was one that focused heavily on the ability of the students to interpret and explain findings from the scientific literature. To assess the validity of the peer assessment marks, we compared these to the marks given by the academics, and assessed if there were inadvertent negative consequences of peer assessment. Because it is important to ensure that the final mark given to the students is as accurate a representation of the quality of their oral presentation as possible, a possible negative consequence would be a dramatic alteration of the marks compared to the marks given by the academic. In particular, we are aware of anecdotal concerns that peers are hard markers (Ballantyne, Hughes, & Mylonas, 2002). In testing whether the marks are significantly altered compared to using no peer assessment, we looked for differences between the final marks (combined peer and academic's marks) and academics' marks alone. This approach therefore uses the academics' marks as the standard against which the other marks are compared. Naturally, there is variation between the marks that different academics would give students (Falchikov & Magin, 1997; Smith, Cooper, & Lancaster, 2002), and academics' marks cannot truly be considered free from variation and bias. The alternative to peer assessment is to have an academic assess each talk and have this taken as the sole mark. For this reason, although the academics' marks may not be perfect, they were taken as the standard against which the students' peer marks were compared.

METHODS

The assessment task

This unit/subject is in first semester, final year, within the Bachelor of Science, and it focuses on human medical genetics. The students were given five weeks to prepare an oral presentation on an assigned genetic disorder. The requirements were to describe the disorder, the gene that is responsible, and the most important research findings from the literature (how the gene was identified, the function of the gene product, and why a mutation in the gene causes this specific disorder). The task was a 10 minute (plus 5 minutes for questions) oral presentation, which was worth 10% of the total mark of the unit/subject in a class of 88 students (17 males/71 females). Each presentation was assessed by 12-19 peers, and the average of these peer marks provided 20% of the students' mark for the oral presentation, with the remaining 80% of the mark given by an attending academic. There were four academics used to mark the classes, with each academic marking two to four groups. The range of marks given by each academic showed a very similar mean and median. Each student was marked by one academic. Both students and academics marked using a rubric of five criteria with a 4-point scale from *poor* to *excellent*, which was provided to students both in advance and at the start of the presentations (Table A1). A brief summary of the rubric was present on all scoring sheets (Table 1). The final score was calculated using the formula: (∑[quantified score for criteria × multiplier])/3 to give a mark out of 10. Scores for quantification of the rubric were: Excellent = 3, Good = 2.25, Adequate = 1.5, Poor/Fail = 0.75. The multipliers for the rubric were: Quality of presentation 1, Identification of key experimental evidence 1, Description of the experiments 4, Explanation of conclusions 3, Ability to answer questions 1, summing to a total of 10.

Statistical analysis

One-Sample Kolmogorov-Smirnov tests were used to determine that the distributions of marks were normal (n = 88, averaged marking by peers, Z = 0.943, p = .337, of peers, Z = 0.509, p = .958, and by academics Z = 0.697, p = .716, asymptotic significance, 2-tailed). As the distributions were normal, parametric tests were used to test for correlation (two-tailed Pearson's product-moment tests). Modifications to the rubrics were assessed using paired t-tests (two-tailed).

The students were given a voluntary questionnaire about the exercise, which 75 students out of 88 (85%) completed (Figure A1). As these were not in normal distribution, correlations between the questions were examined using Spearman's rank order test. Medians were compared using a Mann-Whitney U test. Means of 5-point Likert scales are presented in the text ± standard deviation. Percentages of *agree*, *neutral*, and *disagree* were generated by converting the 5-point scale to a 3-point Likert scale. Statistical analysis was done using Microsoft Excel 2010 and IBM SPSS Statistics 20.

Table 1
Marking rubric used by academics and students, showing marks assigned for each criterion and the conversion multiplier used to quantify each criterion

		Quantification of descriptor for mark				
Multiplier	**Assessment Criteria**	**Excellent**	**Good**	**Adequate**	**Poor**	**Comments**
1	Quality of presentation	3	2.25	1.5	0.75	
1	Identification of key experimental evidence	3	2.25	1.5	0.75	
4	Descriptions of the experiments	3	2.25	1.5	0.75	
3	Explanations of the conclusions	3	2.25	1.5	0.75	
1	Ability to answer questions	3	2.25	1.5	0.75	

Note. The final mark was calculated by multiplying the mark for each criterion by the appropriate multiplier, summing these and dividing the total by 3 to give a mark out of 10. This mark provided 20% of the student's mark for the oral presentation. The other 80% came from the mark given by the academic.

RESULTS

The students have mixed opinions of the peer assessment process

In order to probe the responses of the students to the peer assessment task, a questionnaire was administered at the beginning of the oral presentation sessions (Figure A1, Figure 1, Table 2) Seventy-two (96%) of the students reported that they had peer assessed previously in their degree, while 4% reported never or were unsure. They were asked the maximum proportion of total marks for an assessment, such as this one, that they are comfortable to receive from their peers, with the answers constrained to 0%, 20%, 40%, 60%, 80%, and 100%. As this assessment used 20% peer mark, it was reassuring to see that 35 (48%) were comfortable with a maximum of 20%, and 28 (38%) would be comfortable with higher percentages, resulting in a total of 86% of students comfortable with the level of peer assessment in this study. Interestingly, none were comfortable with 100% peer mark. Ten students (14%) wanted 0%, indicating that they did not want peer assessment at all, with three of these students neutral and four disagreeing with Question 4, "My peers are well qualified to provide an assessment of my talk," showing a mean of 2.89 and a median of 3. This differs significantly from the 63 students who were happy with 20% or more, 23% of whom were neutral and 6.5% disagreed with Question 4 at a mean of 3.55 and a median of 4, *Z(71)* = -2.293, *p* = .022. Two of the students who wanted 0% peer assessment agreed with Question 4, suggesting that they feel their peers are qualified to assess them, but they still do not want peer assessment to happen, possibly because of the perceived responsibility for themselves. Indeed, a student reported that "I don't like having the responsibility of contributing to someone else's mark."

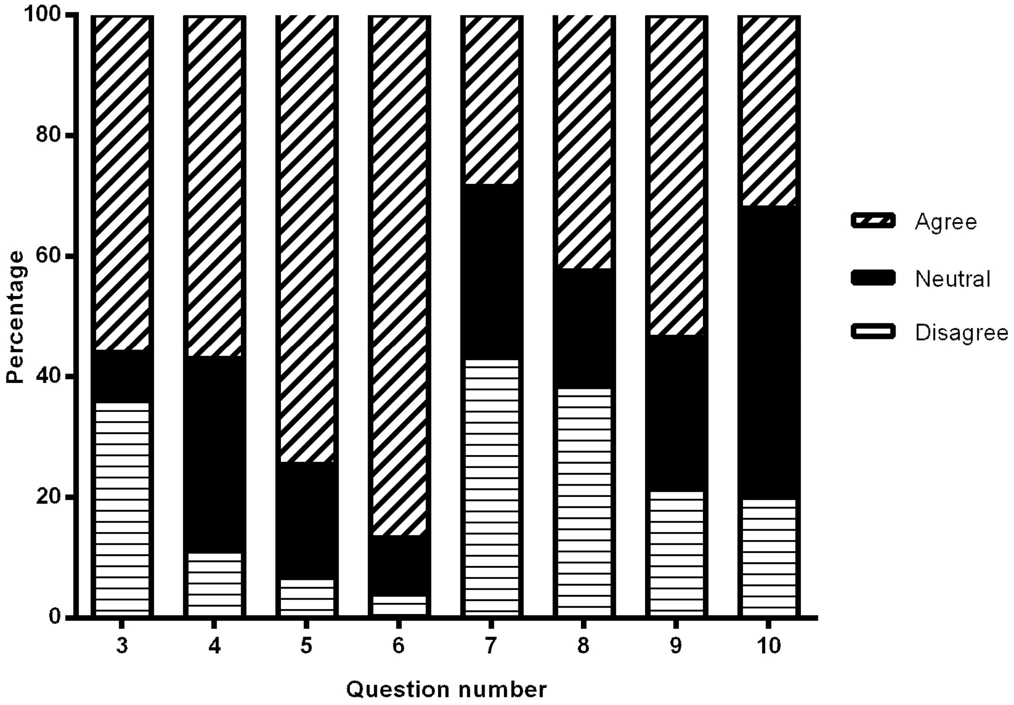

Figure 1. Questionnaire results for questions 3 – 10 in a 3-point Likert scale (See Figure A1 for full questionnaire).

Several students commented that they are concerned about harsh peer marking: “Some peer marking I found to be really critical and unfair,” and “I find that peers can be harsher markers even though they make the same mistakes.” Both these students wanted 0% peer marking, but interestingly were neutral on whether their peers were qualified to assess them. In fact, a student stated that “I agree that it's good and helpful, but some people are VERY harsh markers!” Yet this student also felt that their peers were qualified to peer assess and was comfortable with up to 40% peer marking, suggesting that the feelings of individual students towards peer marking can be complex.

Disappointingly, only 49 (56%) of the students said that they were aware that peer assessment was part of this assessment task (Question 3), despite this being explained in verbal and written instructions, weeks ahead of the task. However, there was no statistically significant correlation between the answers for Question 3 and Question 7, indicating that students who feel the need to modify their talk because of peer assessment are no more likely to take in the information that an assessment task is peer assessed. Overall, 41 (57%) of the students felt that their peers were well qualified to assess their talk (Question 4). Gratifyingly, only eight (11%) disagreed with this question, suggesting that 64 (89%) are at least amenable to the idea of being peer assessed, showing excellent agreement with previous studies (Søndergaard, 2009).

Table 2
Questionnaire results in percentages (See Figure A1 for full questionnaire)

Question	**not sure**	**0**	**1 - 2**	**3+**			***n***	
1	2.7	1.3	38.7	57.3			75	
	0	**20**	**40**	**60**	**80**	**100**	***n***	
2	13.7	48.0	31.5	4.1	2.7	0	73	
	Strongly Agree	**Agree**	**Neutral**	**Disagree**	**Strongly Disagree**		***n***	**Mean ± SD**
3	28.0	28.0	8.0	25.3	10.7		75	3.4 ± 1.4
4	5.6	51.4	31.9	11.1	0.0		72	3.5 ± 0.9
5	20.0	54.7	18.7	6.6	0.0		75	3.9 ± 0.8
6	26.7	60.0	9.3	4.0	0.0		75	4.1 ± 0.7
7	9.5	18.9	28.4	35.1	8.1		74	2.8 ± 1.2
8	13.7	28.8	19.1	34.3	4.1		73	3.1 ± 1.2
9	9.3	44.0	25.3	18.7	2.7		75	3.4 ± 1.0
10	8.0	24.0	48.0	14.7	5.3		75	3.2 ± 1.0

Note. SD = standard deviation.

Students reported increased involvement and reflection due to peer assessment

Two of our primary goals in adding peer assessment were to increase the students' involvement in the other presentations and to increase their level of self-reflection. It is very promising that 56 (75%) students agreed that the way they watch other students' talks changes because they are assessing the talk, and 65 (87%) agreed that they feel a responsibility because they are assessing other students' talks (Questions 5 and 6). These two questions showed significant correlation (r = .389, p < .001), strongly supporting that a large group of the students are altering their behavior while watching their peers give their presentations. The students were mixed as to whether they modified the preparation (Question 7) or presentation (Question 8) of their talk, with only 21 (28%) agreeing to Question 7, while 32 (43%) felt that they did not, and 21 (28%) were neutral. Question 7 and Question 8 showed strong inter-item correlation (r = .728, p < .0001). Interestingly, there was no statistically significant correlation between the students' answers to questions 5 and 6 (a change in the way they watch other talks) when compared to questions 7 and 8 (a change to their own presentation). Pleasingly, 40 (53%) felt that they reflected more on their own talk due to the peer assessment process, with only 16 (21%) feeling that they did not (Question 9), which suggests that one of our major aims was successful. Students commented that peer assessment "gets you involved more!," that you are "more likely to listen, and ask questions," that "it is a good way to ensure [the] audience pays attention and as a result are able to ask meaningful questions," and even that it "forces you to pay close attention to other talks."

The students were equivocal on whether their own talk improved due to peer assessment, with 36 (48%) neutral, 24 (32%) agreeing and 15 (20%) disagreeing (Question 10), but there was a strong correlation between those who felt it improved their talk and those who reflected more on their own talk (r = .569, p < .0001), which was consistent with our expectations of the value of peer assessment. This is despite a lack of correlation between students who changed the way they viewed other's talks, and those who altered their own talks. Interestingly, there was also a highly significant correlation between students who think that their peers are qualified to assess them and those who think that peer assessment improved their talk (r = .376, p < .001), indicating a general positive impression from this subset of students. Certainly, some students shared our view that peer assessment is not just a cheap way to generate a mark, commenting that "it makes us aware of the marking and what requirements there are." Only five students (7%) felt both that their peers were not qualified and that their talk was not improved by peer assessment, indicating that the vast majority of students were accepting of, and valued, peer assessment.

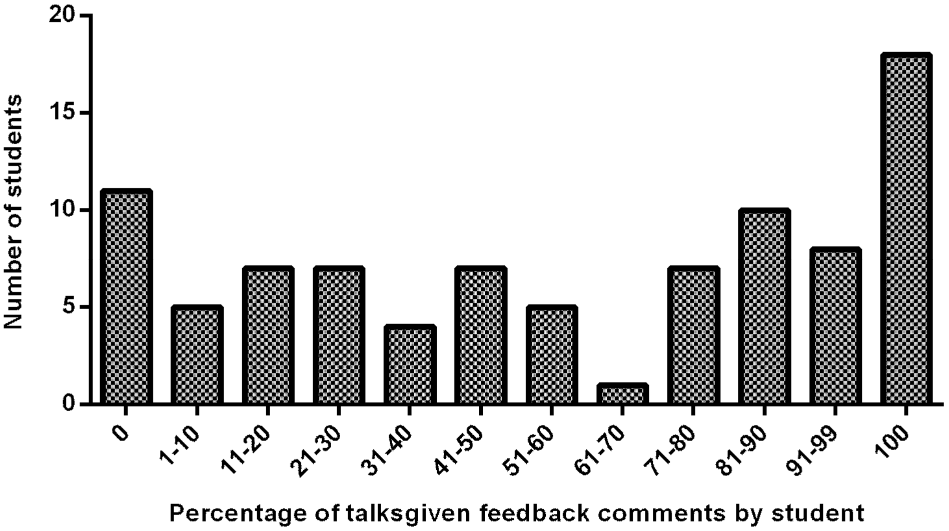

Figure 2. Bar graph representing the percentage of talks for which each student wrote feedback comments, grouped into ranges. This demonstrates the broad range in levels of engagement of the students with providing feedback.

Peer assessment is sometimes introduced to increase the amount of feedback that each student receives. We thought that students who are engaged in the peer assessment process and consider it to have value are likely to give more feedback comments to their peers. The amount of feedback that each student gave their peers varied widely. Twenty-six keen students gave feedback to more than 90% of the talks they assessed, but 16 students gave feedback to only 10% or less. On average, students gave feedback for 56% ± 38% (standard deviation) of the talks they assessed. The broad spread of student engagement with giving feedback can be seen in the large standard deviation of 38%, and also in the number of talks each student gave feedback (Figure 2). Many of the peer feedback comments were on presentation style (e.g., Don't read from notes, slides good or too crowded, voice good or too quiet). Only some students commented on content and explanation.

The peer marks correlate with the academic marks but with different overall distributions

It is important to examine whether including peer assessment resulted in a mark that accurately reflected the quality of the presentations, as defined by the mark given by the academic. Overall, the peer marks showed a much wider range, from 3.25 – 10 (mean 8.7 ± 1.3), in comparison to the marks from the academics, which were 5.5 – 10 (mean 7.8 ± 1.1) (Figure 3). Each individual student received a broad range of marks from their peers, shown by the standard deviations of peer marks received by each student, which varied from 0 to 1.79. The highest mark given to an individual student was 10 from all peer markers and 9.5 from the academic.

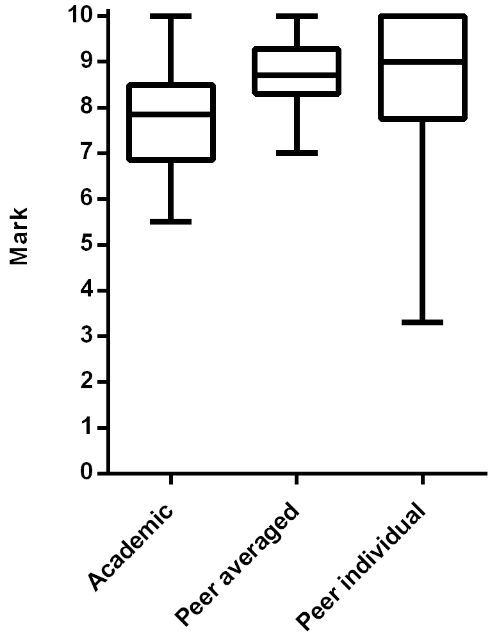

Figure 3. Comparison of the distributions of the academic marking, averaged peer marking, and individual peer marking. Plot shows mean, upper, and lower quartiles, and minimum and maximum. The academic marks were generally lower and show a broader distribution than the averaged peer marks, although the individual peer marks show a very broad distribution with slightly higher median.

As expected, when the peer marks given to each student were averaged, the distribution moved closer to that given by the academics. The averaged peer marks had a mean and median of 8.7 ± 1.0 and a range of 7.0 – 10 (Figure 3). This range is narrower than seen for the academic marks, which had a mean of 7.8 ± 1.1, median of 7.9, and a range of 5.5 – 10. This can be explained by the effect of regression to the mean when averaging many marks (Cheng & Warren, 1999).

In general, the averaged peer marks were higher than the academic marks, which is clearly demonstrated by the 75% quartile of the academics marks being only fractionally higher than the 25% quartile of the averaged peer marks (Figure 3). Indeed, 75 students (85%) were given a higher averaged peer mark than the mark given by the academic. Of the 1,462 peer marks, 389 (27%) were 10, compared to two of the academic marks (2.3%). These data suggest that most students rate peer presentations highly, or are simply loath to mark their peers low. This reticence was reflected in the students' comments about their attitude to marking their peers. Students expressed a strong desire to not disadvantage their fellow students, commenting that "I go very easy and don't want to mark below 'good' on these assessments as I don't want to be mean and hope that they would do the same for me." One insightful student commented that "there tends to be a moral dilemma when you know they are trying hard but they happen to just fall short." One student gave 10 to all presentations, and the broadest spread given by any student was 3.25 – 10 (mean 7.4 ± 2.2), indicating that students differ widely in the level of discrimination they choose to apply to the task of peer marking. There was a strong correlation between the averaged peer mark and the academic mark (r = .523, p < .001). However, the means were significantly different, t(87) = 9.229 p = .0001. From this data it seems that averaging the peer marks brings them closer to the academics marks, and that the averaged marks correlate highly despite more generous peer markers.

Despite contributing only 20%, the peer assessment component does change the overall mark

The final presentation marks (20% averaged peer marks, 80% academic mark) were compared with the academic marks. The final marks were 6.1 – 9.9, with a mean of 7.9 ± 0.96 and a median of 8.0. Despite the high percentage of the academic mark contributing to the final mark, the means of the academic marks and final marks were significantly different, t(87) = 9.234 p = .0001. Overall, 15 (17%) students received a different mark due to the introduction of peer assessment, and only three of these received a lower mark, from 8.4 to 8.3, 10 to 9.9, and 10 to 9.8. The reductions to the mark of 10 are also explainable by the phenomenon of regression to the mean. The greatest increase in marks was 0.6 for five students. It seems, then, that the final processing of the marks brings the final marks much closer to those given by the academic and does not significantly disadvantage any student. It is the potential for disadvantage that is of most concern to both students and academics in their acceptance of a marking method.

Are some students more accurate markers?

It has been reported that students capable of producing a high quality oral presentation may be better assessors of the quality of their peers' presentations (Jacobs, Briggs, & Whitney, 1975; Saavedra and Kwun, 1993). One of the students expressed a similar concern, stating that peer assessment "should only be for presentation style rather than content --> students may not have enough knowledge to make accurate assessment." We calculated the difference between each academic mark and the matching peer mark. We then examined these numbers to see which students marked the closest or furthest from the academic, as an approximation of accuracy. There was no correlation between the size of the academic-peer mark difference and the mark that the peer marking student received from the academic (r = -.37, r^2 = .137, p = .733). This demonstrates that, in this study, there was no difference between the ability of high achieving students and lower achieving students to assess the quality of their peer's talk. Interestingly, there was a small but significant correlation between the mark that each student received from each peer, and the mark that each student gave each peer (r = .323, r^2 = .1, p = .002). An r^2 of .1 indicates a 10% concordance between the mark each student received and the mark each student gave. This suggests that approximately 10% of the peer marks may be influenced by mutual friendship or dislike. This size of effect is largely hidden after averaging the peer marks. This is a very low level of apparent friendship/dislike bias contrasting with several students' perception of peer assessment, as they commented that "also marker[s] are always inclined to mark friends higher or with completely full marks" and "I think that only a very low proportion of marks should be allocated to the peer assessment, as peer review tends to incorporate a lot of bias." We found no correlation of marks with the order in which the oral presentations fell within a session, as had previously been observed (Langan et al., 2005). There was also no discernible difference in the marks given by males and females to same or opposite gender, in keeping with the previous studies (Falchikov & Magin, 1997).

The difference in marks is caused by reticence of peers to give low marks

The strongest difference between the academic and peer marks is at the lower ends of the ranges (Figure 3). We took the difference between each academic mark and each averaged peer mark, and looked for correlations with the region in the marks range. The differences in the marks showed a significant negative correlation with the academic marks (r = -.766, r^2 = 0.59, p < .0001) (Figure 4), indicating that as the academic mark decreases, the difference between the academic mark and the averaged peer mark increases. Indeed, an r^2 of .59 indicates that 59% of the difference between the marks can be accounted for by the level of the mark overall. This was even more extreme if we grouped the students into eight roughly equal size groups based on academic marks. The academic marks for each group were averaged, as were the differences between the academic mark and peer mark. The correlation between these was r = -.959, p < .0001, showing that 92% of the differences between averaged peer mark and academic mark relate to the grade given by the academic. In fact, although the overall means of the academic marks and averaged peer marks were significantly different, the means of the top group of students (grade 9 - 10, n = 13) were not significantly different from the academic mark mean, t(12) = 0.512, p = .612. The differences between the means becomes significant as the grades go lower, with the lowest bracket (5 – 7, n = 22) showing t(11) = 15.534, p < .0001. This strongly supports that the students mark higher at the lower ranges of the marks, but give equivalent marks at the top ranges.

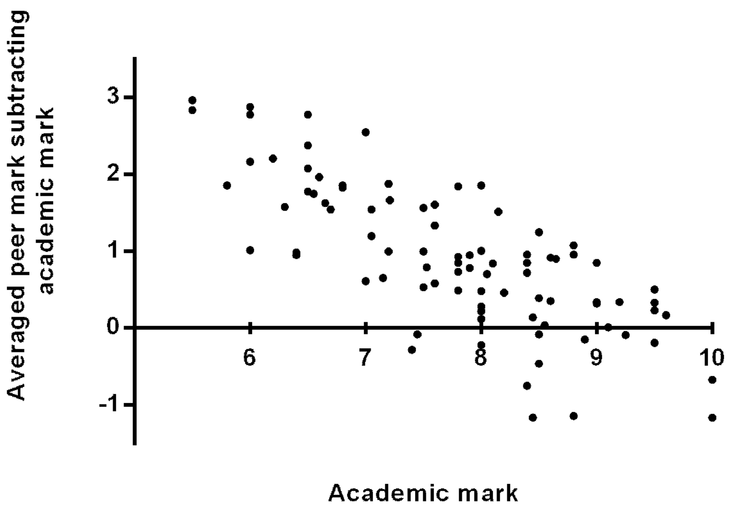

Figure 4. Dot plot comparing the difference between the averaged peer mark and the academic mark given to each student with the academic mark given to each student. As the academic mark increases, the difference becomes smaller. The small negative difference at the top marks is reflective of regression towards the mean caused by averaging multiple marks.

Modulating the rubric to explore the differences in marks

To further examine the difference in peer and academic marks, we determined the effect of modulating the peer marks using a common scaling formula. For any given spread of marks, accurate and fair scaling (altering of the range of marks) can be achieved by applying a multiplier (F) to the distance between each mark and the possible full mark (in this case 10). If F is greater than 1, the marks are scaled down, with the lowest marks moving further, and the higher marks moving proportionally less. Conversely, if F is less than 1, the marks will scale up, with the higher marks moving less than the lowest marks. The formula is (new score = 10 - F*[10-original score]). We trialled F values from 1.05 to 1.7. At F = 1.7 the mean of the peer marks was 7.76 ± 1.21, compared to the mean of the original academic marks, 7.75 ± 1.09 (Table 3). The spread of these scaled marks is much more similar to the academic marks, with similar standard deviations, and these distributions are now highly similar, t(87) = 0.079, *p* = .937. Using these scaled marks as the 20% peer mark with the 80% academic mark generates final marks that are also highly similar to the original academic marks, t(87) = -0.225, *p* = .823. The correlation between the academic marks and the final mark generated using the F = 1.7 scaling is *r* = .979, *p* < .0001, indicating a 96% concordance in the sets of marks.

Table 3
The effect of scaling on the similarity of the distributions of the peer marks and the academic marks (Paired t-tests)

Rubric	Mean	Median	SD	Min	Max	25%	50%	75%
Academic mark	7.75	7.85	1.09	5.50	10	6.85	7.85	8.50
Peer average	8.68	8.66	0.71	7.02	10	8.25	8.66	9.24
Peer average scaled with F = 1.7	7.76	7.72	1.21	4.93	10	7.03	7.72	8.71

Note. SD = standard deviation.

DISCUSSION

Many students reported a change in their behaviour due to the peer assessment process

In previous years, the students filled in a fact sheet based on the talks for a few marks. Although this was designed to increase student engagement, the questions generally required only extraction of simple facts, and the interest of the students was not well sustained. We consider that peer assessment provided much greater student engagement. The majority of the students reported that they changed the way they watched other students' talks and felt a responsibility to do so. Relative to their experience in previous years without peer marking, the academics reported observing increased engagement of most of the students with each other's talks. This has also been observed in other studies (Ballantyne et al., 2002; Orsmond, 2011; Topping, Smith, Swanson, & Elliot, 2000). Interestingly, the students who reported a feeling of responsibility did not necessarily report that they changed their own presentation as a result, suggesting that these two outcomes are not necessarily directly linked.

Another of the major learning outcomes we wanted to achieve from the peer assessment process was that students' would focus more attention on marking criteria, and thus improve this and future presentations. The students were mixed in their opinion about whether they did this, and many were undecided on whether peer assessment motivated them to improve their talk. At this university, it is a requirement that all students receive the marking criteria along with their assessment task instructions, and yet our observations are that the criteria are not well applied by all students. As peer assessment has become a common occurrence in undergraduate courses at this university, we feel confident that repeated exposure to the process will cause students to reflect further on the criteria by which they are marked and by which they mark their peers, and that this will improve their assessments as a result. It is broadly accepted that positive learning outcomes of peer assessment increase with multiple exposure to peer assessment (Orsmond, 2011).

The students are not biased, but are reticent to mark their peers low

Some students were concerned that their peers may be biased due to friendship groups. We found only scant evidence of a bias in the marks, possibly due to mutual friendship or dislike, and so we feel that this is not sufficient to warrant concern as, consistent with other studies (Magin, 2001), the effect was negligible. In addition, although some peers were very harsh on individual presentations, the averaged peer marks generally produced an equal or higher mark to the academic mark, and so the concerns of the students that their peers are hard markers were unfounded. For the most part, the students felt that their peers are qualified to assess their presentations and were generally comfortable with peer assessment as long as the proportion of marks was not too high.

A common concern of the students is that their peers may judge them unreasonably harshly. In this study, the peer marks were found to be most similar to the academic marks at the higher grades, and significantly and consistently over-marked at the lower grades, suggesting that students are very reticent to mark peers low. Indeed, applying a substantial level of scaling (using a factor of 1.7) to the averaged peer marks generated a distribution of marks with a very high correlation, and very similar mean, to the academic marks, indicating that students are very capable of discriminating between high and low quality presentations of fellow students, but are unwilling to use the lower end of the scale. This is consistent with previous studies (Brindley & Scoffield, 1998; Falchikov, 1995).

Application of the rubric

The students showed a strong reticence to mark low, one student noting that, "a peer would very rarely fail another peer's work," and yet we were surprised to see individual peer marks as low as 3.25 and 4, far below the academic marks of 7.2 and 5.8. We take this as evidence that some students were not considering how the rubric weightings multiply to calculate the peer mark. In future assessments we could ask students how much they consider the marking rubric in the preparation of their talk, and whether the peer assessment has altered their intention to consider the rubrics for future assessment tasks. The marking academics did observe that introducing a more explicit marking rubric (as opposed to a general marking scheme) for this year increased the amount of experimental content from the scientific literature presented by the students, although they were not always discerning in their choice of which experiments to include and explain. Interestingly, the students' feedback comments to each other often focused on the more stylistic elements of the presentation rather than scientific content. These results are in keeping with other studies, which indicate that peers tend to mark more similarly to academics if the criterion is overall performance rather than specific criteria (Campbell et al., 2001; Falchikov & Goldfinch, 2000; MacAlpine, 1999). In future, the scientific nature of the rubric needs to be emphasised with the students to alleviate this. We did not see any correlation between the quality of the students' own work and the similarity of their marking compared to the academic marking, indicating that both high and low achieving students are equally capable of applying the rubric and judging their peers' performance.

The final mark is sufficiently unperturbed to cause little negative impact on the overall assessment

The students gave highly variable marks, but the validity of the marks (as compared to the academic marks) improved when averaged, in keeping with observations from other studies (Boud & Falchikov, 2007; Cheng & Warren, 1999; Falchikov & Goldfinch, 2000; Magin & Helmore, 2001; Topping, 1998). To determine the effect of the peer assessment process on the students' marks, we compared the final marks to the marks given by the academics and determined that they are significantly different, although they do correlate strongly. The mean correlation between peer marks and academic marks found in a meta-analysis by Falchikov and Goldfinch (2000) was .69, and our study shows a very similar result. Many of the students indicated that they would be willing to accept a higher percentage of peer mark than the 20% used in this study. In order to maintain this minimal effect of the peer marks on the final mark, we will not alter this percentage in future years.

After rounding, only 17% of the students have different marks, and only 3 (3%) show a lower mark. Our results indicate that it is sufficient to apply a scaling of 1.7 to the averaged peer marks in order to align them closely with the academic mark, which provides a possible method for mitigating the effect of peer over-marking. In order to maintain transparency of marking we have chosen not to use this approach.

Throughout this analysis we have set the academic mark as the standard for accuracy, but it can be argued that an academic mark is not necessarily a reliable or consistent measure of the quality of a talk, as academic bias or variability is also possible (Falchikov & Magin, 1997; Smith et al.,, 2002). Therefore, against this current standard practice of using possibly variable academic marks, we consider that the observed alteration to the final marks by using peer assessment has done no substantial damage to the mark outcomes of the students, and has introduced sufficient positive learning outcomes to warrant continued use in this format, an opinion shared by others (Boud & Falchikov, 2007; Falchikov & Goldfinch, 2000; Orsmond, 2011). Many other studies find peer evaluations reliable and adequately valid (Sadler & Good, 2006), although some others find them too variable. Overall, we consider this experiment in peer assessment in a science setting to be successful as it appears to achieve the positive effects of increased motivation, engagement, and self-reflection, and only causes minor negative outcomes with broadly accurate, though a little more generous, marks in the majority of cases.

CONCLUSION

This study establishes that the students felt a responsibility because of the peer assessment and that they considered that they changed the way they watched each other's talks and reflected on their own. We consider that peer assessment successfully increased the students' attention to each other's talks and increased the motivation of many students in their talk preparation. Some students gain a deeper consideration of their own talk through reflection and may therefore have learned to consider the marking criteria for this and future talks, but these aims were not directly assessed in this study. We consider that this peer assessment task has satisfied the four major goals for its introduction. This study established that the students are capable of discriminating but are unwilling to give their peers low marks. Despite this, the final marks are a suitably accurate representation of the quality of the oral presentations, being minimally altered by the inclusion of 20% peer assessment. As a result of these findings, we will continue to use this peer assessment process in this unit/subject. Furthermore, we conclude that there was a sufficiently positive effect from the introduction of peer assessment to warrant its use in other science based courses.

ACKNOWLEDGEMENTS

This research was carried out with the approval of the Monash University Human Research Ethics Committee, Project number CF12/1179 – 2012000578

REFERENCES

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment & Evaluation in Higher Education, 27*(5), 427-441.

Boud, D., & Falchikov, N. (2007). *Rethinking assessment in higher education: Learning for the longer term.* London, England: Routledge.

Brindley, C., & Scoffield, S. (1998). Peer assessment in undergraduate programmes. *Teaching in Higher Education, 3*(1), 79-89.

Campbell, K. S., Mothersbaugh, D. L., Brammer, C., & Taylor, T. (2001). Peer versus self assessment of oral business presentation performance. *Business Communication Quarterly, 64*(3), 23-42.

Cheng, W., & Warren, M. (1999). Peer and teacher assessment of the oral and written tasks of a group project. *Assessment & Evaluation in Higher Education, 24*(3), 301-314.

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education*, *24(*3), 331-350.

Falchikov, N. (1995). Peer feedback marking: Developing peer assessment. *Innovation in education and training*, *32*(2), 175-187.

Falchikov, N. (2005). *Improving assessment through student involvement: Practical solutions for aiding learning in higher and further education.* London, England: RoutledgeFalmer.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, *70*(3), 287-322.

Falchikov, N., & Magin, D. (1997). Detecting gender bias in peer marking of students' group process work. *Assessment & Evaluation in Higher Education, 22*(4), 385-396.

Goldschmid, B., & Goldschmid, M. L. (1976). Peer teaching in higher education: A review. *Higher Education, 5*, 9-33.

Higher Education Council (Australia). (1992). *Achieving quality: Higher education*. Canberra, Australia: Australian Government Publishing Service.

Jacobs, R. M., Briggs, D. H., & Whitney, D. R. (1975). Continuous-progress education: III. Student self-evaluation and peer evaluation. *Journal of Dental Education*, *39*(8), 535-541.

Langan, M. A., Wheater, C. P., Shaw, E. M., Haines BJ, Cullen WR, Boyle JC, … Preziosi, R. F. (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.

MacAlpine, J. M. K. (1999). Improving and encouraging peer assessment of student presentations. *Assessment & Evaluation in Higher Education*, *24*(1), 15-25.

MacPherson, K. (1999). The development of critical thinking skills in undergraduate supervisory management units: Efficacy of student peer assessment. *Assessment & Evaluation in Higher Education*, *24*(3), 273-284.

Magin, D. (2001). Reciprocity as a source of bias in multiple peer assessment of group work. *Studies in Higher Education*, *26*(1), 53-63.

Magin, D., & Helmore, P. (2001). Peer and teacher assessments of oral presentation skills: How reliable are they? *Studies in Higher Education*, *26*(3), 287-298.

Napan, K., & Mamula-Stojnic, L. (2005). *A process that empowers: Self and peer assessment as a component of education for sustainability.* Paper presented at the Making a Difference: Evaluations and Assessment Conference, November 30–December 1, 2005, Sydney, Australia.

Orsmond, P. (2011). *Self- and peer-assessment: Guidance in practice in the biosciences*. Leeds, England: UK Centre for Bioscience, The Higher Education Academy.

Patton. C. (2012). "Some kind of weird, evil experiment": Student perceptions of peer assessment. *Assessment & Evaluation in Higher Education*, *37*(6), 719-731.

Saavedra, R., & Kwun, S. K. (1993). Peer evaluation in self-managing work groups. *Journal of Applied Psychology, 78*, 450-462.

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

Smith, H., Cooper, A., & Lancaster, L. (2002). Improving the quality of undergraduate peer assessment: A case for student and staff development. *Innovation in Education and Training*, *39*(1), 71-81.

Søndergaard, H. (2009). Learning from and with peers: The different roles of student peer reviewing. In *ITiCSE* (pp. 31-15). Paris, France: ACM.

Stanier, L. (1997). Peer assessment and group work as vehicles for student empowerment: A module evaluation. *Journal of Geography in Higher Education*, *21*(1), 95-98.

Topping, K. (1998). Peer assessment between students in colleges and universities. Review of Educational Research 68(3):249-276.

Topping, K, Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between post-graduate students. *Assessment & Evaluation in Higher Education*, *25*(2), 149-166.

Vickerman, P. (2009). Student perspectives on formative peer assessment: An attempt to deepen learning? *Assessment & Evaluation in Higher Education*, *32*(2), 221-230.

APPENDIX

Figure A1. Questionnaire

1. In your university course have you ever been asked to peer assess other student's work before? Options: 3+ times, 1-2 times, 0 times, not sure
2. In general, what is the maximum proportion of your total mark for an assessment such as this that you are comfortable to receive from your peers? Options: 0%, 20%, 40%, 60%, 80%, 100%

The next three questions relate to this assessment. Please select the most correct response (Options in Likert scale, Strongly agree, agree, neutral, disagree, strongly disagree, not sure)

3. I was aware that I was going to assess other students' talks, and other students were going to assess my talk.
4. My peers are well qualified to provide an assessment of my talk.
5. The way that I watch other students' talks changes because I am assessing the talk.
6. I feel a responsibility because I am assessing other students' talks.
7. The way I prepare my talk changes because I am being assessed by other students.
8. The way I present my talk changes because I am being assessed by other students.
9. The peer assessment process caused me to reflect more on my own talk.
10. The peer assessment process has helped me to improve my own talk.
11. Please write any further comments here or over the page:

Table A1
Full rubric with grade descriptors

	Excellent (80-100)	**Good** (65-79)	**Adequate** (50-64)	**Poor** (0-49)
Quality of presentation (10%)	Clear slide layout Excellent use of images. Minimal text. Clearly readable. Speaks with good pacing. Makes eye contact and does not read information. Uses engaging tone.	Good images but not always well placed. Size and labels are clear. Little text. Speaks well, but often backtracks. Makes good eye contact and looks at notes or screen occasionally.	Labels, text, and legends are a bit unclear or too small. Too much detail. Blocks of text on slides. Some hesitation and uncertainty are apparent. Makes little eye contact. Monotone and non-engaging delivery.	Slides are cluttered. Labeling is not clear. Too small to see. Mostly text and very few images. Makes no eye contact and reads from notes. Hesitation and uncertainty are apparent.
Identification of key experimental evidence (10%)	The primary literature has been critically evaluated and the experimental evidence crucial for our understanding of the topic has been identified.	Clear selection of the most important evidence from analysis of the primary literature. Some inclusion of non-essential information or absence of a small amount of critical evidence.	A mixture of essential and non-essential information provided with limited prioritisation of significance. Limited use of primary literature.	Very limited, or irrelevant, information presented from inappropriate sources.
Description of experiments 40%)	The results are clearly presented with sufficient detail to allow the audience to critically analyse the experiments.	The results are mostly presented clearly but a few details required for the analysis of the experiments are lacking or vaguely described.	Some results presented clearly but others are vaguely or imprecisely described.	Results rarely presented with sufficient detail for their analysis.
Explanation of conclusions (30%)	The conclusions are well explained and are completely justified as a clear and logical interpretation of the available evidence	Conclusions are appropriate and mostly well explained but in a few cases it is not clear how the evidence supports the conclusion or not all of the implications have been considered.	Some accurate conclusions are presented but are not fully justified by the evidence presented or some obvious implications have not been considered.	No conclusions, or incorrect conclusions, presented.
Ability to answer questions (10%)	Understands audience questions. Can integrate knowledge to answer questions. Thoroughly responds to questions.	Understands the audience questions. Can integrate knowledge to answer the question. Thoroughly responds to most questions.	Makes an effort to address question. Can address some questions. Responds poorly to some questions	Responds poorly to questions or makes no effort. Overlooks obvious answers