Implementation of marker training exercises to improve marking reliability and consistency

Gavin Buskes; Huey Yee Chan.
The University of Melbourne
Corresponding Author Email: g.buskes@unimelb.edu.au

STRUCTURED ABSTRACT

CONTEXT
One of the challenges present in teaching a large engineering subject is that of achieving marking consistency of assessments across multiple markers. Several measures of standardising markers exist, such as calibrated review, and are commonly used in the humanities, particularly for assessments that could be prone to a wide variation in marks such as essays. The application of such methods, in an engineering context, is somewhat less documented but of particular importance in the case of reflective writing. This study contrasts the implementation of several different methods of using marker training exercises prior to the actual assessment marking and provides an analysis of the results in order to minimise the effect of multiple marker irregularities and to provide effective high-quality formative feedback on a piece of reflective writing

PURPOSE
This paper presents several different methods of marker training exercises, run prior to the actual assessment marking, and provides analyses to determine the effect of each in terms of minimising marking inconsistency among multiple markers on a piece of reflective writing.

APPROACH
In all three marker training exercises, markers are given samples of a piece of reflective writing, of differing quality, along with a rubric outlining the marking criteria for the piece of writing and exemplars for indicative marking standards. Each of the methods employed differ in how the reference standard was set and how feedback was delivered to the markers. Statistics comparing the marking results across markers from before the introduction of the training exercises and between each of the three training methods were analysed to investigate marking reliability and consistency.

RESULTS
A significant reduction in the spread of the marker means has been achieved through the introduction of the marker training, indicating an improvement in consistency. Some differences in results between the alternative methods employed has also been observed.

CONCLUSIONS
Marking consistency can be improved with the introduction of a marker training exercise prior to the actual assessment marking. Different methods of implementing the marking training exercise, and how the feedback is provided, can have an effect of the amount of improvement in terms of consistency and reliability.

KEYWORDS
Marking consistency, marking reliability, multiple markers, reflective writing.

Introduction

One of the challenges present in teaching a large engineering subject is that of achieving consistency of marking assessments across multiple markers, particularly for a piece of written assessment. Many studies, including Blok (1985), Engelhard (1994) and Hughes and Keeling (1984), discuss the subjectiveness in marking written pieces of assessment, which are thereby more prone to marking consistency issues in a subject with multiple markers. For this reason, marking scales or rubrics are used in order to apply a standard of consistency that would be absent due to markers' subjective judgments and marking according to their "general impression" of a piece of writing (Kayapinar, 2014).

Even with a rubric and marking scale, there is still no guarantee that markers will share a common view of the value of a given mark for a piece of assessment. This is particularly relevant to novice markers, most of whom are capable of making gross judgments about a piece of writing and placing it into broad categories from good to bad, but assessing based on a rubric with a breakdown of multiple criteria could be a new and complex task for them. Experienced markers may have developed bad habits that need to be addressed. Therefore, the need exists to train markers, regardless of their experience level, to learn to make an informed judgement on a piece of writing that is of a common standard within all the criteria of a rubric. Price (2005) argues that the key to consistency in marking is the application of a set of assessment criteria and standards which has been agreed upon by markers, that is, markers develop their understanding of the assessment criteria by discussing marking with their peers.

Putting together identified best practices, Bird and Yucel (2013) outlined a comprehensive program aimed to improve consistency and efficiency of marking a first-year science report by a large team of laboratory demonstrators. The program practices included elements such as an assessment rubric, annotated exemplars, feedback code and sample comments, complemented with a discussion among markers to exchange knowledge of standards and decide on a common approach before assessments are marked.

The intention of this study is to document several methods of using marker training exercises similar to that of Bird and Yucel's (2013), implemented prior to the actual assessment, and provide an analysis on their efficacy in minimising the effect of multiple marker irregularities for a piece of reflective writing in a large first-year engineering subject. While the methods in this study were not as comprehensive as that of Bird and Yucel's (2013), most of the elements were taken into consideration, such as the rubric, exemplars, feedback code and sample comments. A different agreement on marking standards among markers approach, however, was taken. The group discussion among markers element was not practiced; instead, various methods of one-way feedback directly to markers were explored.

Purpose

This study investigates marking consistency by means of statistical analysis on actual marks of a piece of reflective writing over the duration of four years, from 2015 to 2018, by looking at:

1. the spread of marks across independent marker groups.
2. evidence of statistically significant difference in marks between independent marker groups and the population as well as significant differences within marker groups.

Approach

In order to investigate the effect on marker consistency of introducing a marker training exercise prior to actual assessment marking, the research was conducted in a first-year engineering subject at The University of Melbourne during semester 1, over a four-year period. The typical subject enrolment is more than 600 students per semester, which are divided into workshop class groups with a maximum of 60 students. Individual tutors mark assessment tasks for students in their allocated workshop class. The assessment of interest in this study is a piece of individual reflective writing submitted during the middle of the semester.

Two weeks prior to the actual assessment marking, all markers undergo the marker training exercise. Three different methods of marker training exercise were employed over three years and results were compared with results from the year immediately prior to the introduction of the training exercise.

In all three of the marker training exercises, markers were given the same four samples of a piece of reflective writing, of different qualities. Markers graded the papers against 9 specified criteria and were given 2 exemplars that were marked by the experienced coordinator for indicative marking standards.

Method 1 – Feedback on overall standard of marking based on coordinator benchmark

In 2016, the 14 markers in the subject carried out the trial marking exercise through an online platform. Markers rated 4 samples of the reflective writing paper based on a Likert scale of 0 to 4 for 9 specified criteria, which totalled 36 marks for each paper. Before markers rated the 4 samples, they were provided with two exemplars marked by the coordinator as reference. Upon completion of assessing the 4 samples, individual marker's ratings on each paper were assessed by an automated online grading system based on the coordinator's benchmark standards. Immediate feedback on the standard of marking and an overall recommended rating on each paper (out of 36 marks) were provided.

Method 2 – Feedback on 9 criteria of marking based on a derived average benchmark

In 2017, the 15 markers in the subject carried out the trial marking exercise using a spreadsheet for mark entry. For each of the 4 samples of reflective writing paper, markers recorded their mark, minimum of 0 to maximum of 4, for each of the 9 criteria along with a descriptive statement to justify the mark awarded into a spreadsheet. The spreadsheet automatically summed the marks of the 9 criteria bringing a total out of 36 marks for each paper. Before markers rated the 4 samples, they were provided with two exemplars marked by the coordinator as a reference similar to that of Method 1. At the end of the assessment period, rating information from all markers was compiled and a reference standard of marking was derived from the average of all markers. Charts comparing how tutors fared amongst each other and feedback on the individual 9 criteria that were considered extreme ratings were provided as feedback to individual markers.

Method 3 – Feedback on 9 criteria of marking based on a combination of derived average and coordinator benchmark

In 2018, 14 markers participated in the trial marking exercise that was a combination of Method 1 and Method 2. Similar to Method 2, markers recorded their marks and comments on 4 sample papers into a spreadsheet. This time, however, the coordinator also entered their assessment into a spreadsheet. Feedback in the form of comparison charts and extreme marker ratings derived from the average of all markers was provided at the end of the assessment period. In addition to that, the coordinator's spreadsheet as the benchmark marking standard was also provided. The key difference between the coordinator benchmark this time and that of Method 1 was the thorough feedback on the 9 individual criteria for the 4 sample papers being supplied instead of an overall paper mark out of 36.

For all of the marker training methods, there was no process to follow up on how the markers chose to review the feedback provided.

Data analysis

Descriptive and inferential statistical analyses were carried out on the actual marks of the piece of reflective writing assessment from independent markers across 4 years, from 2015 to 2018, as a holistic perspective to investigate the effect of marker training exercise on marking consistency. All data was analysed using IBM SPSS 24 software.

For this study, the actual marks of the reflective writing paper from the whole cohort for each of the 4 years were split according to the respective marker who evaluated the papers. Each marker was set up as a marker group within their respective year. The number of marker groups ranged from 1 to 15, based on the number of markers in the particular year. The number of samples in each marker group was not fixed; it depended on the number of papers marked by the particular marker. The sample size ranged from a minimum of 23 to maximum of 86 papers per marker group, shown in Table 1.

Table 1: Summary of marker groups from 2015 to 2018

	2015	2016	2017	2018
Marker training exercise	Not carried out	Method 1	Method 2	Method 3
Number of marker groups	15	14	15	14
Minimum number of samples in marker groups	23	25	25	25
Maximum number of samples in marker groups	86	68	52	55

Observations of the spread of medians across marker groups and differences between marker groups were used as indications of marker consistency in a holistic manner. It is prudent to note that the analysis methods used to provide comparison from a holistic and relative point of view does not take into account how the actual differences in student quality might affect marker group marks. Student academic quality was assumed to be fairly normally distributed over a large sample size of 600 students in a cohort (Field, 2009). With large enough sample sizes for each marker group, the quality of students within each group was assumed to be fairly evenly distributed as well, and that this factor would not affect the relative differences between marker groups.

In order to select the appropriate statistical analysis methods for the data, the overall reflective writing paper marks for each of the 4 years were tested for normality. The Shapiro-Wilk test identified that the overall marks did not follow a normal distribution to a significance level of 0.05. The evidence of non-normality of the overall mark distribution, coupled with the small sample sizes of the marker groups led to the decision to select non-parametric tests and use of medians for further evaluation of marking consistency. Non-parametric techniques should show superior power to parametric techniques when they are compared to a markedly non-normal distribution (Rasmussen and Dunlap, 1991) and non-parametric tests are most useful for small studies (Fagerland, 2012).

Box plots were used as a measure of dispersion to compare the spread of the medians from all marker groups across 4 years, and are shown in Figure 1.

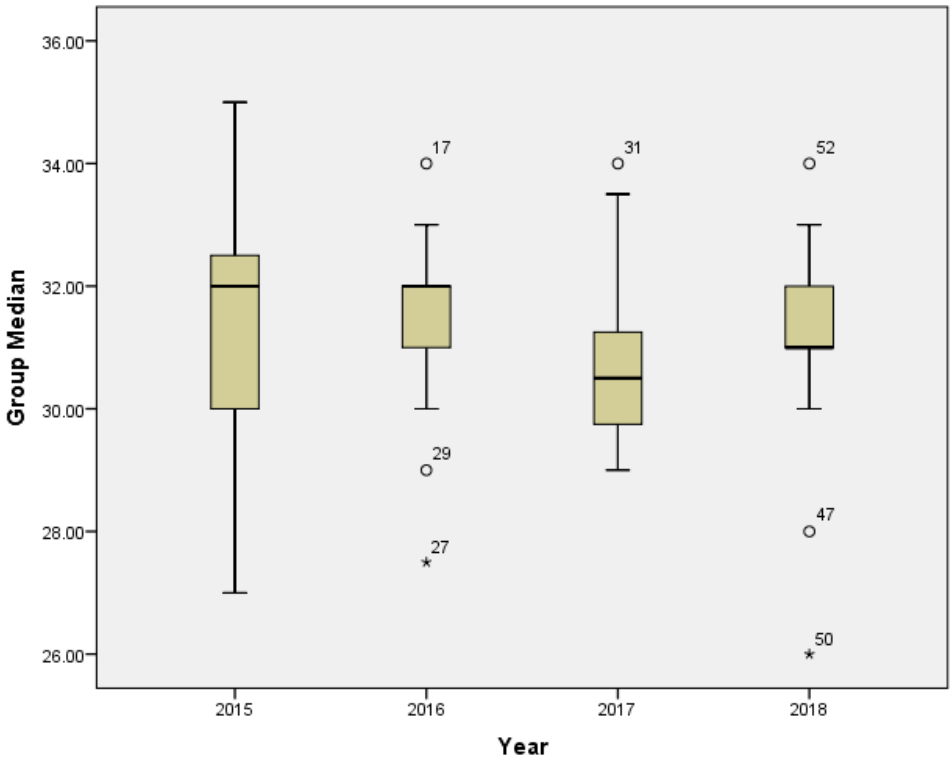

Figure 1: Box plot of marker group medians

The plots revealed a larger spread of medians among marker groups in 2015 as compared to the other 3 years. Interquartile range based on the 25% to 75% percentiles was highest at 3.00 for 2015, the year before marker training exercise was introduced. There were observed differences in marker medians over the 4 years, but a Kruskal-Wallis test confirmed the differences were not significant at the 95% confidence interval. Due to the reduction in spread, outliers were observed in the years with marker training.

The non-parametric Mann-Whitney U test was carried out to identify all independent marker groups that are statistically significantly different from the population for each year and shown in Table 2. The consolidated pairwise comparisons from the Mann-Whitney U test, with a 95% confidence interval for significance, provided evidence that the implementation of marker training resulted in the reduction in the proportion of marker groups that significantly differed from the population. The 60% of marker groups with significant difference from the population in 2015 was reduced to approximately half in subsequent years when the marker training was implemented.

Table 2: Mann-Whitney U test to identify differences between marker groups and population

	2015	**2016**	**2017**	**2018**
Marker training exercise	Not carried out	Method 1	Method 2	Method 3
Number of marker groups	15	14	15	14
Number of groups significantly different from the population at significance level of 0.05	9	4	5	5
Percentage of groups with significant difference	60%	28.6%	33%	35.7%

The Kruskal-Wallis test provided strong evidence ($p << 0.05$) of statistically significant differences between at least two independent marker groups for all the 4 years. Further post-hoc analysis using pairwise comparisons from the Dunn-Bonferroni tests, shown in Table 3, showed that the number of group-pairs that were significantly different had been reduced by more than half from 28.6% significantly different pairs in 2015 to a worse case of 12% in 2018 with the implementation of marker training exercise.

Table 3: Post-hoc analysis on marker group pairs with significant difference

	2015	**2016**	**2017**	**2018**
Marker training exercise	Not carried out	Method 1	Method 2	Method 3
Number of marker group-pairs	105	91	105	91
Number of marker group-pairs that are significantly different at significance level of 0.05	30	9	8	11
Percentage of group-pairs with significant difference	28.6%	9.9%	7.6%	12%

Although results indicated an improvement in consistency with the implementation of marker training exercises, it was inclusive of the results across all markers. It could be reasonably expected that novice markers who were new to the subject in a particular year, and thus marking the papers for the first time, would be contributing a higher number of group-pair differences. In this light, Table 4 consolidates the data from Table 2, and shows the Mann-Whitney U test data limited to only first-time novice markers. In 2015, all markers were new to marking the assessment as the reflective writing was introduced for the first time in the subject. From 2016 to 2018, markers were made up of a

combination of experienced markers who have marked the same assessment in previous years and novice markers who were marking the assessment for the first time. No significant improvement was observed in the percentage of novice marker groups differing from the population for 2015 and 2016 when Method 1 was introduced. However, a consistent reduction in the number of novice marker groups different from the population, from 60% in 2016 to 0% in 2018, suggested that enhancements to the marker feedback technique for Method 2 in 2017 and Method 3 in 2018 has made a bigger difference to the consistency of novice markers as compared to Method 1 in 2016. Looking from a different perspective, this implied a decline in the effectiveness of marker training exercise for experienced markers, since the total number of differences remained somewhat constant over the 3 years.

Table 4: Mann-Whitney U test to identify differences between new marker groups and population

	2015	2016	2017	2018
Marker training exercise	Not carried out	Method 1	Method 2	Method 3
Number of novice marker groups	15	5	9	6
Number of novice groups significantly different from the population at significance level of 0.05	9	3	1	0
Percentage of groups with significant difference	60%	60%	11%	0%

Discussion

Observations from the comparison of marker medians and non-parametric tests strongly suggested significant improvement in marker consistency with the implementation of marker training exercises in the subject. The identification of outliers in the years with marker training could serve as information to the coordinator in identifying marks from marker groups that might need moderation or further review, though it was not carried out in the subject. Improvement in marking consistency was also apparent from the significant reduction in the number of relative differences between markers, by approximately half, from the non-parametric Mann-Whitney U and Kruskal-Wallis tests. However, there is likelihood that the marked improvement from 2015 to 2016 could have been attributed to the fact that all markers were new to marking the assessment in 2015 and hence making it difficult to ascertain the efficacy of the marker training exercise when implemented for the first time in 2016.

In order to investigate the tangible impact of the different marker training methods, novice marker groups were isolated for analysis. It was interesting to note that marker training Methods 2 and 3 were more effective in providing a benchmark marking standard for novice markers as compared to Method 1, which did not make a significant difference when compared to the year without marker training. One key difference between Method 1 and the other two lies in the detail of feedback provided after the marker training exercise. Marking of the reflective writing was based on 9 specific criteria but the feedback provided to markers in Method 1 was based on the overall paper mark (out of 36) rather than the 9 individual criteria, hence there might have been insufficient breakdown within the marking standards for meaningful feedback.

Another factor that could have contributed to the effectiveness of the marker training exercise was the format of the feedback provided to markers. In Method 1, feedback presentation was simple - a few lines of text informing the marker whether they marked above or below the expected standard and a recommended range of marks that were appropriate. In Methods 2 and 3, more thorough feedback in the form of comparison charts and highlighted sections of extreme ratings along with recommendations of expected ratings were sent to individual markers in a spreadsheet via email. The spreadsheet also contained anonymised ratings and comments from all markers and the experienced coordinator. Markers were able to review ratings and comments from other markers for reference. A suggested improvement strategy would be to integrate Methods 2 and 3 into an online platform to take advantage of technology to provide practice and feedback in a timely and efficient manner (Dempsey, PylikZillig, & Bruning, 2009).

An interesting observation from the non-parametric Mann-Whitney U tests was that the improvement in consistency between novice markers was contrasted by some increasing differences with experienced markers. Does this imply marker training was not as effective for experienced markers? In the authors' opinions, this could be attributed to how markers reviewed the provided feedback. It was questionable if markers were thorough enough in scrutinising and absorbing the details in the feedback. There is a possibility that novice markers who were not as confident paid more attention to the feedback compared to experienced markers, who perhaps felt they could draw on past experience. Since there is no evidence to support this assertion, this presents an opportunity in the future to review the current process and consider a more thorough marker training exercise to include tutor surveys, meetings to discuss marker ratings and perhaps a process of subsequently collaboratively re-grading the papers until all markers come to a common agreement, a practice presented by Willey & Gardner (2010) that was found to be effective in improving marking consistency.

Conclusion

The marker training exercises were shown to have a positive impact in improving marking consistency in assessing a piece of reflective writing in a large first-year engineering subject. The amount of detail coupled with the different methods of how feedback was delivered to markers had an effect on the efficacy of the training exercise. Overall, the authors of this paper were pleased with the improvements achieved though the evolution of the marker training exercises over the 4 years. Not surprisingly, the outcome of the study revealed a number of inadequacies within the marking methods that could be further addressed in future implementations. One recommendation would be to implement measures to ensure a more comprehensive feedback review process among markers as a follow up after the completion of the training exercise.

References

Bird, F. L., & Yucel, R. (2013). Improving marking reliability of scientific writing with the Developing Understanding of Asessment for Learning programme. *Assessment & Evaluation in Higher Education, 38:8*, 536 - 553.

Blok, H. (1985). Estimating the reliability, validity and invalidity of essay ratings. *Journal of Educationsl Measurement 22(1)*, 41-52.

Dempsey, M., PylikZillig, L., & Bruning, R. (2009). Helping preservice teachers learn to assess writing: Practice and feedback in a web-based environment. *Assessing Writing, 14*, 38-61.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted rasch model. *Journal of Eductional Measurement 21(3)*, 93-112.

Fagerland, M. W. (2012). t-tests, non-parametric tests, and large studies—a paradox of statistical practice? *BMC Medical Research Methodology, 12, 78.* Retrieved from http://doi.org/10.1186/1471-2288-12-78

Field, A. (2009). Discovering statistics using SPSS. 3 ed. In A. Field, *Discovering statistics using SPSS. 3 ed.* (p. 822). London : SAGE Publications Ltd.

Hughes, D., & Keeling, B. (1984). The use of model essays to reduce context effects in essay scoring. *Journal of Educational Measurement 21(3)*, 277-281.

Kayapinar, U. (2014). Measuring essay assessment:Inrea-rater and inter-rater reliability. . *Eurasian Journal of Educational Research, 57*, 113-136.

Price, M., Rust, C., & O'Donavan, B. (2005). A social constructivist assessment process model: how the research literature shows us this could be the best practice. *Assessment & Evaluation in Higher Education 30(3)*, 231-240.

Rasmussen, J. L., & Dunlap, W. P. (1991). Dealing with nonnormal data: Parametric analysis of transformed data vs nonparametric analysis. *Educational and psychological measurement [0013-1644] vol 51 issue 4*, 809 -820.

Willey, K., & Gardner, A. (2010). Perceived Differences in Tutor Grading in Large Classes: Fact or Fiction? *40th ASEE/IEEE Frontiers in Education Conference.* Washington DC.