Getting tutors on the same page

Keith Willey
University of Technology, Sydney, Sydney, Australia
Keith.Willey@uts.edu.au

Anne Gardner
University of Technology, Sydney, Sydney, Australia
Anne.Gardner@uts.edu.au

***Abstract**: In large engineering subjects, it is common to have multiple tutors where each tutor is responsible for grading the assessment tasks for students in their tutorial. An issue regularly faced by subject coordinators is how to achieve a consistent standard of marking and feedback quality amongst different tutors. To address this issue the authors initially used a number of methods including double blind marking to support consistent grading. However, with increasing demands on academics these time-consuming activities became an unrealistic option. This process was improved by using a software tool to compare both the marking and feedback provided by different tutors for a number of randomly selected project tasks. In this paper, we report using new software features developed as a result of this previous research to quickly establish and build a community of assessment practice amongst subject tutors. The reported process promotes inclusiveness by using a software tool to anonymously record and report tutor assessments allowing all opinions to be considered during a subsequent discussion activity. Even though this pilot exercise was undertaken by experienced tutors it significantly influenced their feedback skills and to a lesser extent their marking standards.*

Introduction

As a result of changes in the last two decades, Australian and UK universities have seen a reduction in staff–student ratios often resulting in large classes. Furthermore, research funds are often used to buy permanent academic staff out of teaching, resulting in an increasing number of less experienced casual or sessional teaching staff being used to conduct core teaching activities such as tutorials and marking of student work (Price 2005, White 2006).

Grading is often an activity that results in anxiety for both teachers and students. This is especially so for less experienced staff when holistic marking is used in part due to the difficulties in justifying grading decisions to students. This issue is further complicated in large classes by the fact that often a number of staff are used to mark the same activity for different students. Even experienced staff differ in their understanding of academic standards. The fact that increasing marking is being undertaken by less experienced sessional teachers and tutors only compounds this problem.

For consistent marking amongst tutors, it is important for all assessors to share a common view of the value of a given grade. Tomkinson & Freeman (2007) suggest that some form of induction, for example, a small number of 'yardstick' assessments be used as a basis for discussion about standards.

Several researchers including Price report that *"An assessment standards discourse is needed to support the functioning of assessment communities of practice…"* (Price 2005, p. 226). That is, tutors develop their understanding of the assessment criteria and language of feedback by discussing marking with other academics. This aligns with a social constructivist view of learning, that is, learning requires "*active engagement and participation*" (White 2006, p.237) this being true for both tutors and students.

In large engineering subjects, it is common to have many tutors where each tutor is responsible for marking the assessment tasks for students in their tutorial. An issue regularly faced by subject coordinators is how to achieve a consistent standard of marking and feedback quality amongst different tutors.

To address this issue the authors initially used a number of methods including double blind marking and re-marking a random sample of assessments to support consistent grading. However, with increasing demands on academics these time-consuming activities became an unrealistic option. In our first effort to improve the process we used the benchmarking tool in SPARK^{PLUS} to allow tutors to compare both the average mark and feedback provided by tutors for a number of selected project tasks (Willey & Gardner 2010, a & b)

This activity proved effective in reducing the variability in marking amongst different tutors. Furthermore, we found that using a software tool to record tutor assessments and feedback before exploring their understanding in a subsequent discussion activity promoted inclusiveness of less experienced and less confident tutors by allowing all points of view to be considered. As a result of this initial research, we developed a number of new SPARK^{PLUS} features to promote more in depth conversations to improve both the standard and consistency of tutor marking.

In this paper, we report a pilot study conducted using the new SPARK^{PLUS} features to examine the mechanisms by which tutors develop a shared understanding of assessment requirements through these collaborative conversations.

New Software Features

The new multiple assessor mode in SPARK^{PLUS} (SPARK^{PLUS} 2011) allows participants to rate work and provide written feedback on categories of criteria. After the activity, participants can compare their rating and feedback to those of other participants that is provided anonymously. Individual ratings are displayed by using colour-coded triangles superimposed on a rating slider (Figure 1).

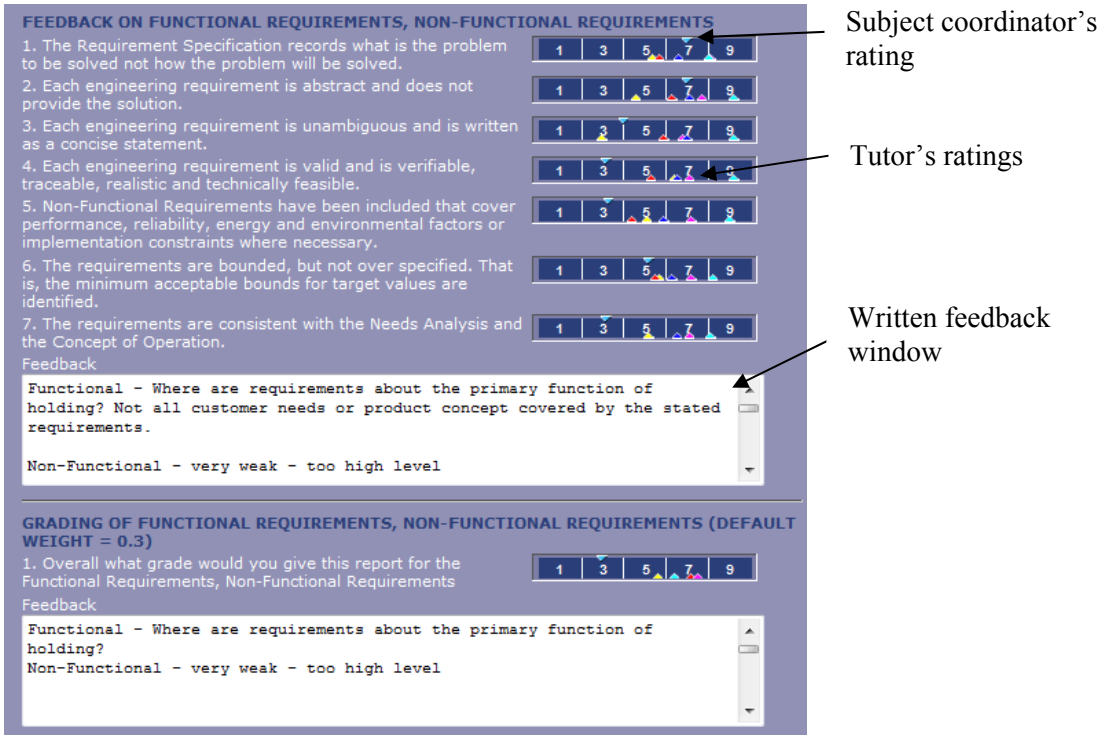

Figure 1: Benchmarking results screen in SPARK^{PLUS}: Upper triangle shows coordinator's marking, lower triangles show individual tutors' marking of this report.

Written feedback from different participants is displayed anonymously in viewing windows provided for each category of criteria. In the instance shown in Figure 1, the ratings of the subject coordinator are shown on the top of each slider while those of the participating tutors are shown on the bottom.

Participants also receive feedback via the rating and feedback summary screens (Figures 2 & 3 respectively). The rating summary screen shown in Figure 2 provides histograms (which expand when clicked) showing the distribution of ratings across the rating scale frequency bins. An associated slider also shows the minimum, maximum, average and standard deviation of participants' ratings making it easy to identify the criteria where there is general agreement and those where participants have quite different opinions. This allows discussion to focus quickly on those areas that need to be addressed.

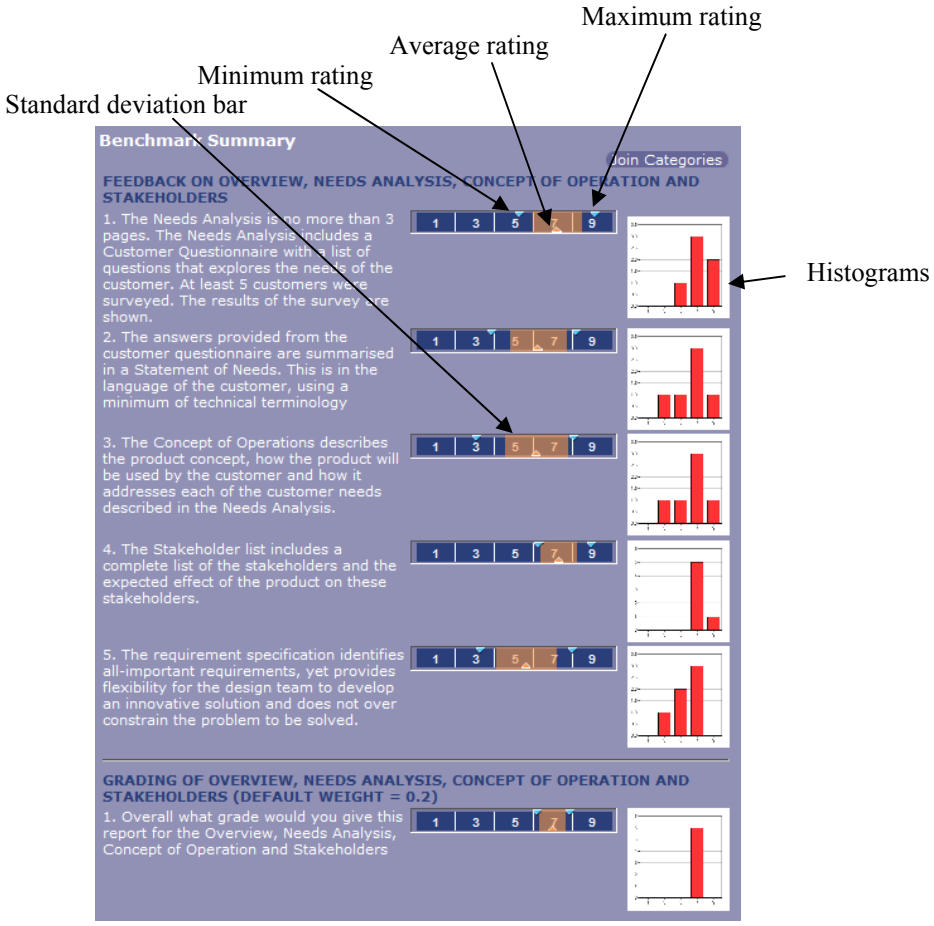

Figure 2: SPARK[PLUS] Benchmarking rating summary screen (only the first two categories have been shown)

While the feedback summary screen allows participants' comments for each category to be viewed for either all or one particular rating level (Figure 3).

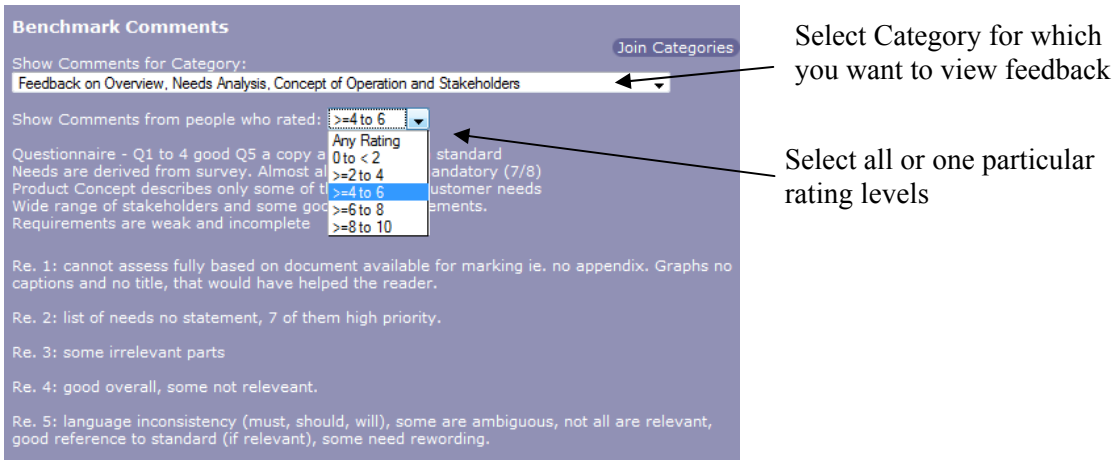

Figure 3: SPARK[PLUS] feedback summary screen

Method

A second year core (all disciplines) engineering degree subject, at the University of Technology, Sydney, typically has an enrolment of approximately 300+ students per semester. In addition to lectures, students are distributed amongst ten tutorials where individual tutors are responsible for grading assessment tasks. One of the assessment tasks for this subject is the Requirements Specification report which was used in the benchmarking activity described in this paper.

The reported investigation, conducted during Autumn semester 2011, consisted of a number of stages:

Prior to Tutor meeting
Stage 1: Tutors were provided with a copy of two Requirements Specification reports from the current semester. (These functioned as the 'yardstick' assessments as suggested by Tomkinson & Freeman (2007)). Tutors marked these reports against specified criteria and entered their assessment (marks and feedback comments) into the multiple assessor tool in SPARK^{PLUS} (Figure 1). The subject co-ordinator also entered his assessment (marks and feedback comments) into SPARK^{PLUS} to allow comparison with the tutors' assessment.

During Tutor Meeting
Stage 2: Tutors logged on to SPARK^{PLUS} and compared their marking and feedback to that of the other tutors (displayed anonymously) and the subject coordinator.

Stage 3: Tutors were formed into a group and asked to discuss their individual marking (previously recorded in SPARK^{PLUS}) and subsequently to collaboratively re-mark one report i.e. they were required to reach a consensus about the appropriate mark for each assessment criterion and agree on an overall holistic grade for the submission. (For this part of the exercise, we chose the Requirements Specification report with the most variation in ratings and feedback comments)

Stage 4: The subject co-ordinator explained his marking of the reports which was compared to that of the tutors. Subsequent discussions explored the reasons for particular ratings and grading differences.

At various stages in the project (pre, during and post meeting) tutors were asked to complete a series of reflective questionnaires that consisted mainly of open-ended questions. Subsequently, tutors were interviewed to further explore the impact of the reported exercise. The authors also observed the interaction amongst tutors and kept notes during the tutor meeting.

Results and Discussion

In the reported semester there were ten tutorials taught by six staff (five tutors and the subject coordinator) all of which agreed to participate in the pre-tutor meeting activities. However, only four tutors and the subject coordinator participated in the tutor meeting.

Prior to the tutor meeting, these four tutors were asked to assess their perceived level of expertise and confidence in the subject material, understanding of the assessment criteria and their capacity to grade and give feedback to students on their reports. The results of these assessments are shown in Table 1.

Table 1: Results of pre-activity survey.

Selected Questions From Survey 1	n = 4		
	Low	Intermediate	High
My expertise in the subject material covered in these reports is:			4
I am confident in my ability to grade these reports to the required standard.			4
I am confident that I understand / interpret the assessment criteria.		1	3
I am confident that I can clearly articulate and explain the strengths and weaknesses of these reports to students when I provide them with feedback.		2	2

All the tutors rated themselves as having high subject material expertise and confidence in their ability to mark the reports. This is not surprising as each was experienced, having tutored the subject for at least three semesters. Some of the tutors were less confident with their understanding of the assessment criteria and their ability to provide student feedback.

Each of the two reports was assessed against 20 criteria spread across four different categories. Despite the relative experience of the four tutors, they reported that during the conversations they questioned or changed their understanding of both the subject material and the assessment criteria for at least one criterion and up to three criteria as shown in Table 2.

Table 2: Impact of activity conversations for each tutor.

UNDERSTANDING OF SUBJECT MATERIAL	Tutor 1	Tutor 2	Tutor 3	Tutor 4
For how many criteria did the conversations with others cause you to **question or change** your **understanding** of the subject material?	1	2	1	3
INTERPRETATION/UNDERSTANDING OF ASSESSMENT CRITERIA				
For how many criteria did the conversations with others cause you to **question or change** your **interpretation/understanding** of the assessment criteria?	2	1	1	3

Arguably more important, three tutors (one tutor did not respond to these questions) and the subject coordinator reported that the discussions with their peers would impact/alter the way they mark their reports, the issues on which they would give feedback and the language they will use to provide this feedback (Table 3). Interestingly, the subject coordinator who arguably knew more about the subject than the tutors, also reported that the discourse increased his understanding of the subject material "A Lot".

Table 3: Impact of focussed conversations amongst tutors

My conversations about the Requirement Specifications report with the other tutors has impacted/altered	**n = 3 + Subject Coordinator (SC)**		
	Not at all	**A Little**	**A Lot**
The way that I will mark the reports in my tutorial group			3, SC
The aspect/topics/issues on which I will give feedback		2	1, SC
The language that I will use to provide feedback to students		2	1, SC
Increased my understanding of the subject material	2	1	SC

Three of the tutors reported that they learnt equally from explaining their opinions to others as they learnt from listening to the opinions of their peers. The remaining tutor reported his learning usually occurred when listening to other members of the group. Not surprisingly, we observed this tutor to be the least confident in their discussions with the others.

The tutors and subject coordinator also commented that while their ability to mark the reports and provide appropriate feedback to students increased through reading the ratings and feedback provided by others in SPARK^{PLUS}, the most significant improvement resulted from the discussions. However, all participants agreed that without the availability of this data and the clear way in which SPARK^{PLUS} presented it, they felt the discussions would have taken much longer and been less focused and effective.

When asked to describe the best thing about having a focussed conversation with other tutors about grading and feedback typical responses were:

> *"It validated my understanding of the subject and fine tuned a few concepts"*
>
> *"I realising that I marked too easily"*
>
> *"It was a fast way to get issues discussed and resolve differences"*

After the collaborative re-mark, it was interesting to note that both the grade and mark given to the report reduced significantly to a 65% (mark) Credit (grade), from an average 77% (mark) Distinction (grade) (individual tutor marks pre-activity were 81%, 75%, 78% and 74%). On investigating this difference, we found that it was mostly a result of changes in the tutors' expectation of what was

required in the test plan and the report presentation including the standard of grammar, spelling etc. Through their discussion, the tutors raised their expected standard for these two aspects of the report. Furthermore, the tutors reported that the activity would have a significant impact on the topics, language and breath of the feedback that they would give students. One tutor commented that through the discussions they "*learnt strategies/tactics for dealing with student issues",* while another tutor remarked that *"talking first -hand and listening to the other tutors thinking"* helped him clarify what was required by the students to demonstrate the subject learning outcomes.

The participants agreed with our observations that having all points of view recorded and available anonymously promoted inclusivity as it ensured all differences in opinions were available to be discussed irrespective of the participant's expertise. A number of the participants commented that in the absence of this reporting they may not have been confident to report their opinion especially where it differed significantly from the rest of the group. Furthermore, having these differences so clearly reported discouraged group think or being distracted by and subsequently focusing discussions on only one or two issues.

In summary, although only a pilot trial the results support the findings of other researchers (Tomkinson & Freeman 2007, Price 2005, White 2006) that conversations between academics are an effective method of developing a shared understanding of assessment criteria and improve marker consistency. In addition, this study suggests that the addition of the software tool has potential to make the process more inclusive, whilst enabling conversations to focus quickly on the pertinent issues.

Conclusion

Our findings support the conclusions of other researchers who found that conversations with other academics about assessment standards and marking is an effective method of developing a shared understanding of assessment criteria and improving marker consistency. In addition, the results suggest that the software tool has the potential to make the process both more inclusive by providing the opportunity for all to present their opinion irrespective of their perceived expertise or confidence and efficient by clearly identifying differences in opinions, enabling conversations to focus quickly on the pertinent issues.

References

McCallum N, Bondy J., & Jollands M. (2008) Hearing each other - how can we give feedback that students really value. *Proceedings of the Nineteenth Annual Conference of the Australasian Association for Engineering Education* , Yeppoon, Australia, 7-10 December. Faculty of Sciences, Engineering & Health, CQU University.

Price M. (2005) Assessment standards: the role of communities of practice and the scholarship of assessment. *Assessment & Evaluation in Higher Education* Vol. 30, No. 3, pp. 215-230.

SPARK^{PLUS} (2011), http://spark.uts.edu.au last viewed 9th August, 2011.

Tomkinson B & Freeman J. (2007) Using portfolios for assessment: problems of reliability or standardisation? in *Enhancing Higher Education, Theory and Scholarship, Proceedings of the 30th HERDSA Annual Conference*, Adelaide, Australia, 8-11 July.

White, N. R. (2006) Tertiary education in the Noughties: the student perspective, *Higher Education Research & Development*. Vol. 25, No. 3, pp. 231-246.

Willey, K. & Gardner, A.. (2010a) Perceived Differences in Tutor Grading in Large Classes: Fact or Fiction? *Proceedings of the 40th ASEE/IEEE Frontiers in Education Conference.* Virginia, USA, 27 – 30 October.

Willey, K & Gardner A. (2010b) Improving the standard and consistency of multi-tutor grading in large classes, *Assessment: Sustainability, Diversity and Innovation, Proceedings of the ATN Assessment Conference*, The University of Technology, Sydney, 18 & 19 November.