# Building a community of practice to improve inter marker standardisation and consistency

Keith Willey
University of Technology, Sydney
PO Box 123
Broadway, NSW. 2007
+61 2 9514 7605
Keith.Willey@uts.edu.au

Anne Gardner
University of Technology, Sydney
PO Box 123
Broadway, NSW. 2007
+61 2 9514 2622
Anne.Gardner@uts.edu.au

## ABSTRACT

Over several years the authors have coordinated engineering subjects, with large cohorts of up to 300+ students. In each case, lectures were supported by tutorials. In the larger subjects it was not uncommon to have in excess of 10 tutors, where each tutor is responsible for grading the assessment tasks for students in their tutorial. A common issue faced by lecturers of large multiple tutor subjects is how to achieve a consistent standard of marking between different tutors. To address this issue the authors initially used a number of methods including double-blind marking and remarking. This process was improved by using the benchmarking tool in SPARK$^{PLUS}$ [1] to compare both the grading and feedback provided by different tutors for a number of randomly selected project tasks. In these studies we found that while students' perception of difference in grading was not unfounded, the problem was exacerbated by inconsistencies in the language tutors use when providing feedback. In this paper, we report using new SPARK$^{PLUS}$ features developed as a result of this previous research to quickly establish and build a community of practice amongst subject tutors. We found that in just one session these processes assisted tutors to reach a higher level of shared understanding of the concepts and practices pertinent to the subject assessment activities. In addition, it enabled tutors to gain an appreciation of the grading issues frequently reported by students. This resulted in not only improving both the understanding and skills of tutors but changing the way they both marked and provided feedback.

## 1. 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 [2],[3].

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. These issues contribute to the fact that some students feel that grades are a function of who's tutorial they find themselves in, described as "tut lotto" [4].

For consistent marking between tutors it is important for all assessors to share a common view of the value of a given grade. Tomkinson [5] suggests 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 [2] report that: "An assessment standards discourse is needed to support the functioning of assessment communities of practice…"(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" this being true for both tutors and students [3 p.237]

The authors have regularly coordinated engineering subjects, with large cohorts (up to 300+ students). In each case, lectures were supported by tutorials. In the larger subjects it was not uncommon to have in excess of 10 tutors, where each tutor is responsible for grading the assessment tasks for students in their tutorial. A common issue faced by lecturers of large multi-tutor subjects is how to achieve a consistent standard of marking between different tutors.

To address this issue the authors initially used a number of methods including double-blind marking and remarking to support consistent grading. However, with both increasing student numbers and teaching loads these activities are fast becoming an unrealistic option. These processes were improved by using the benchmarking tool in SPARK$^{PLUS}$ to allow tutors to compare both the average grading and feedback provided by tutors for a number of selected project tasks [6 -7].

These activities were effective in reducing the variability in marking between different tutors. Furthermore, we found that using a software tool to record tutor assessments and feedback before exploring their understanding in a subsequent discussion

1st WORLD ENGINEERING EDUCATION FLASH WEEK LISBON 2011

*WEE2011*, September 27-30, 2011, Lisbon, Portugal.
Editors: Jorge Bernardino and José Carlos Quadrado.

activity promoted inclusiveness of less experienced and less confident tutors. In addition, we found that while students' perception of difference in grading between tutorials was not unfounded, the problem was exacerbated by inconsistencies in the language tutors use when providing feedback.

These studies supported the conclusions of other researchers that conversations about assessment standards and marking is an effective method of developing a shared understanding of assessment criteria and improving the standard, and consistency with marking [2-7].

As a result of this previous research [6] we developed a number of new SPARK$^{PLUS}$ features to promote further improvement in both the standard and consistency of tutor marking and the quality of student feedback.

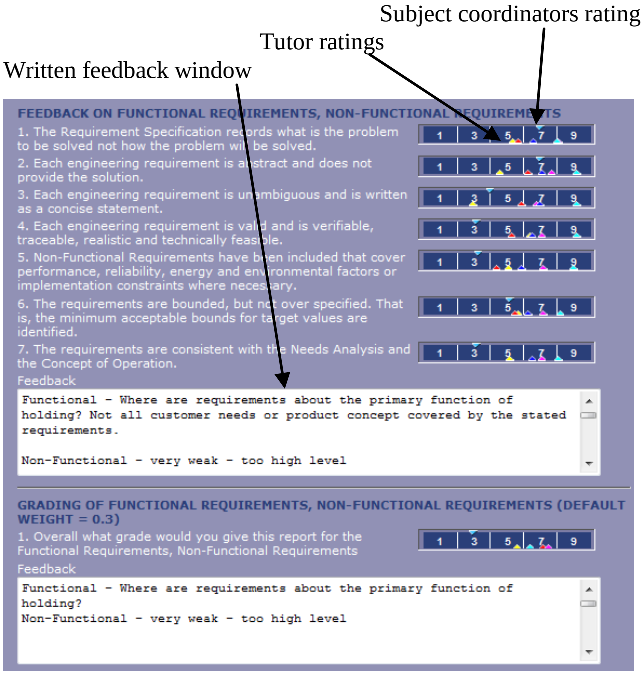
A subsequent study was conducted to:

- investigate the impact of these new SPARK$^{PLUS}$ features on tutor learning and understanding of the issues associated with using multiple markers from both an academic and student's point of view.

and

- examine the mechanisms by which tutors learn through collaboration

In this paper we report on the former and find that in a single session these features assisted tutors to reach a higher level of shared understanding of the concepts and practices pertinent to the subject assessment activities.



FEEDBACK ON FUNCTIONAL REQUIREMENTS, NON-FUNCTIONAL REQUIREMENTS

1. The Requirement Specification records what is the problem to be solved not how the problem will be solved.
2. Each engineering requirement is abstract and does not provide the solution.
3. Each engineering requirement is unambiguous and is written as a concise statement.
4. Each engineering requirement is valid and is verifiable, traceable, realistic and technically feasible.
5. Non-Functional Requirements have been included that cover performance, reliability, energy and environmental factors or implementation constraints where necessary.
6. The requirements are bounded, but not over specified. That is, the minimum acceptable bounds for target values are identified.
7. The requirements are consistent with the Needs Analysis and the Concept of Operation.

Feedback

```
Functional - Where are requirements about the primary function of
holding? Not all customer needs or product concept covered by the stated
requirements.

Non-Functional - very weak - too high level
```

GRADING OF FUNCTIONAL REQUIREMENTS, NON-FUNCTIONAL REQUIREMENTS (DEFAULT WEIGHT = 0.3)

1. Overall what grade would you give this report for the Functional Requirements, Non-Functional Requirements

Feedback

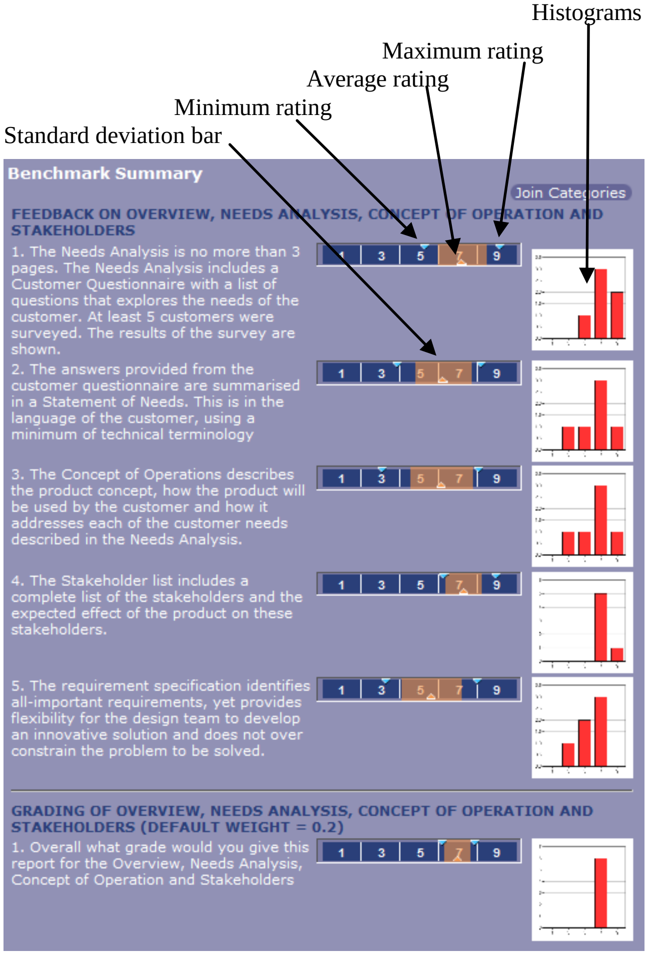
```
Functional - Where are requirements about the primary function of
holding?
Non-Functional - very weak - too high level
```

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

## 2. NEW SPARK$^{PLUS}$ FEATURES

The new multiple assessor mode in SPARK$^{PLUS}$ 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 are provided anonymously. Individual ratings are displayed by using colour-coded triangles superimposed on a rating slider. Feedback from different participants is also displayed anonymously in viewing windows provided for each category of criteria. In the instant shown in figure 1 the ratings of the course coordinator are shown on the top of each slider while those of the participating tutors are shown on the bottom. The previous version of SPARK$^{PLUS}$ only showed participants their rating, the average rating of their peers and the instructors rating. Also written feedback could only be provided overall and not on a category basis.

Participants may also receive feedback by viewing either the rating or feedback summary screens. The rating summary screen shown in figure 2 provides histograms (which expand when clicked) showing the distribution of ratings across a maximum of five frequency bins. An associated slider also shows the minimum, maximum, average and standard deviation of participants' ratings.



Benchmark Summary

FEEDBACK ON OVERVIEW, NEEDS ANALYSIS, CONCEPT OF OPERATION AND STAKEHOLDERS

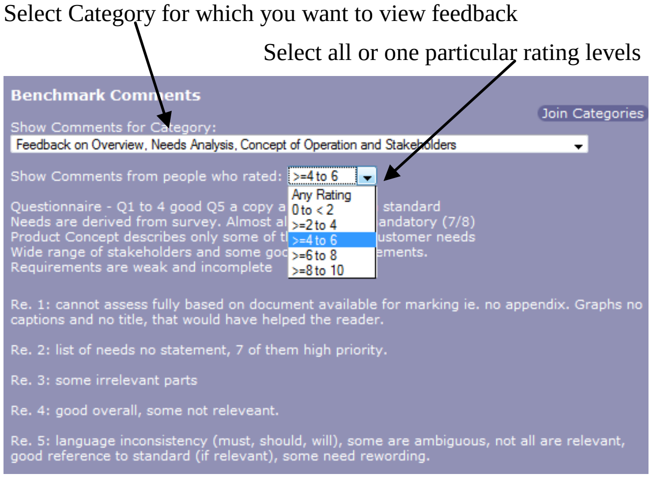
1. The Needs Analysis is no more than 3 pages. The Needs Analysis includes a Customer Questionnaire with a list of questions that explores the needs of the customer. At least 5 customers were surveyed. The results of the survey are shown.
2. The answers provided from the customer questionnaire are summarised in a Statement of Needs. This is in the language of the customer, using a minimum of technical terminology
3. The Concept of Operations describes the product concept, how the product will be used by the customer and how it addresses each of the customer needs described in the Needs Analysis.
4. The Stakeholder list includes a complete list of the stakeholders and the expected effect of the product on these stakeholders.
5. The requirement specification identifies all-important requirements, yet provides flexibility for the design team to develop an innovative solution and does not over constrain the problem to be solved.

GRADING OF OVERVIEW, NEEDS ANALYSIS, CONCEPT OF OPERATION AND STAKEHOLDERS (DEFAULT WEIGHT = 0.2)

1. Overall what grade would you give this report for the Overview, Needs Analysis, Concept of Operation and Stakeholders

**Figure 2: SPARKPLUS Benchmarking rating summary screen (only the first two categories have been shown)**

The rating summary screen makes it easy for both participants and the instructor to observe the criteria where there is general agreement and those where participants have quite different opinions allowing discussion to focus on those areas that need to be addressed the most. Another advantage of this screen is that it allows participants to see the value of using a combination of a formative feedback rubric and holistic marking. In the activity shown in figure 2 the first five criteria capture participant's opinions as to the standard of work in regard to each criterion. These criterion were formative (zero weighted) that is, their value was not taken into account in calculating the overall mark. The purpose is to provide feedback on the strengths, weaknesses and quality of the work being assessed without the limitations imposed when one has to consider how each criterion contributes to a student's overall summative grade. The sixth slider (slider one in the second category) was used to holistically rate the quality of the work for the assessment category. Notice the histograms for the sixth slider shows that holistically all the participants (tutors in this case) rated the work to be relatively the same (all ratings fell in the same frequency bin). Conversely, the histograms for the first five formative sliders show the variety of opinions held by the different participants as to the quality of work against individual criterion. If these criteria had been summative students would have with some justification argued about the differences in ratings provided by different tutors. Being formative, they exist to provide students with feedback as to the quality of their work with only the holistic slider contributing to their final grade (the sum of the parts do not necessarily make the whole).


Select Category for which you want to view feedback

Select all or one particular rating levels

Benchmark Comments

Show Comments for Category: Feedback on Overview, Needs Analysis, Concept of Operation and Stakeholders

Show Comments from people who rated: >=4 to 6

Questionnaire - Q1 to 4 good Q5 a copy a standard
Needs are derived from survey. Almost al andatory (7/8)
Product Concept describes only some of tl ustomer needs
Wide range of stakeholders and some goc ements.
Requirements are weak and incomplete

Re. 1: cannot assess fully based on document available for marking ie. no appendix. Graphs no captions and no title, that would have helped the reader.

Re. 2: list of needs no statement, 7 of them high priority.

Re. 3: some irrelevant parts

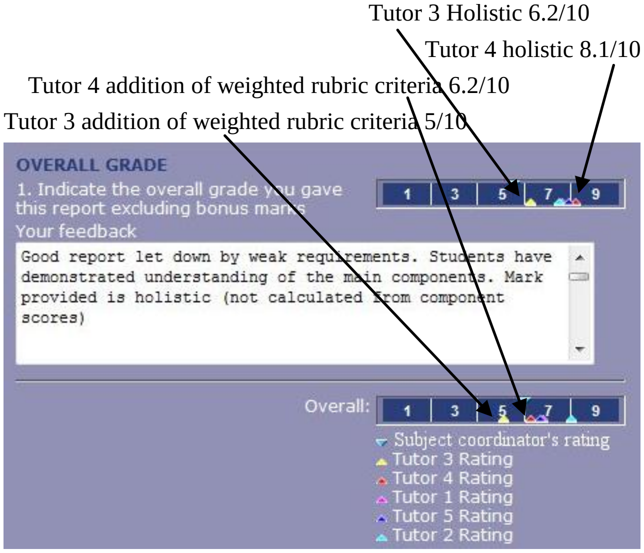
Re. 4: good overall, some not releveant.

Re. 5: language inconsistency (must, should, will), some are ambiguous, not all are relevant, good reference to standard (if relevant), some need rewording.

**Figure 3: SPARK$^{PLUS}$ Benchmarking comment summary screen**

The comment summary screen shown in figure 3 provides participants with the opportunity to view anonymously the feedback comments provided by their peers. This screen is particularly useful for participants to view the feedback of those with opinions different from their own. For example, let us say that on a particular criterion you rated the piece of work high while a number of other participants rated it low. By selecting an individual rating range you can view all the feedback comments of participants who rated within this range. At first this may not appear to be a very significant feature, however, consider an activity in a class with a thousand students. It is very useful to be able to have the program automatically enable you to view the comments of people who for example disagreed with your opinion.

The final feature that will be discussed in this paper is the capacity of the program to provide a comparison between the overall holistic grade provided by participants and the grade determined if each of the individual criterion with appropriate waiting contributed summatively. This comparison is shown in figure 4 where it can be seen that the lowest rating in each case is provided by tutor 3. When marking holistically Tutor 3 awarded 6.2/10. Conversely when marking using the weighted rubric tutor 3 awarded a 5/10. Similarly, tutor 4 provided the highest holistic mark (8.1/10 top slider) and a comparatively low 6.2/10 using the weighted rubric (bottom slider). Again demonstrating that the sum of the parts (marks awarded using a multi-criteria rubric) often do not reflect the overall grade that would be awarded using holistic judgement.


Tutor 3 Holistic 6.2/10

Tutor 4 holistic 8.1/10

Tutor 4 addition of weighted rubric criteria 6.2/10

Tutor 3 addition of weighted rubric criteria 5/10

OVERALL GRADE

1. Indicate the overall grade you gave this report excluding bonus marks

Your feedback

Good report let down by weak requirements. Students have demonstrated understanding of the main components. Mark provided is holistic (not calculated from component scores)

Overall:

**Figure 4: SPARK$^{PLUS}$ comparison of holistic and summative criteria grading.**

While SPARK$^{PLUS}$ has a number of other new features to provide feedback to assist the subject coordinator or instructor to be aware of marking issues in this paper we restrict our discussion to the features described above as they are available to all participants (tutors and instructor) in the activity.

# 3. METHOD

A second year 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.

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

Stage 1: Tutors were provided with a copy of two reports from the current semester. Tutors graded these reports against specified criteria and entered their assessment (grading and feedback comments) into the new benchmarking task in SPARK$^{PLUS}$ (partial screen shot Figure 1). The course co-ordinator also entered their assessment (grading and feedback comments) into SPARK$^{PLUS}$ to allow comparison with the tutors' grading and feedback.

Stage 2: At a tutor meeting, tutors were asked to read a list of feedback comments provided by tutors for different criteria categories and indicate the grade/mark they would have given in each case if they had given this comment.

Stage 3: Tutors were then asked to logon to SPARK$^{PLUS}$ and compare their marking and feedback to that of the other tutors (that was displayed anonymously) and the course coordinator.

Stage 4: Tutors were formed into a group and asked to discuss their individual grading (previously recorded in SPARK$^{PLUS}$) and subsequently collaboratively re-grade the reports ie they were required to reach a consensus about the appropriate grade for each assessment criterion and agree on an overall holistic grade for the submission.

Stage 5: The course co-ordinator then explained how they had graded the reports and the group compared their grading with the course co-ordinator's. Subsequent discussions were held in which the differences in grading were explored and discussed.

Stage 6: Tutors were guided through a discussion where difference in the grading and feedback comments provided by individual tutors were compared and examined from a student's point of view.

Stage 7: The Tutor meeting concluded with tutors being encouraged to go home and use SPARK$^{PLUS}$ to further analyse, examine and reflect on their individual grading and their feedback comments for each report compared to that of both the course coordinator's and the other tutors' to benchmark their judgement.

At various stages in the project (pre, during and post activity) 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 between tutors and kept notes during the Tutor meeting.

## 4. RESULTS / DISCUSSION

In the reported semester there were ten tutorials taught by a total of six tutors. Of these, five tutors and the subject coordinator agreed to participate in the pre-tutor meeting activities (marking and providing feedback on the reports in SPARK$^{PLUS}$). However, only four of these tutors and the subject coordinator attended the tutor meeting. All of the tutors had prior experience using the earlier version of SPARK$^{PLUS}$ which was used to facilitate student collaborative learning activities for their tutorial four times a semester.

Prior to the exercise, tutors were asked to assess their level of expertise and confidence in the subject material, understanding the assessment criteria, capacity to grade and give feedback to students on their reports. The results of these assessments for the subject coordinator and the four tutors who participated in the Tutor meeting activities are shown in table 1 (five participants in total, hence single participant response is equivalent to 20%).

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

In the Tutors meeting Tutors were asked to logon to SPARK$^{PLUS}$ and compare their marking and feedback to that of the other tutors (that was displayed anonymously) and the course coordinator (stage 3). They were encouraged to use both the results (figure 1) and summary screens (figure 2 and 3) in making their comparisons.

**Table 1: Results of pre-activity survey.**

| Selected Questions From Survey 1 | Low | Intermediate | High |
|---|---|---|---|
| My expertise in the subject material covered in these reports is: | | | 100% |
| I am confident in my ability to grade these reports to the required standard. | | | 100% |
| I am confident that I understand / interpret the assessment criteria. | | 20% | 80% |
| 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. | | 40% | 60% |

Afterwards tutors were asked what impact did being able to see everyone else's individual criterion, category and overall ratings and feedback as opposed to only the average rating of their peers (previous version of SPARK$^{PLUS}$) have on their confidence in marking.

All the tutors reported a positive impact saying that the screens made it easy for them to observe where their opinions differed from the other tutors. They were able to clearly identify where there was the most disagreement and where they agreed. These opinions are reflected in the following comments:

*"I was able to see what they (tutors) were thinking"* to both *"learn and improve my own technique".*

*"I was able to get a feel for how others mark ... I found it a learning experience"*

*"I could see that we all marked the overall score the same.. giving me more confidence in the task of marking"*

Furthermore, from just viewing the screens tutors formed opinions about their marking ability. For example, one tutor commented that:

*"I'm too lenient. I need to put more effort into marking the content of the reports-rather than the style."*

Tutors reported the feedback comments summary screen helped them to understand the reasons for marking differences.

*"I was able to see what they were thinking and learn and improve my own (feedback) technique."*

While the results were reported anonymously, there were instances of Tutors feeling some anxiety when their assessment and/or feedback differed significantly from the rest of the cohort.

*"I felt disappointed that I wasn't close to the average with the marking".*

*"I felt a bit worried when my feedback differed from the common/majority"*

Furthermore, tutor's agreed that observing the differences between their grading and feedback comments helped them understand both the issues involved in having multiple tutors and students concerned with inconsistent marking and feedback from different tutors

*"I can see consistency across the tutors is important"*

*"Greatly assisted in my understanding of the different emphasis markers are providing".*

*"I can see the potential for frustration by the students"*

In the author's experience in large classes using multiple markers students often focus on differences in marking on individual criteria between student submissions. Their focus typically being to argue for an increase mark or in some cases a fraction of a mark rather than focusing on the feedback provided and the overall quality of their work. In response to this, the authors changed their marking to provide formative feedback on individual criterion and overall holistic grades. In addition, grades were not released until after student had been given an opportunity to reflect on the feedback provided. Understandably, some academics who have been using detailed marking rubrics are somewhat reluctant to move to holistic grading. This is particularly apparent in subjects where assessment tasks are mainly analytic in nature. It is the author's opinion that even for analytic questions where for example a minor calculation error has been made, if the student provides an obviously incorrect answer they have not demonstrated both the required judgement nor capacity to satisfactorily meet the associated subject learning outcomes and hence should be graded accordingly (unsatisfactory). Alternatively, it could be argued a student that identifies they know their answer must be incorrect and can explain why even though they cannot find the error has demonstrated the judgement and capacity to meet the associated learning outcomes and hence receive a passing grade.

The authors have found that holistic marking is more likely to be adopted if an academic discover the benefits themselves. To assist them in this process we deliberately designed the new features of SPARK$^{PLUS}$ to help academics appreciate these benefits.

In the reported activity tutors were asked to observe the difference between the individual criteria ratings and the holistic ratings using both the results and summary screens. Afterwards tutors were asked how their observations impacted their understanding of students concerns that the sum of the individual criteria feedback does not always match their final mark (figure 2 and 4). For example referring to figure 2 tutors marked quite differently against the individual criterion within the requirements category (first five histograms) while their overall grade for this category (bottom histogram in the figure) was relatively the same.

The subject coordinator reported that *"observing the difference between the individual criteria and the holistic overall mark showed me the variation in the feedback vs the mark. Seeing this across several tutors explained (to me) why students see variation between the tutors (markers)".*

While tutors commented that:

I was *"surprised how closely aligned the overall ratings were"*.

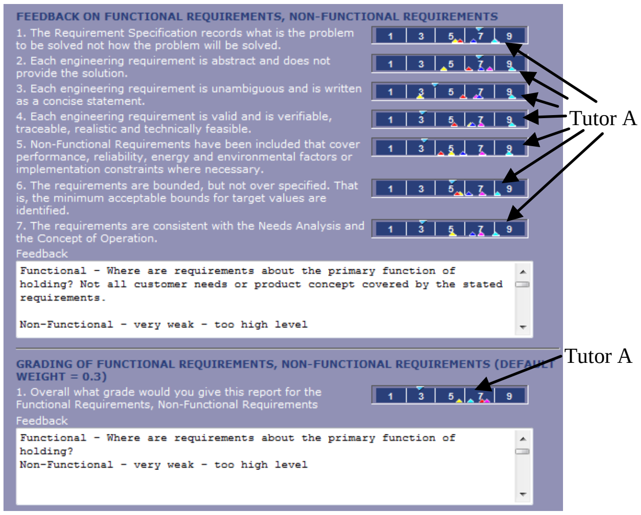
*"Using grades and marking criteria holistically is better than using numerical methods"*

I now *"won't give a numerical mark for the subsections but a grade"*

*"I was pleased that our final marks were reasonably close but can understand why students may be upset by the variation in the subsections"*

At the end of the session the authors took the opportunity to highlight anonymously different tutor ratings that had inconsistencies and would have likely led to students being dissatisfied with their marking and/or feedback. For example figure 5 shows that Tutor A gave the highest rating of any participant on each of the individual criteria for this category (average 8.9/10) and provided feedback that indicated the submission was "*very good*". However, their overall category rating was the third lowest being only 6.1/10. It is these differences between feedback on criteria and overall grade that gives students the perception that their grade is unfair, even if it is not.

It should be noted that Tutor A was an experienced tutor who provided fair overall grades. In fact there overall submission mark in the reported exercise was the median (middle) of all the participants. Hence, it was not the standard of their overall marking that was an issue just the inconsistencies between both the criterion feedback and the overall grade.


FEEDBACK ON FUNCTIONAL REQUIREMENTS, NON-FUNCTIONAL REQUIREMENTS

1. The Requirement Specification records what is the problem to be solved not how the problem will be solved.
2. Each engineering requirement is abstract and does not provide the solution.
3. Each engineering requirement is unambiguous and is written as a concise statement.
4. Each engineering requirement is valid and is verifiable, traceable, realistic and technically feasible.
5. Non-Functional Requirements have been included that cover performance, reliability, energy and environmental factors or implementation constraints where necessary.
6. The requirements are bounded, but not over specified. That is, the minimum acceptable bounds for target values are identified.
7. The requirements are consistent with the Needs Analysis and the Concept of Operation.

Feedback

```
Functional - Where are requirements about the primary function of
holding? Not all customer needs or product concept covered by the stated
requirements.

Non-Functional - very weak - too high level
```

GRADING OF FUNCTIONAL REQUIREMENTS, NON-FUNCTIONAL REQUIREMENTS (DEFAULT WEIGHT = 0.3)

1. Overall what grade would you give this report for the Functional Requirements, Non-Functional Requirements

Feedback

```
Functional - Where are requirements about the primary function of
holding?
Non-Functional - very weak - too high level
```

**Figure 5: Benchmarking results screen in SPARKPLUS : Upper triangle shows coordinator's marking, lower triangle shows individual tutor's marking against the "Requirements" category. The ratings of Tutor A have been identified.**

All the participants commented that the discussion helped them appreciate the benefits of holistic marking. They all indicated that as a result of the reported activity they would make changes to the way they provide feedback and present their marks in the future. Most comments related to combining the use of formative feedback rubrics with holistic grading.

The authors suggest using criteria to provide formative feedback on the strengths and weaknesses of a submission. This releases the academic from trying to balance a summative rubric to add up to the holistic grade their judgement tells them the submission deserves. In addition, students are free to focus and reflect on the feedback provided discussing specific issues highlighted with the tutor to build their understanding and learn rather than focusing on increasing their mark. We recommend providing these formative feedback rubrics before students are given their final grade. Only after students have reflected on and discussed the feedback with the tutor should grades be released.

If after getting their grade students wish to argue for an increase in their mark then they must do it holisticly, e.g. explain why their submission is satisfactory (pass), credible (credit), distinctive (distinction) or highly distinctive (high distinction) as described by the Grades and Descriptors used at the University of Technology, Sydney shown in Table 2.

**Table 2: Grades and Descriptors used at the University of Technology, Sydney [8]**

| Grade | Descriptor |
|---|---|
| High Distinction | Work of outstanding quality on all objectives of the subject, which may be demonstrated by means of criticism, logical argument, interpretation of materials or use of methodology. This grade may also be given to recognise particular originality or creativity. |
| Distinction | Work of superior quality on all objectives, demonstrating a sound grasp of content, together with efficient organisation and selectivity. |
| Credit | Work of good quality showing more than satisfactory achievement on all objectives, or work of superior quality on most of the objectives. |
| Pass | Work showing a satisfactory achievement on the overall objectives of the subject. |
| Fail | Unsatisfactory performance in one or more objectives of the subject as contained within the assessment items. |

In summary, we found that the new SPARK$^{PLUS}$ features assisted tutors to reach a higher level of shared understanding of the concepts and practices pertinent to the subject assessment activities in a single session. In addition, they enabled tutors to gain an appreciation of the grading issues frequently reported by students. This resulted in not only improving the understanding and skill of individual tutors but changed the way they both marked and provided students with feedback.

## 5. CONCLUSION

The activity was effective in helping tutors to benchmark and reflect on their marking judgement.

The presented process promoted inclusiveness by using an anonymous software tool to record tutor assessments and feedback before exploring their understanding in a subsequent discussion activity. The benchmarking activity was particularly effective in helping to develop marking standards and feedback skills.

We found that the new features of SPARK$^{PLUS}$ helped even experienced tutors. Tutors reported that the new screens helped them to learn and improve their marking. They made it easy for them to observe differences in opinion between tutors, enabling them to quickly identify where there was the most disagreement and where they agreed.

Furthermore, being able to observe the differences between grading and feedback comments helped them understand the issues involved in having multiple tutors and students concerned with inconsistent marking and feedback. All participants indicated that as a result of these observations in the future they intended to use a combination of formative feedback rubrics and holistic marking.

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 both marker consistency and student satisfaction with feedback.

## 6. REFERENCES

[1] SPARK$^{PLUS}$ http://spark.uts.edu.au last viewed 9th May, 2011.

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

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

[4] McCallum N, Bondy J., & Jollands M.. “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, 2008). Faculty of Sciences, Engineering & Health, CQU University.

[5] Tomkinson B & Freeman J. “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, 2007).

[6] Willey K & Gardner A. “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, 2010)

[7] Willey K & Gardner A., "Improving the standard and consistency of multi-tutor grading in large classes", Assessment: Sustainability, Diversity and Innovation. A conference on assessment in higher education, The University of Technology Sydney Nov 2010

[8] UTS Grades and Descriptors http://www.gsu.uts.edu.au/rules/s3.html