Calibrating Assessment Literacy Through Benchmarking Tasks

Simon Knight[a]*, Andrea Leigh[b*], Yvonne C. Davila[b], Leigh J. Martin[b], Daniel W. Krix[b]

[a]Faculty of Transdisciplinary Innovation, University of Technology Sydney, Sydney, Australia; [b]Faculty of Science, University of Technology Sydney, Sydney, Australia University, City, Country

University of Technology Sydney, PO Box 123, Broadway, NSW 2016, Australia, Simon.Knight@uts.edu.au; Andrea.Leigh@uts.edu.au

Note the first author led analysis and preparation of the manuscript, the second author led the design and implementation of the benchmarking tasks described herein.

Calibrating Assessment Literacy Through Benchmarking Tasks

In calibration tasks students assess exemplar texts using criteria against which their own work will be assessed. Typically these tasks are used in the context of training for peer assessment. Little research has been conducted on the benefits of calibration tasks, such as benchmarking, as learning opportunities in their own right. This paper examines a dataset from a long-running benchmarking task (~500 students per semester, for four semesters). We investigate the relationship of benchmarking performance to other student outcomes, including ability to self-assess accurately. We show that students who complete the benchmarking perform better, that there is a relationship between benchmarking performance and self-assessment performance, and that students appreciate the support for learning that benchmarking tasks provide. We discuss implications for teaching and learning flagging the potential of calibration tasks as an under-explored tool.

Keywords: peer and self assessment; feedback; assessment literacy; benchmarking; calibration tasks;

Introduction: Feedback for Learning

Feedback is fundamental to students' learning (Hattie and Timperley 2007). However, the provision of targeted, timely, and actionable feedback is challenging given constrained resources, diverse student needs, and growing enrolments. As a result, a number of pedagogically motivated strategies to provide students with feedback have been investigated. This paper introduces and analyses a naturalistic dataset regarding a particular strategy – calibration tasks – that has been understudied as a standalone task design with the potential to support learning.

The Effects of Peer and Self Assessment

One method for providing feedback is via peer and self-assessments; exercises that develop student's critical evaluation and reflection skills through engaging them in applying assessment criteria to their own or their peer's work. Peer-assessment – the evaluation of a peer’s work for formative or summative purposes – is a pedagogic strategy supported by the literature (see, for overviews, Strijbos and Sluijsmans 2010; Topping 1998). Across this research, both benefits and challenges in effective use of peer-assessment are highlighted, flagging the need for well-designed pedagogic strategies in deploying peer assessment. Importantly, though, evidence indicates that peer-assessment activities can provide ratings as reliably as an instructor (K. Cho, Schunn, and Wilson 2006).

The pedagogic benefit of peer-assessment appears to arise particular from the activity of students *giving* (rather than receiving) feedback. Evidence for this benefit has been observed across qualitative analyses of feedback and focus group data (Nicol, Thomson, and Breslin 2014), experimental comparison of peer-commenting versus peer-reading (without comment) (K. Cho and MacArthur 2011), and comparison of solely receiving versus solely giving feedback (Lundstrom and Baker 2009). Moreover, emphasizing the benefits of peer-review for reviewers, Wooley, Was, Schunn, and Dalton (2008) found that students who gave written feedback and a numerical rating, versus those who gave only a numeric rating, performed better in their own writing. As Cho and Cho note, it appears that "reviewers learn by explaining what makes peer texts good or bad, by identifying problems that exist in those peer texts, and then in devising ways in which those problems can be solved" (2011, 630). Indeed, a similar effect has been proposed for self-assessment (in which students assess their own work), with research suggesting that students become more aligned with tutor-judgements of their own work over multiple semesters (Boud, Lawson, and Thompson 2013; Carroll 2013). As we discuss further below, one strategy to support students in their understanding and provision of high quality feedback has been to design tasks to ‘calibrate’ student’s judgement against that of an instructor. The contribution of this paper is to investigate the potential of these calibration tasks to provide feedback, and their relationship to other learning outcomes, and thereby to contribute to the – as yet, limited – body of knowledge regarding the use of such tasks to support student learning.

Calibration Exercises as Feedback

In calibration tasks, students assess a range of exemplar illustrative assignments, rather than peer's work, typically as a training stage prior to peer assessment. This gives novice writers and reviewers the opportunity to learn about the criteria and to apply them on a common piece of work. Thus, there is potential for calibration tasks to develop students' capacity to give feedback. Indeed, independent of peer assessment, this capacity may develop students' understanding of the purposes of assessment (their assessment literacy), and develop both their evaluative judgement of other's work, *and* increase their capacity to critically assess – and thus improve – their own work (Boud 2000). 'Calibration' tasks may bring about many of the same benefits of peer assessment, while avoiding concerns regarding the time required for peer assessment, and the equitableness of quality peer-feedback provision.

Thus, in developing pedagogic models to improve students' ability to effectively apply assessment criteria to improve the quality of their work, we can consider two types of 'calibration' model (Song et al. 2016). Song et al.,'s analysis, conceptualized calibration in terms of training for peer-assessment, with (1) a 'stand-alone' peer-review pre-task targeting review reliability, compared to (2) a 'mixed' approach in which students calibrated against real-peer work that had also been marked by a tutor in order to flag poor/good reviewers. Their analysis of experimental data indicated that calibration improved review quality, with stand-alone calibration being superior to mixed approaches in which the calibration conducted in tandem with peer assessment.

However, to date calibration exercises have typically not been investigated as a means to develop evaluative judgement independent of peer-assessment. The focus of calibration tasks in current research is their potential in training raters for peer review, and filtering poor raters, in order to ensure reliable peer grading. However, a richer model of calibration tasks frames them as a pedagogically valuable activity to support learning in their own right. The value of such an approach arises from students' provision of feedback – as in peer-assessment models – through the application of criteria to exemplar assignments. Moreover, in calibration tasks, this feedback is on a shared set of exemplars. As such, the activity provides a dataset that can be explored by the students and – in large classes – teams of tutors, who can use the comments to understand the qualities and deficits identified, and to illustrate and build capacity in the provision of quality written feedback. Indeed, such activities that engage students in interacting with each other's reviews may lead to learning gains. One method that has been used is to ask reviewers to read other reviews, and create a summary of that feedback (Goldin and Ashley 2010). This method combines both peer-assessment, and the use of exemplars of feedback to support learning, producing significantly clearer reviews than the original review submissions (although the impact on the learning of reviewers and reviewed was not assessed) (Goldin and Ashley 2010).

Peer assessment learning designs are well researched and understood. However, the role of calibration tasks as an independent activity has not been well investigated, despite their potential to match the benefits of peer assessment. Indeed, calibration tasks also hold a number of benefits over peer assessment, namely: all students can be exposed to exemplars that exhibit important features students should attend to (such as diversity of quality), rather than the more ad-hoc selection in peer assessment; feedback on misalignments between student and instructor numeric-judgements of the exemplars can be more readily provided to both students and instructors; and because a limited set of exemplars are assessed, their features (for example, particular concepts or phrases) can be readily identified in the student feedback, providing a mechanism to give feedback on the quality of student feedback, and more general diagnostic data into feedback appropriateness. Given the task's potential both to provide the gains seen in peer-assessment, and to support the calibration of both tutors and students in developing their evaluative judgement, research is required to understand the impact of these tasks.

The Context of This Study: Benchmarking Tasks

Calibration tasks have the potential to support students in their feedback (on their own and other's work), and assist in training tutors to provide high quality feedback. Moreover, given a restricted domain – the analysis of feedback on known exemplar texts – calibration tasks afford deeper discussion of the qualities of the shared texts being assessed, thus building further student and tutor ability to make such evaluations. To support these features, we have developed a calibration model that we call 'benchmarking', supported by a software tool, SPARKPlus (Willey and Gardner 2008). In contrast to simple calibration models, in which students complete calibration-as-training prior to peer assessment, in our benchmarking model, students:

(1) assess a set of preselected exemplars of varying quality (a Pass, Credit, and Distinction or High Distinction), against a criterion-based rubric, prior to their own writing task;

(2) receive feedback on the quality of their feedback, and their judgement of the exemplars provided compared to that of the lead-academic for the unit;

(3) undertake some reflective action to calibrate their judgement in support of writing their own task;

(4) at the time of submitting their own assignment, engage the same assessment skills in a reflective self-assessment exercise.

This approach builds on the use of exemplars, which have a demonstrated impact on student outcomes with a series of interview and survey studies by Hendry and colleagues indicating that: students appreciate marking exemplars together; student interaction with assessments standards and their application to exemplars is associated with improved student outcomes (Hendry, Armstrong, and Bromberger 2012); and that these marking sessions are well liked by students, who suggest the sessions help them understand expectations and improve the quality of their work (Hendry and Jukic 2014). Related research has indicated that even in the absence of exemplars, the provision and discussion of assessment criteria in advance leads to improved grades (Payne and Brown 2011). Indeed, in a novel application of a calibration assessment task, Wimshurst and Manning (2013) compared two cohorts of students who had, and had not, been asked to assess texts that had been previously graded by an instructor. That comparison found that 7% of the variance in final mark could be attributed to being in the calibration year's cohort. Moreover, their qualitative analysis of feedback comments, and feedback on the task, indicated that the task had supported students in developing their understanding of their assignment, and how they might go about creating a coherent and integrated piece of writing.

Peer assessment and 'mixed' approach calibration exercises can be seen as a special case of exposure to exemplars in which the application of the assessment criteria is targeted at authentic and diverse student texts. Further work is needed to understand the design of these tasks to develop effective pedagogic models and a research agenda that can test differences between various task configurations – as has been conducted in the peer assessment literature. When students provide feedback, they learn. Therefore, there is a need to develop tasks that can provide effective support to them in learning how to give feedback, through exposure to exemplars and feedback on the quality of their own feedback.

The aims of this study, then, are (1) to investigate the relationships between the benchmarking task, student self-assessment, and learning outcomes, and (2) to investigate the diagnostic potential of data arising from the benchmarking task. In order to do this, the specific pedagogic-aims described above will be investigated, using a four-year authentic dataset from a large first-year undergraduate life sciences module (a single unit of study, class, or course).

Methodology

The Context

The dataset analysed in this paper was obtained over four teaching sessions between 2012 and 2015, from a large first year life sciences module *Biocomplexity,* which has a typical enrolment of roughly 500 students. The module is taught in an Australian metropolitan institution, with work assessed using the following grading scheme: Fail: 0-49; Pass: 50-64; Credit: 65-74; Distinction: 75-84; High Distinction: 85-100. In this module, students develop both their evolutionary biology and ecology knowledge, and their academic and communication skills within that context. A key assignment for the module is a written report (worth 40% of the total grade for the module), which follows the structure of a typical scientific report, and is assessed against the following six criteria, with a further 10% separately allocated to a lab notebook submission:

(1) Comprehension, knowledge and synthesis: demonstrated understanding of ecosystem ecology and an ability to place own data in this context (17%)

(2) Methodology and data handling: detailed attention to methodological process and presentation of data (10%)

(3) Layout and protocol: instructions on report compilation followed correctly (15%)

(4) Referencing: quality and use of references is of a high standard (15%)

(5) Writing: demonstrated ability to communicate clearly to a scientific audience (17%)

(6) Scientific enquiry: reasoned question/s posed and pursued in context (16%)

To support student understanding of these criteria, an assessment structure has been implemented which includes students undertaking a benchmarking task as described above, specifically, over a 12 week semester:

(1) In week 4 students assess three exemplars, selected to display the range of grades both overall and on specific target criteria. The exemplars used in the task were obtained from previous students' reports from 2012 and 2013, for which permission from the authors was obtained. The use of previous students' work provides novice writers, in this case first year students, with authentic texts to review. The students are specifically provided with the discussion and references sections, and asked to grade the exemplars using criteria four and five (see above), and to give written feedback against these criteria, using the SPARKPlus tool.

(2) In week 5 the benchmarking results are released, with the SPARKPlus system showing students: (1) the instructor grades and written feedback for the criteria they have assessed; (2) the average and range of grades given by the student cohort for the criteria; (3) all of the written feedback given by the cohort for each of the exemplars. These results give insight to the students and teaching-

team into how the exemplars were assessed and their respective strengths and weaknesses. The instructor feedback (grades and written comments) and its comparison to the student provided feedback are used in class to discuss the criteria, and how to use the feedback to understand the expectations of the assignment (in week 6).

(3) In week 9 students submit their own assignment, and self-assessments of their reports against the criteria; students who do not complete the self-assessment task are assessed out of a maximum of 35 points rather than 40 points.

(4) In week 11, students receive the grades for their assignment and written feedback on the criteria most in need of improvement. Students are given the opportunity to resubmit their report to improve their grade (by 10%), only if they have completed the week 4 benchmarking assignment. Final assignment grades are released at the end of semester and before the final exam (typically week 12).

As described above, the task is intended to, (1) engage students with the assessment criteria and their application; (2) expose students to exemplars of varying quality, and the evaluation of these exemplars; and (3) provide diagnostic information to the students and teaching team regarding the calibration of their evaluative judgement against the assessment criteria. In addition to these benefits, the students are made aware that the tutor team (of approximately 20 staff) also undertakes the benchmarking activity, grading the same exemplars against all six of the criteria. Students are asked to grade against two criteria only in order to limit the load on them, while still engaging them with the application of the criteria, the range of criteria (and their distinct contributions to the grade), and the range of assignment qualities to which they apply the criteria. Using two specified sections also reduces the risk of academic integrity breaches (notably, plagiarism). This shared experience inducts students into an important professional practice, and re-assures them that although their assignments may be graded by different people, significant work goes into ensuring that there is a shared understanding of the assessment criteria across both teaching team and student cohort.

Analysis and Research Questions

The tasks provide a rich set of data, for which ethics approval was granted (HREC-ETH15-0078) to analyse the historic data (i.e., no data of students currently undertaking the target module was analysed). For each student the following data was obtained, from which two measures were calculated for the purposes of analysis:

(1) Whether the student completed the benchmarking and self-assessment tasks (True/False);

(2) The feedback the student gave, both quantitative (for two target criteria, grades, assigned using a slider that maps position to a mark) and written comments (against the two criteria separately) on the benchmarking task;

(3) Student accuracy on the benchmarking and self-assessment tasks (the difference between the grade-mark they assigned, and that of the instructor);

(4) And student criterion-level grade-marks on their submitted assignments.

Distance scores: First, to understand calibration effects, a ‘distance’ measure was calculated for both the benchmarking task and self-assessments. The distance represents the difference between the tutor assigned mark (i.e., the benchmark, for an assignment criterion mark) and the student mark; as such, it can be seen as a proxy accuracy measure where positive scores indicate the student gave a mark lower than that of the instructor (they under-marked), and negative scores that they were higher (they over-marked). We hypothesised that students who are more accurate on the benchmarking task – i.e., are better calibrated – would achieve a higher grade on their own assignment, as well as have smaller distance scores on their self-assessments (i.e., their assessment of their own report would be closer to the tutor assigned marks). That is, we hypothesised that being able to identify the qualities of a submission, would be related to better ability to identify and develop those qualities in one's own work. For the benchmarking task, the distance score is the average distance over the three benchmarked exemplars for both criteria assessed.

Consistency measure: Second, in order to give an indication of the variability of these distances, the standard deviation of the distances was calculated. This measure provides insight into whether students show internal consistency (low standard deviations across distances), indicating that they can apply judgement across criteria and assignments, but – as shown by their distance scores – need to calibrate up or down, or whether they are internally inconsistent (higher standard deviations), in which case alternative interventions may be required.

In addition, a survey based evaluation was conducted on the feedback supports provided to students in 2012 and 2013 with the following seven questions (on a 1-5 Likert scale), and two open ended questions:

(1) The SPARK benchmarking process (week 4) helped me to engage early with the report assessment criteria

(2) The report assessment criteria helped me to understand what was expected in my report

(3) I followed the assessment criteria closely when writing my report

(4) I understood how each assessment criterion contributed to a particular Graduate Attribute

(5) Self-assessing my report helped me to critically evaluated my own academic performance in this task

(6) I have a better understanding of why scientific writing skills are important for a scientific career

(7) Overall I was satisfied with the report-writing learning process

(8) What was most useful about the report-writing learning process for you?

(9) If you could make improvements to the process, what would they be?

The survey was deployed as an optional activity in class, with feedback collected via an anonymous survey.

These datasets were analysed to address the following questions:

(1) How accurate are students in their self-assessments, and what is the relationship of this to their grade?

(2) Do students who complete the benchmarking task perform better in their assignment than those who do not?

(3) Is accuracy on the benchmarking predictive of final assignment mark?

(4) Are students who complete the benchmarking significantly more accurate in their self-assessment of their own assignment?

(5) Is accuracy on the benchmarking related to self-assessment accuracy?

(6) What are student perceptions of feedback structures to support their assignment completion?

Results

Q1: How accurate are students in their self-assessments, and what is the relationship of this to their grade?

As Figure 1 indicates, although the distribution of student self-assessments is generally consistent with staff-assessments, overall the students self-assessed higher than staff assessments, with a larger number of self-assessments falling into the ‘distinction’ boundary for overall grade on the assignment.

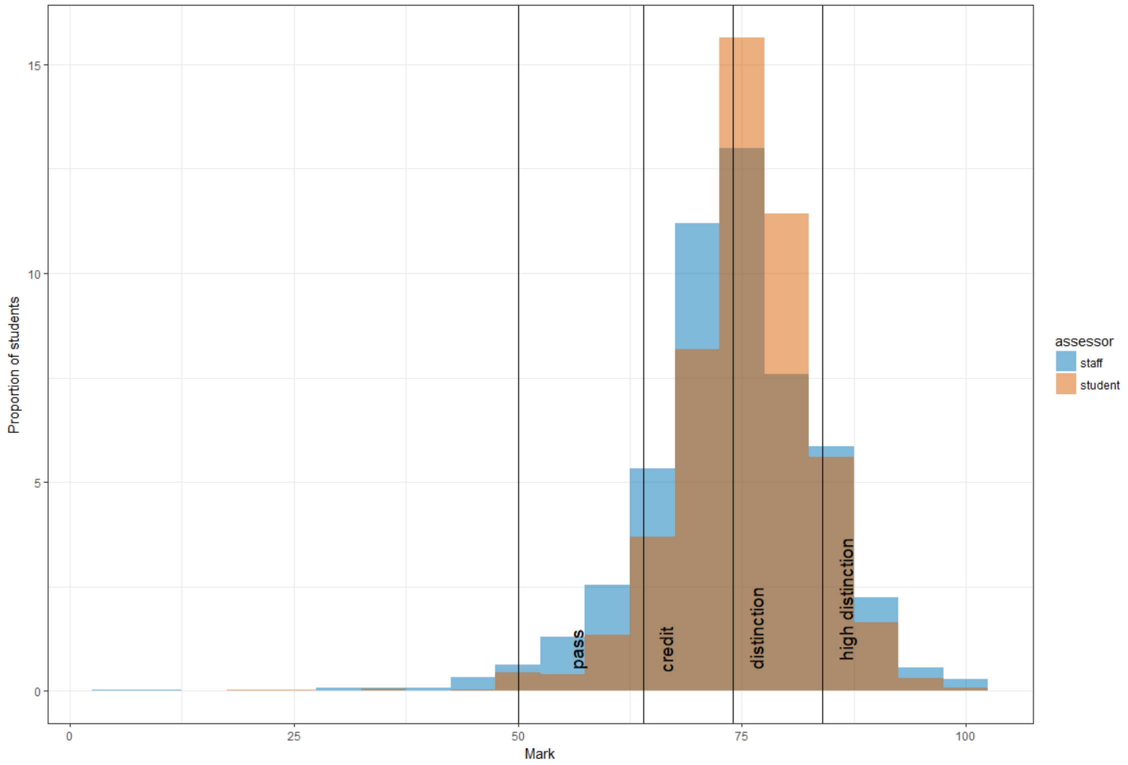

Figure 1 - Histogram comparing the distribution of staff marks to student self-assessments

Self-assessment mark was significantly related to higher final report mark, indicating that students are broadly accurate in their self-assessment, as shown in Figure 2; although this was not a 1:1 – i.e. perfect relationship – which the figure indicates with the black line. As the line of best fit (in red) shows, compared to that 1:1 relationship, there is a clear gap between self-assessment marks and the staff marks. That this relationship is markedly shallower than a 1:1 slope, indicates a small trend towards students with lower marks tending to overestimate their performance, and students with higher marks underestimating their performance.

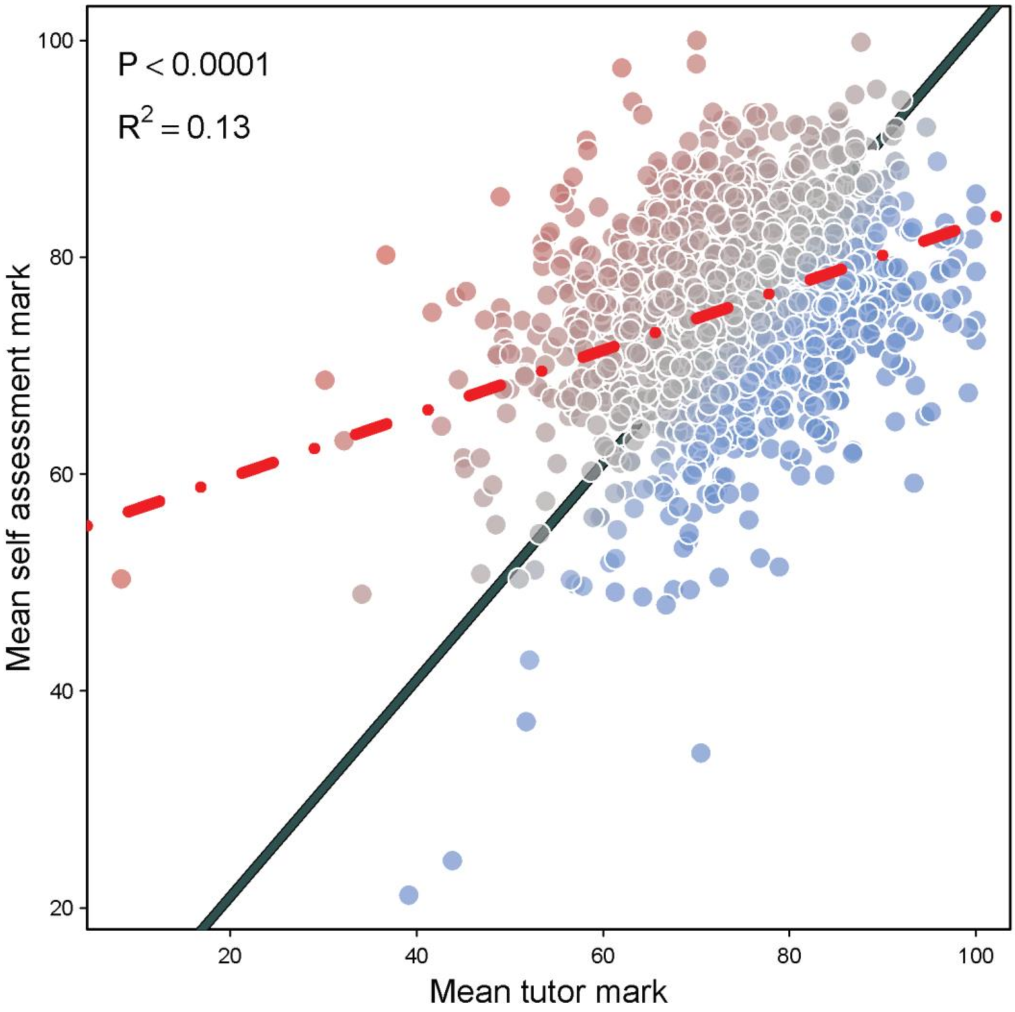

Figure 2 - Relationship between tutor and self-assessment marks; black line indicates 1:1 (i.e., if the tutor and self-assessments were identical), the red line shows the line of best fit indicating a high intercept and low gradient for self- assessments, implying students typically overestimate their own grades, but that this is more pronounced for lower grades than higher.

By looking at the standard deviation of marks over criteria, we can explore how variable students were in their success on the criteria, and their own judgements of this variability. Notably, there was a significant but small relationship between variability (standard deviation) across criterion-level self-assessments and variability in the tutor-marks; *r*(2002) = .12, *p* < .0001; that is, more variable marks across criteria as assessed by tutors are associated with more variable marks across criteria in the self-assessments.

This small relationship indicates that although student self-assessments reflect variability in criterion success (i.e., they do not grade themselves the same across all criteria), their judgement of how variably they perform across criteria has only a small relationship to tutor-assessments of that cross-criterion variability.

Indeed, an analysis of the distances between self-assessment and tutor marks indicates a strong positive relationship between distance and mark, such that students who overestimate their mark (i.e., have negative distances) are more likely to have lower overall marks, while students who underestimate their mark significantly (i.e. have positive distances), are more likely to have higher overall marks; *r*(2012) = .68, *p* < .0001.

Q2: Do students who complete the benchmarking perform better in their assessment than those who do not?

To assess this question, the marks of students who did and did not complete the benchmarking, all of whom still submitted a final report, were compared, with a significantly smaller group in the latter cohort. There was a significant difference in the overall marks of students such that those who completed the benchmarking task scored higher (M = 74.14, SD = 9.28, N = 1979), compared to those who did not (M = 68.24, SD = 12.16, N = 129); *t*(137.88) = 5.41, *p* < .0001. *d* = 0.62[1]. In addition, students who completed the benchmarking task had significantly lower mark variability among criteria (computed by calculating the standard deviation of their marks across the criteria) (M = 5.54, SD = 3.28, N = 1972), than those who did not (M = 6.73, SD = 5.59, N = 129); *t*(133.82) = 2.41, *p* = .01734, *d* = 0.20. That is, students who did not complete

[1] *d* (or Cohen's *d*) is an effect size measure representing the difference between the two group means divided by the average of their standard deviations, thus a *d* of 1 represents that the two groups differ by 1 SD, .5 by half an SD, etc., with .2 considered small, .5 medium, and .8 large.

the benchmarking task performed significantly poorer overall, and achieved less consistent marks across the criteria, implying a poorer ability to calibrate against these criteria.

Q3: Is accuracy on the benchmarking predictive of final mark?

There was no significant relationship between the benchmarking distance scores and student final marks, *r*(1896) = .03, *p* = .23, this was true across comparisons, with very low, non-significant, relationships identified, as indicated in Table 1.

Table 1 Relationships between benchmarking distances and student outcomes

	r (*df* = 1896)	*p*
Benchmarking distance and final marks	.0273	.23
Benchmarking distance and writing mark	.0003	.99
Benchmarking distance and referencing mark	.0396	.08^
Benchmarking writing distance and writing mark	.0254	.27
Benchmarking referencing distance and referencing mark	.0297	.20

Q4: Are students who complete the benchmarking significantly more accurate in their self-assessment

To assess this question, the self-assessment distances of students who did and did not complete the benchmarking were compared, with a significantly smaller group in the latter cohort. There was a significant medium effect difference in the average distance scores of students, such that those who completed the benchmarking task had lower distances (M = 1.15, SD = 16.13, N = 1979), compared to those who did not (M = 8.13, SD = 27.47, N = 129); *t*(133.84) = 3.00, *p* = .0032. *d* = 0.62.

Q5: Is accuracy on the benchmarking related to self-assessment accuracy?

There was a small significant relationship between the benchmarking distance scores and student self-assessment distances, *r*(1887) = .10, *p* < .0001. That is, (in)accuracy in the benchmarking task and (in)accuracy in self-assessment are related, such that those who were more inaccurate in the benchmarking were also more inaccurate in their self-assessment. This was also true for the relationship between benchmarking distance scores and self-assessment distances within the two benchmarking criteria: writing, *r*(1887) = .06, *p* = .0145, and referencing, *r*(1887) = .06, *p* = .0061.

Q6: What are student perceptions of feedback structures to support their assignment completion?

In 2012 n = 210 students completed the feedback survey, from a total cohort of 482, with n = 220 of 511 in 2013. The feedback from these cohorts was generally positive (>75% agree or strongly agree on all questions), with the distributions for each question indicated in .

Table 2.

Table 2 Student Feedback Survey*

	Year	Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly agree
	%	%	%	%	%	%
The SPARK benchmarking process (week 4) helped me to engage early with the report assessment criteria	2012 2013	1.43 0.47	3.81 6.16	12.38 13.74	62.85 60.19	19.52 19.43
The report assessment criteria helped me to understand what was expected in my report	2012 2013	0.95 0.00	1.90 1.82	4.76 8.18	56.19 54.09	36.19 35.91
I followed the assessment criteria closely when writing my report	2012 2013	0.95 0.91	2.86 3.18	9.52 17.27	57.14 60.00	29.52 18.64
I understood how each assessment criterion contributed to a particular Graduate Attribute	2012 2013	0.48 0.00	1.91 3.18	21.53 22.72	61.24 62.27	14.83 11.82
Self-assessing my report helped me to	2012 2013	1.44 1.36	7.67 6.36	21.05 20.91	46.89 49.09	22.97 22.27

critically evaluated my own academic performance in this task						
I have a better understanding of why scientific writing skills are important for a scientific career	2012	0.96	4.31	11.00	57.89	25.84
	2013	0.45	1.81	14.55	60.45	22.73
Overall I was satisfied with the report-writing learning process	2012	0.96	0.96	7.18	68.42	22.49
	2013	0.91	0.91	13.81	16.27	22.72

*Note the small discrepancies in sample size are due to some students omitting answers on some questions

Of the 430 respondents, 312 responded to the question “What was most useful about the report-writing learning process for you?”, and 212 to the question “If you could make improvements to the process, what would they be?”. These comments generally had a low number of characters in them (i.e. they were short, M = 46.91, SD = 49.58 for the first question, and M = 32.38, SD = 59.04 for the second). These comments generally made reference to one or more of the activities referred to in the questions. Of these responses, 26 comments explicitly referred to the benchmarking or SPARK as the most useful learning feature, and 8 responses in the improvements. Of these improvements, only 2 suggested that the exercise was not useful, with the others indicating other suggestions (regarding timing, instructions, or other features of the task). One student indeed said:

> *Benchmarking helped me to understand what level of writing was expected for each grade. The feedback and re-submission really helped me to better my writing and to understand how I could improve*

A surface analysis of comments not explicitly mentioning the benchmarking also indicated broader reference to the task, saying for example:

> *That it forced me to be familiar with the marking criteria BEFORE writing the assignment. Usually I look at the criteria after writing the assignment and seeing whether it met the criteria, but with this method I made sure to incorporate the points whilst writing.*

And

> *Having previous reports to look at and gain understanding how to write and what the markers are looking for.*

Discussion & Conclusion

While peer and self-assessment have received considerable research and teaching attention, calibration exercises such as the benchmarking approach investigated in this paper have not been well explored. This paper demonstrates a number of key findings.

First, there is a need for calibration tasks. Students display a significant gap in the accuracy of their self-assessments (Figure 1), with students who overestimate their performance more likely to have lower grades, and students who underestimate their performance more likely to produce better work (Figure 2). Students vary in their ability to judge quality work across criteria, and this is reflected in their grades across criteria as indicated by relationships between the standard deviation of criterion-marks and distances on those marks. Exercises like the benchmarking task may be important in addressing these gaps in student evaluative judgement.

Second, data from benchmarking tasks may provide diagnostic information regarding student performance, and impact on learning. At a crude level, we have shown (in RQ2) that students who complete the task do better overall. This finding alone should be treated with caution, as of course this may be an effect of student motivation or various other performance factors unrelated to the benchmarking itself. Importantly, our findings indicate no clear relationship between accuracy on the benchmarking task and final mark (RQ3). The challenges of identifying such relationships are clear; if the benchmarking task impacts on learning (i.e., it does indeed have a calibration effect), then performance on the benchmarking should support change in final outcome, rather than being predictive of that final outcome.

Third, then, the benchmarking data may provide insight into self-assessment performance, and that performance is important given – as discussed above – the relationship between self-assessment distances and performance. Again, on a crude level, students who completed the tasks were better self-assessors than those who did not (RQ4). As above, although this is an important finding, the sample size is very small and there are many possible explanations for this finding, most of which are beyond the scope of this paper. More importantly, though, there was a small relationship between the benchmarking distances and self-assessment distances (RQ5). That is, underestimating the quality of the benchmarked works (higher distance scores) was associated with underestimating the quality of one's own work, with lower scores – i.e., overestimation – associated with the converse. This suggests that the benchmarking and self-assessment tasks may tap into similar skills, skills that we know are important for improving student outcomes. Therefore, it may be possible to develop data from the benchmarking tasks to provide diagnostic insight to both students and instructors. For example, interventions could target different kinds of support to students who under or over-mark in a benchmarking task, to support them in improving their own work.

These findings are an important step in understanding benchmarking and calibration tasks as significant pedagogic tools in their own right, independent of training for peer assessment. As RQ6 indicates, students see the value of these approaches, and appreciate the targeted feedback to support their learning. However more work is required to understand the impact of calibration on student learning. The analysis reported here is from a single module, with a limited set of benchmark texts, criteria, and student cohort. Further work should be conducted to understand how interventions – for example, to target calibration support for over, under, and inconsistent raters – might impact student outcomes. While test-retest models, in which students undertake similar tasks twice to investigate change in accuracy, may be desirable in experimental settings, such work is challenging in authentic time-stretched classroom contexts. A significant contribution of this work is in its analysis of a practical, authentic, pedagogic intervention. Further work may also explore the richer kinds of feedback data, including the written comments that students provide.

In this paper we have taken a principled approach to the evaluation of a teaching and learning innovation, to understand how the impact of the approach on learning and how it might be further developed to gain insight on, and support, that learning. As outlined above, there are challenges in evaluating this kind of authentic data. However, it is crucial that our understanding of these kinds of authentic practice is developed, because in so doing researchers have the potential to understand, and improve, existing practices without requiring the uptake of novel forms of assessment by lecturers. Students learn through providing feedback, and by engaging with assignment exemplars. Calibration exercises such as benchmarking tasks have potential to support development of evaluative judgement towards self-assessment and improved performance on assessment criteria, and to provide diagnostic information to both instructors and students regarding this development.

Acknowledgements

The authors gratefully acknowledge support of the University of Technology Sydney, and the DVC(education) Professor Shirley Alexander in implementing and researching the benchmarking tasks, including through funding via UTS Vice-Chancellor's

Learning and Teaching grants. Our thanks to Simon Housego who was instrumental in developing the benchmarking tasks and their evaluation. Thanks too to the tutors, and students.

References

Boud, David. 2000. “Sustainable Assessment: Rethinking Assessment for the Learning Society.” *Studies in Continuing Education* 22 (2): 151–167. doi:10.1080/713695728.

Boud, David, Romy Lawson, and Darrall G. Thompson. 2013. “Does Student Engagement in Self-Assessment Calibrate Their Judgement over Time?” *Assessment & Evaluation in Higher Education* 38 (8): 941–956. http://www.tandfonline.com/doi/abs/10.1080/02602938.2013.769198.

Carroll, Danny. 2013. “Benefits for Students from Achieving Accuracy in Criteria-Based Self-Assessment.” In . Sydney. https://www.researchgate.net/profile/Danny_Carroll/publication/264041914_Benefits_for_students_from_achieving_accuracy_in_criteria-based_self-_assessment/links/0a85e53c9f80a21617000000.pdf.

Cho, Kwangsu, and Charles MacArthur. 2011. “Learning by Reviewing.” *Journal of Educational Psychology* 103 (1): 73. http://psycnet.apa.org/journals/edu/103/1/73/.

Cho, Kwangsu, Christian Schunn, and Roy W. Wilson. 2006. “Validity and Reliability of Scaffolded Peer Assessment of Writing from Instructor and Student Perspectives.” *Journal of Educational Psychology* 98 (4): 891. doi:10.1037/0022-0663.98.4.891.

Cho, Young Hoan, and Kwangsu Cho. 2011. “Peer Reviewers Learn from Giving Comments.” *Instructional Science* 39 (5): 629–643. doi:10.1007/s11251-010-9146-1.

Goldin, Ilya, and Kevin. Ashley. 2010. “Learning by Reviewing through Peer Feedback Refinement.” In *Proceedings of the Workshop on Computer-Supported Peer Review in Education*. http://www. cspred. org/proceedings/cspred-2010-proceedings. pdf.

Hattie, John, and Helen Timperley. 2007. “The Power of Feedback.” *Review of Educational Research* 77 (1): 81–112. http://rer.sagepub.com/content/77/1/81.short.

Hendry, Graham D., Susan Armstrong, and Nikki Bromberger. 2012. “Implementing Standards-Based Assessment Effectively: Incorporating Discussion of Exemplars into Classroom Teaching.” *Assessment & Evaluation in Higher Education* 37 (2): 149–161. http://www.tandfonline.com/doi/abs/10.1080/02602938.2010.515014.

Hendry, Graham D., and Katherine Jukic. 2014. “Learning about the Quality of Work That Teachers Expect: Students’ Perceptions of Exemplar Marking versus Teacher Explanation.” *Journal of University Teaching and Learning Practice* 11 (2): 5. http://eric.ed.gov/?id=EJ1040741.

Lundstrom, Kristi, and Wendy Baker. 2009. “To Give Is Better than to Receive: The Benefits of Peer Review to the Reviewer’s Own Writing.” *Journal of Second*

Language Writing 18 (1): 30–43. http://www.sciencedirect.com/science/article/pii/S1060374308000313.

Nicol, David, Avril Thomson, and Caroline Breslin. 2014. “Rethinking Feedback Practices in Higher Education: A Peer Review Perspective.” *Assessment & Evaluation in Higher Education* 39 (1): 102–122. http://www.tandfonline.com/doi/abs/10.1080/02602938.2013.795518.

Payne, Emma, and George Brown. 2011. “Communication and Practice with Examination Criteria. Does This Influence Performance in Examinations?” *Assessment & Evaluation in Higher Education* 36 (6): 619–626. http://www.tandfonline.com/doi/abs/10.1080/02602931003632373.

Song, Yang, E. F. Gehringer, J. Morris, J. Kid, and S. Ringleb. 2016. “Toward Better Training in Peer Assessment: Does Calibration Help?” In . http://ceur-ws.org/Vol-1633/ws1-paper10.pdf.

Strijbos, Jan-Willem, and Dominique Sluijsmans. 2010. “Unravelling Peer Assessment: Methodological, Functional, and Conceptual Developments.” *Learning and Instruction* 20 (4): 265–269. doi:10.1016/j.learninstruc.2009.08.002.

Topping, Keith. 1998. “Peer Assessment between Students in Colleges and Universities.” *Review of Educational Research* 68 (3): 249–276. doi:10.3102/00346543068003249.

Willey, Keith, and Anne Gardner. 2008. “Improvements in the Self and Peer Assessment Tool SPARK: Do They Improve Learning Outcomes?” *ATN Assessment* 1 (1). http://ojs.unisa.edu.au/index.php/atna/article/viewFile/343/258.

Wimshurst, Kerry, and Matthew Manning. 2013. “Feed-Forward Assessment, Exemplars and Peer Marking: Evidence of Efficacy.” *Assessment & Evaluation in Higher Education* 38 (4): 451–465. http://www.tandfonline.com/doi/abs/10.1080/02602938.2011.646236.

Wooley, R., C. Was, Christian D. Schunn, and D. Dalton. 2008. “The Effects of Feedback Elaboration on the Giver of Feedback.” In *30th Annual Meeting of the Cognitive Science Society*. http://www.academia.edu/download/35347035/p2375.pdf.