Accepted version of Boud, D., Lawson, R. and Thompson, D. (2015). The calibration of student judgement through self-assessment: disruptive effects of assessment patterns, *Higher Education Research and Development,* 34, 1, 45-59. DOI:10.1080/07294360.2014.934328

The calibration of student judgement through self-assessment: disruptive effects of assessment patterns

David Boud[1], Romy Lawson[2], Darrall G Thompson[1]
[1]University of Technology, Sydney
[2]University of Wollongong

Abstract

Can extended opportunities for self-assessment over time help students develop the capacity to make better judgements about their work? Using evidence gathered through students' voluntary self-assessment of their performance with respect to assessment tasks in two different disciplines at two Australian universities, the paper focuses on the effects of sequences of units of study and the use of different types of assessment task (written, oral, analysis, project) in the development of student judgement. Convergence between student criteria-based gradings of their own performance in units of study and those allocated by tutors was analysed to explore the calibration of students' judgement over time. First it seeks to replicate analyses from an earlier smaller-scale study to confirm that students' judgements can be calibrated through continuing opportunities for self-assessment and feedback. Second, it extends the analysis to coherently designed sequences of units of study and explores the effects of different types of assessment. It finds that disruptive patterns of assessment within a sequence of subjects can reduce convergence between student and tutor judgements.

Keywords – Self-assessment, judgement, assessment, student learning.

There are many attributes that higher education courses seek to develop in students. These are normally well articulated in course aims and learning outcomes. However, there are attributes that are foundational to all discipline-specific outcomes. One of the most central is the ability of students to make effective judgements about their own work. Without such ability students will not be effective learners—*how can they judge whether the study they are doing is appropriate?* —Nor as graduates will they be able to function effectively in the workplace—*how will they be*

able to judge whether what they have done is sufficient to meet the requirements of the task? This idea has been represented in many different ways: as self-regulation (Nicol & Macfarlane-Dick, 2006) and meta-cognition (Knapper & Cropley, 2000). This paper is located itself in the tradition of self-assessment (eg. Boud, 1995), an approach well accepted in higher education.

There have been studies of student self-assessment over many years and considerable advocacy for the effectiveness of practices in which students review their own work (eg. Boud, 1995; Dochy, Segers & Sluijsmans, 1999). It has been well argued that students need to develop the capacity to make judgements about their own work if they are to be effective learners both as students and following graduation (Boud & Falchikov, 2007).

It is commonly assumed that the development of students' ability to make judgements about their own work develops as a residual effect from the normal tasks students undertake as part of their studies, but the systematic facilitation of self-assessment opportunities is rare in most courses. Engaging students in improving their capacity to make good judgements over different tasks is difficult without providing a comparison with criteria-based assessment by expert tutors. Through such a process with appropriate scaffolding, it has been found that students can move progressively towards the kinds of quasi-independent judgements about self-performance needed for effective lifelong learning (Boud & Falchikov, 2007).

Assessment processes tend to be oriented around the needs of ease of certification rather than the effectiveness of provision of feedback about students' judgement (Boud & Molloy, 2013). Problems arise related to the embedding of student self-assessment opportunities in the normal assessment of tasks. There is often little opportunity for feeding back to students criteria-based comparisons of their own and the tutor's judgements, to enable the calibration of their own judgements.

Conventional assessment advice has for many years recommended that a variety of assessment methods be used in order that the wide variety of learning outcomes for any given course are adequately assessed (eg. Fazey & Marton, 2002). While this is clearly an advance on common assessment practices which focus on end of year written examinations, it is important to ensure that the sheer diversity and complexity of many different assessment approaches does not distract students from the

substance of what they are learning and disrupt the development of their judgement.

This paper explores the calibration of undergraduate students' judgements about their own work through an analysis of their criteria-based self-assessments over time in given units of study. It tackles two problems. First, it seeks replicate with new and more substantial data from a different discipline some tentative findings about the improvement of student judgement over time reported in an earlier paper (Boud, Lawson & Thompson, 2013). Secondly, it examines the development of judgement skills in coherent sequences of subjects and analyses these in terms of types of assessment method used and the assessment criteria related to these forms of assessment.

Framing of the study

Unless students are able to make effective judgements about the quality of their own work beyond the end of the program in which they are enrolled, the assessment within that program is not sustainable. That is, the programs' assessment is unable to meet the needs of students in supporting future learning (Boud, 2000). If assessment is sustainable, as students progress towards graduation, there should be an observable convergence between the grades allocated to their work by subject-matter experts and the grades they judge for themselves to be appropriate for the same work as students develop the capacity to make increasingly effective judgements about their own work. If this effect is not observed, then either the assessment tasks or the associated learning activities need to be redesigned so as to not only maintain their summative purpose but also enable students to develop their own judgements.

Sadler suggests that self-evaluative skills need to be developed 'by providing direct authentic evaluative experience for students' (Sadler, 1989, p.119). That is, they involve students making specific judgements about particular work they have completed. Like any form of expertise, these skills require development over time (eg. Ericsson, Krampe & Tesch-Romer, 1993). Single or even multiple examples of self-assessment activity deployed in particular course units, in themselves are likely to have little impact. As students will encounter new domains of knowledge that require new approaches to study, these changes are disruptive for students. It is therefore unlikely that judgement will improve continuously as novel challenges are encountered.

The role of feedback in the development of judgement is important (Boud & Molloy, 2012). Students need to have some means of knowing whether their judgements are accurate and thus be able to calibrate their own judgements in the light of evidence. Through such calibration they can identify the areas in which they need to improve and see shifts in their capacity over time. This evidence is commonly provided by teachers who can provide useful information about whether work meets required standards and if it does not, how it can be improved. However, Sadler suggests that students should develop means of evaluating the quality of their own work through moving beyond 'teacher-supplied feedback to learner self-monitoring'. He proposes that the course in which they operate needs to 'make explicit provision for students themselves to acquire evaluative expertise' (Sadler, 1989, p.143). While feedback information from teachers may be necessary, it is not enough on its own for evaluative expertise to develop.

We suggest that practice in this evaluative expertise can be developed through systematic self-assessment activities involving the integration of criteria-based assessment as a normal aspect of intermittent assessment and feedback. Students need to be actively making judgements about their own work and relating these to the evaluations of others. They would examine the accuracy of their judgements and use the insights of others, whether teachers or others such as peers, to find reasons to explain poor judgements and for ways to improve future judgements. While some students might engage in these activities regardless of course interventions, there may be merit for others in making this process more systematic and formalised.

However, evaluative expertise alone is not sufficient for improvement, as Ramprasad (1983) has argued. Drawing on Ramprasad, Sadler (1989) identified three conditions for effective feedback, that is feedback that influences learning: (1) a knowledge of the standards; (2) having to compare those standards to one's own work; and (3) taking action to close the gap between the two (Sadler, 1989, p.138). Standards not only need to be explicit, but students need to appreciate how these standards can be manifest in work of the kind in which they are engaged. Relating these standards to one's own work needs an ability to see in one's own work the indicators of achievement. Finally, closing the gap requires opportunities for subsequent performance in which this knowledge can be translated into action. These opportunities may be compromised however if assessment tasks change very substantially from occasion to occasion, or if the forms in which they are represented require the learning of new skills of production, or if the combination of new subject matter with new

forms of assessment task leads to disruption and excessive cognitive load (van Merriënboer & Kirschner, 2007). In circumstances such as these the development of evaluative expertise can be delayed or postponed.

Structuring learning to help students understand the criteria and standards required in their learning is vital for students to maintain a realistic perception of their achievements (O'Donovan, Price & Rust, 2008). This is achieved by making assessments and objectives transparent to students, by providing easily understood feedback that relates to the objectives, and by promoting self-awareness in students (Boud, 1990).

One feature of course design that might be required to aid this process is for student judgements to be calibrated against experienced judges of the kind of work being considered. Qualities of work are hard to discern and the availability of the judgements of others with respect to the very criteria needed to judge one's own work could be facilitative. In such situations the discrepancies between the student judgement and that of the expert other are important pointers in raising the students' awareness about what they need to do to improve their work.

The most readily available information for students to help calibrate their judgements is the information they receive as part of the normal marking process for their assignments. While a global grade for an assignment is low in information content, if this can be broken down into explicit descriptive assessment criteria used for a given task and if this can be supplemented by additional qualitative comments, then the information value of the activity can be increased. If in addition to this, it can be arranged for students to make judgements against criteria about their work and for them to view their own judgements compared to the judgements of others after they have done this, then the information value of the activity can be dramatically increased without other changes to course conduct and organisation. However, this requires an ongoing and sustainable engagement involving tutors, academic coordinators as well as the students themselves. The facilitation of this engagement can be achieved through an online criteria-based marking system accessed alternately by markers and students for a wide range of assessment tasks throughout a course of study.

The study

An earlier investigation using data from such an online system was undertaken across units of study over a number of semesters in an undergraduate course in an Australian university (Boud, Lawson &

Thompson, 2013). The findings showed that across multiple tasks within a given course unit, although students initially struggle to accurately self assess, with further self-assessment opportunities and comparing criteria-based grades from their tutor they become more accurate. When this activity was extended through subsequent semesters and other course units, it was found that there was convergence between student and tutor marks for each first task in a new course unit. Students' overall performance had a strong effect on convergence in that while high and medium performing students marks converged with those of tutors over time, those of the poorest performers did not, and their absolute marks failed to improve over time. The findings of general overall agreement between student and tutor marks and the pattern of over-and under-rating by ability were in accord with earlier studies (Falchikov & Boud, 1989; Dochy, Segers & Sluijsmans, 1999). The current paper builds on this previous investigation to examine the research question:

- Does students' continuing engagement with judging grades through voluntary criteria-based self-assessment lead to convergence between teacher-generated grades and student-generated grades?

Within this overall question, data were examined to explore the following sub-questions. They were grouped into two sets. The first of these sought to replicate findings from our earlier study using a different discipline area in a different Australian university. The second extended the investigation to explore whether the organising of units of study in coherent sequences has an impact on student judgment and whether variation in the types of assessment task and related criteria influences the development of student judgement.

Repeat questions:

1. Does accuracy in students' self-assessed grades vary by student performance?
2. Does students' accuracy in self-assessment relate to improved performance?
3. Do students' marks converge with tutors within a unit of study?
4. Does the gap between student grades and tutor grades reduce over time?

New questions:

1. Does the gap between student grades and tutor grades reduce across units of study designed as a sequence?
2. Does mode/type of assessment task (eg. written assignment, project, presentation, etc.) influence students' judgement of grades?
3. Does analysis of criteria that relate to type of assessment task influence students' judgement of grades?

Through the new questions, the paper aims to investigate whether curriculum design that shows the deliberate sequencing of knowledge and skills across study units has an effect on students' ability to calibrate their own judgements or whether assessment methods within these sequences may negate their impact. Areas of consistency across the two programs are identified and differences between them are discussed. In particular, a discrepancy between the two studies led to the more detailed focus on how different types of assessment task might disrupt the development of students' judgement.

Methodology

The ability of students in undertaking self-assessment using a grading scale for criteria to assess different forms of assessment task was taken as a proxy for student self-judgement. The study operated in an authentic context using data from high stakes assessment tasks in which grades from tutors contributed directly to students overall performance. The tasks were not generated as part of the study: all were the normal assessment tasks students were expected to complete. The criteria being used were also those applied to these tasks. The only addition to the normal assessment activity was that students voluntarily made criteria-based judgements about their own performance using a web-based marking system.

Data was obtained through the medium of ReView™, an online assessment system that allows students to grade themselves on a continuous sliding scale that recorded a specific percentage mark in the software's database. The gradings were made with respect to each explicit criterion for each assessment task in a unit of study (as well as an overall grade). Student self-assessment was voluntary. Tutors were not able to view students' grading judgements until they had completed their own grades on the continuous sliding scale in ReView™. The software also stored percentage marks for each of the gradings made by tutors for each assessment criterion and these were compared with students' self-assessments.

Analysis has been undertaken of differences between student and tutor percentage marks for each element of each assessment task for courses over at least two separate semesters with a view to exploring changes within units of study and over time. Arising from our earlier study it was hypothesised that as students engage in making their own judgements, followed by their viewing of tutor's judgements, there would (a) be a

convergence of marks for each iteration of assessment within a module, and (b) be an initial divergence at the beginning of each new unit of study which reduces in size over time.

Context and data used

We used data from students who voluntarily self-assessed in units of study conducted in a Bachelor of Design degree over a five-year period between 2006 and 2010 and a Bachelor degree in Business over a three year period between 2008 and 2011 in another Australian university. These data sets were selected as they were degree programs that had been using the assessment ReView™ system over an extended length of time. The programs were also structured in a scaffolded nature and so progressive subjects could be identified in order to analyse for the effect of sequencing.

The original study involved 182 Design students who had voluntarily self-assessed in three or more course modules over the course of their degree (data was collected from a total of 49 units of study). This second study used a larger data set of 1162 Business students who had self-assessed in more than two progressive course modules over the course of their degree (data was collected from a total of 7 units of study). Three streams in the Business program were examined—Management, Personal Development and Finance. The course modules selected within each represented a progression: Management had three modules over three years which included a core subject in both the first and second year and then a capstone subject in the final year; Personal development and Finance which both had a core first year module and a final year capstone. These units of study were deemed to have been designed in a sequenced fashion that developed related subject content and skills.

The data used in the analysis consisted of student self-assessment gradings for each criterion for the task, recorded in a database as percentage marks, together with teacher / tutor grades for students for each task against each criterion, also recorded as percentage marks. Data was only obtained from those who chose to self-assess. Data on the proportion of students who undertook self-assessment was not available to us at the time of finalising this paper. The implications of missing data are discussed later.

Results

Repeat Q1. Does accuracy in students' self-assessed grades vary by student performance?

Students were divided into three ability groups dependent on their overall mark for a course module. These groups were classified as low ability for those students who achieved fail grades, mid ability for those students gaining a pass or credit grade and high ability for those who fell into the distinction or high distinction grades—these are the common grade bands used across Australian universities. The original data from the design undergraduate degree showed that when students were examined according to their ability level, i.e. high ability, low ability and mid ability it was found that the low ability students significantly overestimated their ability on all their assessment tasks. The high ability group significantly underestimated their grades on all stages of assessment. However, the mid- range group were significantly higher than the tutors at the beginning task but by the end task there was no significant difference between themselves and the tutors. The patterns were similar for the Business data. However, whilst the high performing Business students underestimated their ability on the first task ($t(1, 3137)=12.095$; $p<0.00$) this difference was much smaller by the end task difference ($t(1, 121)=-2.029$; $p<0.045$). This repeat analysis confirmed that ability level did have an effect on students' accuracy of judgement with students from both data sets in the low ability group showing no improvement in their judgements over time. This is an important finding to note as it means that these students are at risk in terms of both their academic performance and their competency to self-assess.

Repeat Q2. Does students' accuracy in self-assessment relate to improved performance?

Students were categorised as over, accurate or under estimators based on the difference between self-assessed marks and tutor marks. Those who were within 4% of the marks of the tutor (either above or below) were classified as accurate; those over 4% higher than the tutors were grouped as over estimators and those more than 4% below the tutors as under estimators. A one-way ANOVA was conducted to look at difference in performance scores in relation to ability to calibrate, i.e. over estimators, under estimators and accurate estimators. The design data showed accurate estimators with a significant increase in scores across all tasks, but the under and over estimators marks did not significantly alter over time.

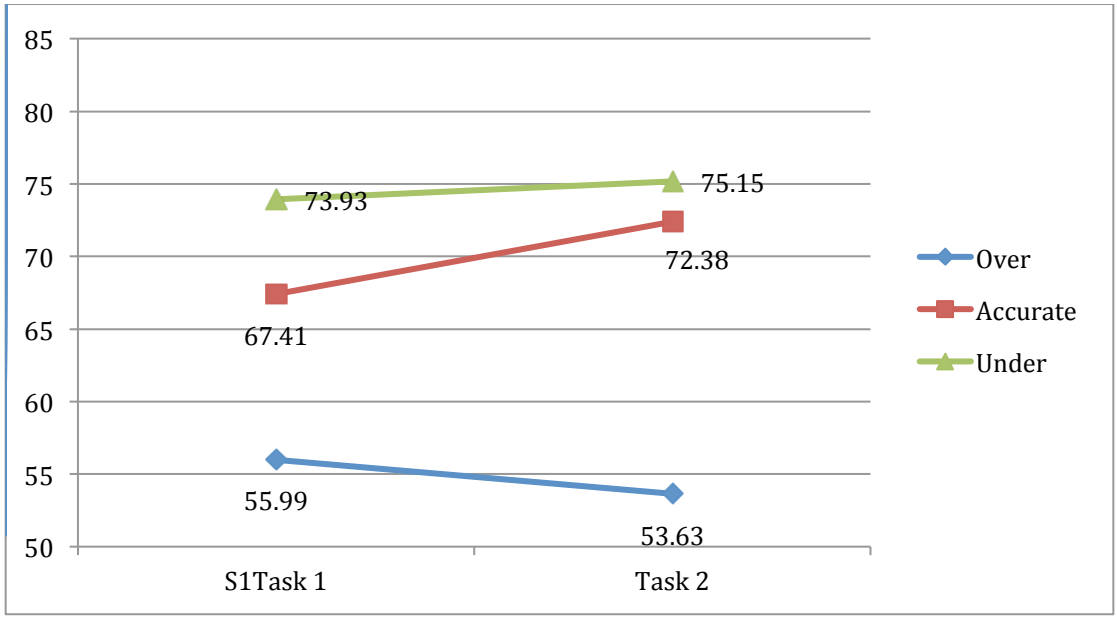

Figure 1. Students' ability to calibrate impact by performance – Design

Again, the Business data showed that accurate estimators improved their performance over time ($F(2,5402)= 12.967$; $p<0.00$), however with this data the under-estimators also improved their performance ($F(2,426)=27.01$; $p<0.00$).

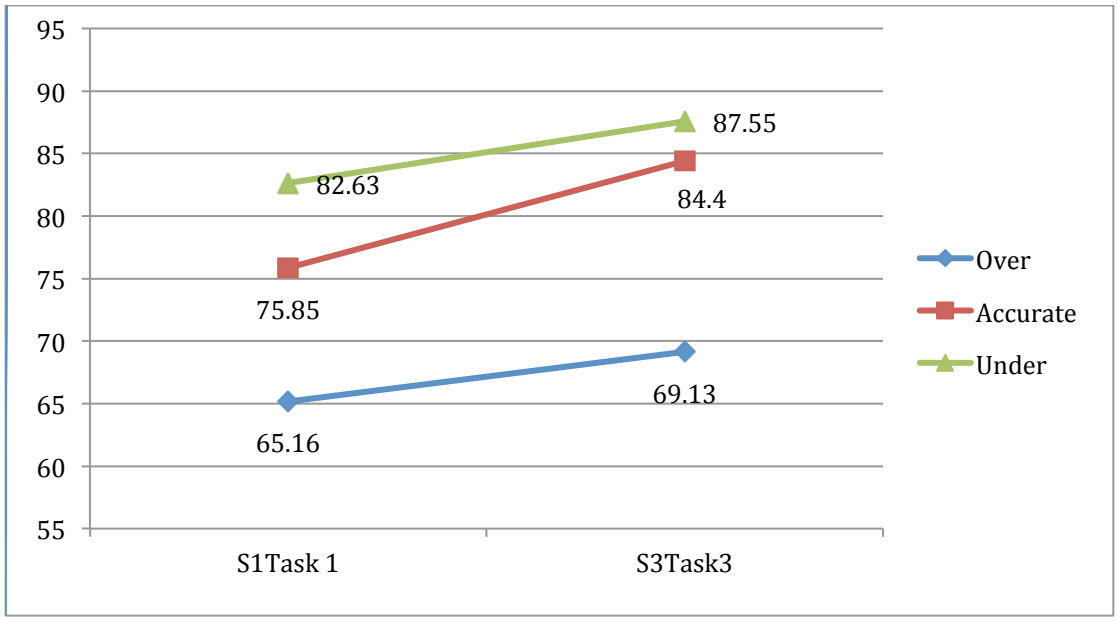

Figure 2. Students' ability to calibrate impact by performance - Business

This again confirms an important aspect of accurate judgements in students, showing that those students who had a good understanding of their ability levels in relation to the expectations of the assessment tasks were able to use this knowledge to manage their learning and show improvement.

Repeat Q3. Do students' marks converge with tutors within a course module?

Using the self-assessment percentages and the tutors marking percentages, a series of paired t tests were conducted to look for overall differences between tutor and students' scores on tasks within a single unit of study. There was a significant difference found between the student and the tutor at the first task, with students rating themselves higher than the tutors. By the second task this significant difference was no longer evident and remained non-significant for third and later tasks where present. The Business data however did not show this convergence. There was a decrease in the difference between staff and students by task two, but not a significant one ($t(1, 5165)=16.360$; $p<0.00$). Then the difference increased by the third task (($t(1, 5826)=34.093$; $p<0.00$).

This result was not as expected and led to our new question 2. Does mode/type of assessment task (eg. written assignment, project, presentation, etc.) influence students' judgement of grades? This was asked on examining the assessment tasks set for students in the design program compared to the business program. On inspection it appeared that the design course modules were designed to use a scaffolded approach to assessment with each task building from the previous, whereas the business degree used a range of different assessment tasks.

Repeat Q4. Does the gap between student grades and tutor grades reduce over time?

Again using percentage marks from self-assessment and tutor marking, a series of paired t tests was undertaken to ascertain whether the difference between students and tutors marks at task 1 decreased with further iterations of self-assessment over semesters. It was found that Design student grades were significantly higher than tutors in the first task of their initial three semesters of self assessing (Semester 1 ($t(1, 909)=8.259$; $p<0.00$); Semester 2 ($t(1, 1170)=3.878$; $p<0.00$); Semester 3 ($t(1, 435)=3.365$; $p<0.01$)). By the fourth semester there was no significant difference between students and tutors ($t(1, 216) =1.956$; $p>0.05$).

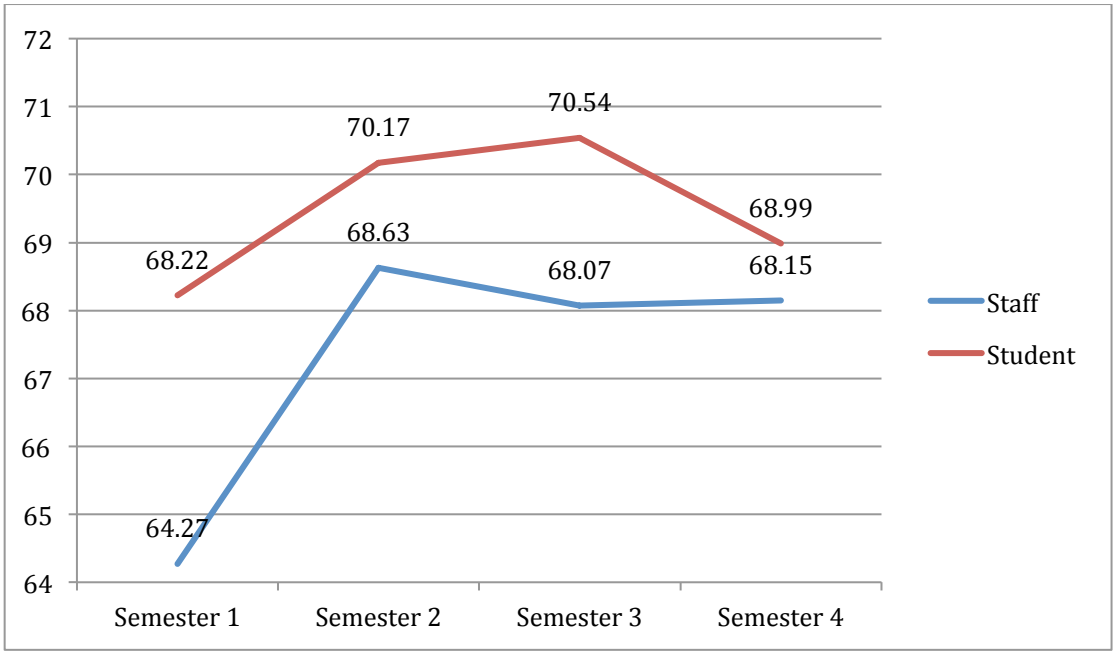

Figure 3. Student and tutor grades for the first task in each module in each semester—Design

This analysis was also conducted for the Business cohort finding that students were significantly higher than tutors in the first task of their initial two semesters (t=(1,6861) 34.069; p=0.00) (t=(1, 1524) 15.552=0.00) of self assessing but by the third year subject this difference was no longer evident (t=(1, 23) 1.157; p=0.259).

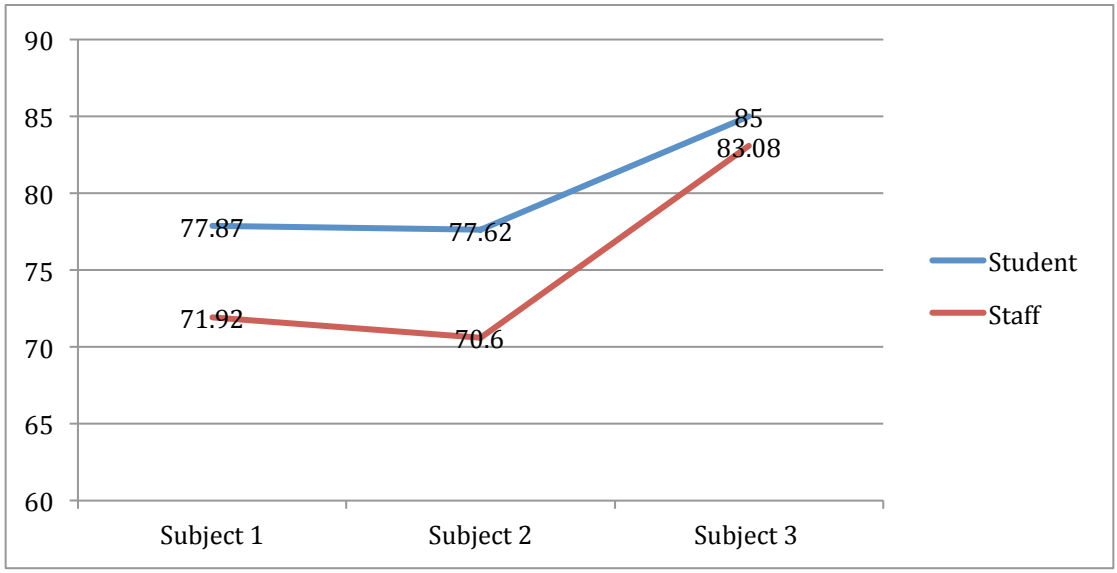

Figure 4. Student and tutor grades for the first task in each module in each semester—Business

This data shows that over time students were becoming more accurate in their judgement of their ability but this improvement was over the

majority of the degree program and so students were not benefitting from this skill in order to manage their learning until the latter stages of their degree.

New Q1. Does the gap between student grades and tutor grades reduce across modules designed as a sequence?

As the design students did not follow a sequential path in their program there was no data for undergraduate design students in regard to this question. However the Business cohort was following a sequenced set of units and when the individual streams were examined (management, personal development and finance) to see if the gap between students and tutors marks reduced, the results did not differ from the totality of streams with convergence occurring by the third year course module (see repeat question 4). When all the tasks were considered however the patterns were erratic showing no gradual reduction in difference between the tutors and students.

These results along with the results from repeat question 3 led us to investigate the type of assessments that students were being asked to undertake within their modules. It was found that in the seven Business modules the first and second tasks were primarily of a similar nature but the third task was usually different in nature, for example, Second year Management core unit required a written Essay followed by a Research paper and then a group based presentation for the final assignment; Management capstone had presentations for Tasks 1 and 2 and then a quiz; Personal development capstone required a portfolio for Task 1 followed by a personal development plan based on the portfolio with a project assessment for the students last piece.

The data were then examined to discover the influence of different types of assessment task.

New Q2. Does mode/type of assessment task (eg. written assignment, project, presentation, etc.) influence students' judgement of grades?

Although convergence was found to occur earlier when students undertook a sequenced set of units of study, that is in the third unit of study for the business students rather than the fourth for the design students, the convergence was not as pronounced as expected; therefore the question of the similar assessment types rather than sequenced subjects was used to investigate convergence between tutors and students over time.

When written assignments were considered (which included essays, reports, briefs, research papers), students' marks converged with their tutors by the third iteration of a written task (t=(1,116) 2.57; p=0.052) for business and by the second iteration (t=(1,68) 1.702; p= 0.093) for design (Figures 5 and 6).

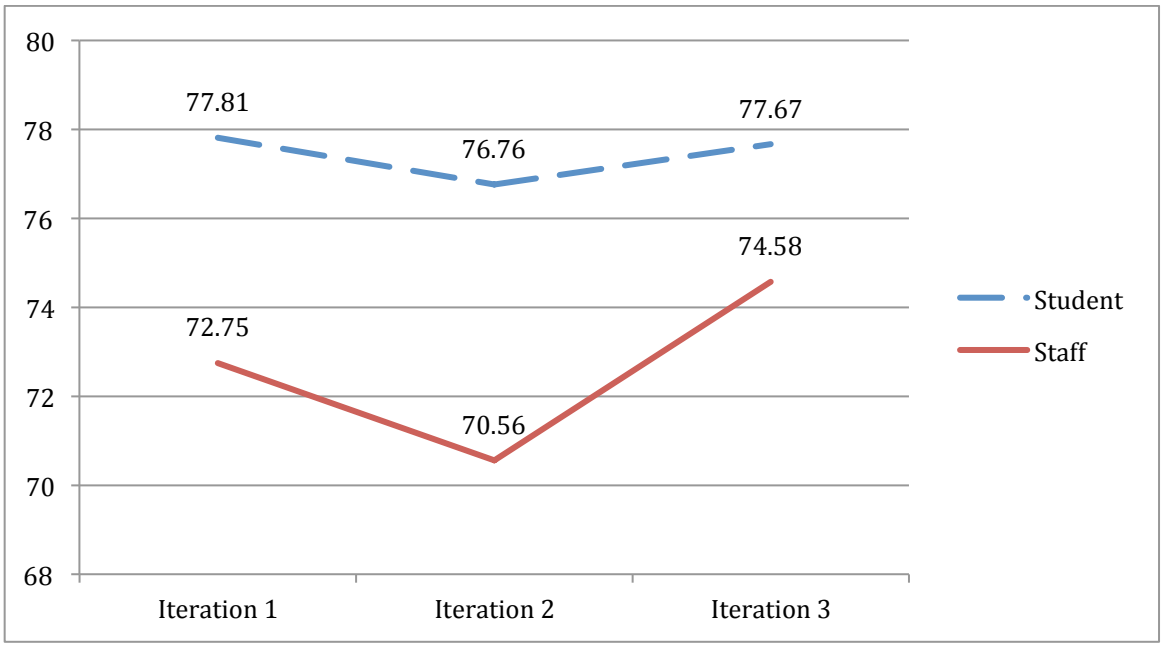

Figure 5. Student and tutor grades for each assessment type–Written assessment (Business)

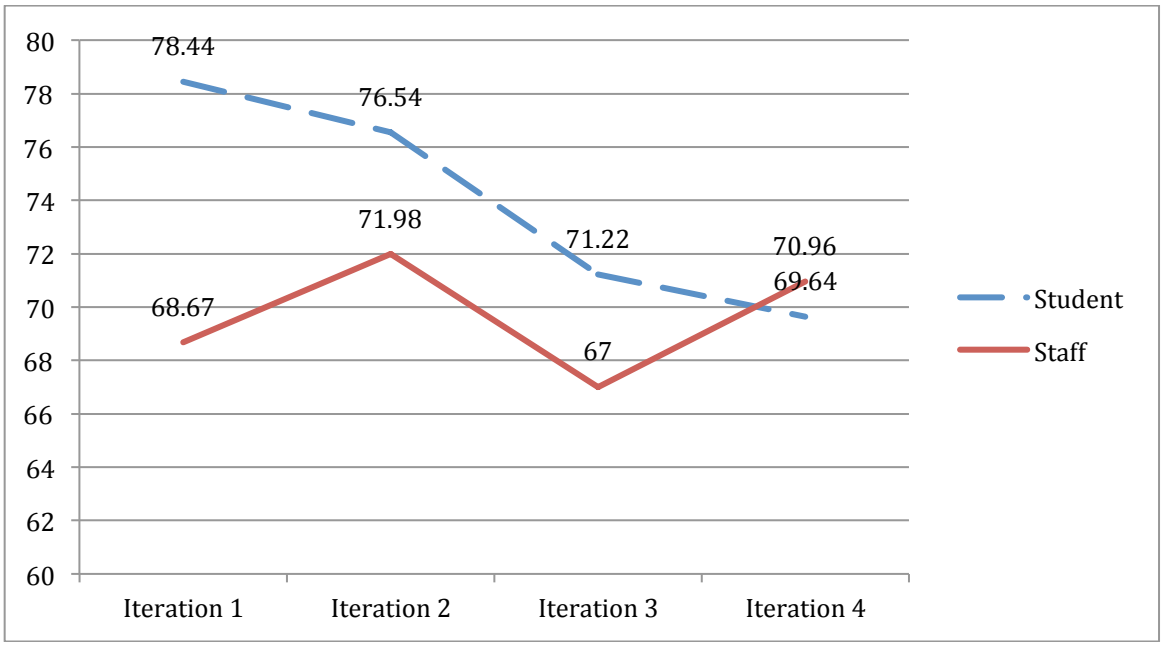

Figure 6. Student and tutor grades for each assessment type–Written assessment (Design)

This shortened length of time to find convergence between students and tutors was also found for the other assessment modes (Table 1).

Table 1. Iterations required for convergence by type of assessment

Assessment Type	Design		Business	
Oral Style Assessments	Iteration 3	(t=(1.15) 2.097; p= 0.053)	Iteration 2	(t=(1,498) -1.603; p=0.109)
Projects	Iteration 3	t=(1,90) 1.992; p=0.059	Iteration 2	t=(1,97) 1.182; p=0.244
Analysis/Problem solving	Iteration 2	t=(1,124) -0.418; p=0.676		
Reflective Journals	Iteration 2	t=(1,139) -0.203; p=0.839	Iteration 3	t=(1,38) -1.891; p=0.066
Skill based	Iteration 1	t=(1,18) 1.563; p=0.135		

Although these results did show earlier convergence for some assessment types, the data was not consistent. This could be due to variation of assessments with a type, for example the difference between an essay and a report. This led us to examine specific criteria that related to the assessment type, for example written communication related criteria for written assignments and presentation skills criteria for oral presentations, rather than all the criteria for each assessment mode.

The data were examined to define categories in order to interrogate this further granularity relating to specific assessment criteria in our new question 3.

New Q3. Does analysis of criteria that relate to type of assessment task influence students' judgement of grades?

The type of assessment tasks had various assessment criteria that related to written or oral communication, critical thinking, practical or technical skills criteria for skill-based assessments and innovative or creative criteria in project assignments. When this analysis was conducted the results were found as in Table 2.

Table 2. Iterations required for convergence by type of assessment and type of criteria

Criteria type	Assessment Task	Design		Business	
Written communication	Written Task	Iteration 2	t=(1,18) -0.186; p=0.854	Iteration 1	t=(1,223) -5.372; p= 0.00
Oral communication	Oral Presentation	Iteration 2	t=(1,8) 0.032; p=0.976	Iteration 2	t=(1,64) -2.854; p= 0.047
Critical thinking	Analysis/ Problems	Iteration 2	t=(1,46) 0.295; p=0.769	Iteration 2	t=(1,40) -26.807; p= 0.00
Practical skills	Skill based	Iteration 1	t=(1,13) -0.853; p=0.409		
Creative and Innovation	Projects	Iteration 2	t=(1,25) 0.911; p=0.371		

These results show that consistent criteria relating to attributes intended for development through a particular type of assessment task will foster calibration of judgement in a reduced amount of time, with results showing that accuracy is obtained by the second iteration. This supports the notion that familiarity in assessment type leads to accelerating students' ability to make accurate judgements. Disruptive assessment patterns where there is no consistency of assessment type or use of criteria across tasks, appears to delay students' development of evaluative expertise.

In summary, these findings address two sets of questions: those replicating the earlier study, and the new ones. In regard to our four repeated questions findings are consistent in many respects with the outcomes of the earlier study. The analysis indicates the following:

1. It was confirmed through the extended data that the accuracy of students' self-assessment varies according to their performance with high ability students underestimating and low ability students overestimating in their gradings against criteria. However, the degree of over or underestimation diminished over the modules studied.
2. The extended data set also confirmed that accurate estimators showed a significant increase in scores across all tasks, and that under-estimators also showed improvement in performance in the Business set.

3. It was confirmed that there is some convergence in differences between student and tutor marks from the first assessed task to the final task within a module or unit of study. However, the Business student data did not show consistent convergence. This led to further interrogation of the data through new questions 2 and 3.
4. It was confirmed that there is convergence between student and tutor marks over time across course modules and semesters. The gap between student and tutor marks lessens in the first assessed task in each subsequent module and then remains constant.

With respect to the new questions the findings indicate that:

1. The gap between student self-assessments and tutor assessments reduced more quickly across modules designed as a sequence of modules that build on each other in knowledge and skills.
2. Early convergence of student and tutor marks was more apparent where similar modes of assessment task were used. However, the data was inconsistent in the design discipline and led to a closer analysis of the marks for individual assessment criteria.
3. Convergence of marks occurred sooner on assessment criteria that were related to the type of assessment task used.

Discussion

The process of having students judge their own grades against criteria provides a way of tracking the development of their skills of self-judgement, one of the key features sustainable assessment is designed to promote. Without a measure of this kind, we have no evidence that assessment tasks designed to be sustainable are having the desired effect. The study of Design and Business undergraduate degree programs was determined by their extended use of the software (ReView™) that allowed interrogation of comparative student and tutor data at a level of granularity necessary for the findings to be further explored in new questions extending the initial study.

The framing of the study mentioned Sadler's three conditions for effective feedback: (1) a knowledge of the standards; (2) having to compare those standards to one's own work; and (3) taking action to close the gap between the two (Sadler, 1989, p.138). Whilst knowledge of standards was not addressed as part of an intervention the study does show that students can become more accurate in judging the standards of their own performance when given extended opportunities to self-assess over two or more modules using normal assessment tasks. Self-assessment and tutor feedback against criteria used as a basis in the study

clearly helped some students to understand the criteria and standards giving a realistic perception of their achievements (O'Donovan, Price & Rust, 2008). However, there is considerable variation in the extent of this understanding and there are many other factors including the nature of the assessments and/or the subject matter involved. It is also noted that using assessment criteria to compare students' self-assessments with tutors may have limitations. As Sadler (1989) discusses, academics are often found to use holistic judgement which they subsequently justify by resort to criteria. Our assumption is that students are not expert enough to do this, particularly in the earlier stages of their courses, and so need criteria to scaffold their judgement. All comparisons in this study are between criterion-specific judgements.

The study does not show that improvement in judgement is necessarily due to criteria-based self-assessment *per se*. However a study by Lawson (2011) found that with development of competency the perception levels rose to more realistic judgements of their learning by the end of the module. In observing improved judgement we recognise that it is particularly challenging to obtain independent measures of such improvements and identify the causes involved.

One of the most educationally interesting findings of the study as in our previous study is the effect on mid-range students. High performing students were most likely to have developed skills of judging their own work already and there is relatively little improvement in their judgement. Low performing students start with poor self-judging skills, and show a modest if any improvement through self-assessment over time. However, mid-range performing students show the most significant improvement. They over-estimated their performance in beginning tasks but by the end tasks there was no significant difference between their own and tutors assessments.

It should be noted that the assessment activities for both the Design and Business programs were not designed to develop student judgement, rather on reviewing the learning activities and assessment tasks it was evident that they were designed for academics to judge the knowledge and related skills for their courses. In Design there had been a complete move away from final examinations, whereas in Business about one third of the assessment weighting was on examinations which the opportunistic data set did not allow us to explore. These disciplinary variations between examinations and other forms of assessment are not unusual in the Australian higher education context.

The other intriguing finding relates to how disruptive patterns of assessment can have an impact on the calibration of judgement. The findings suggest that students develop their judgement less rapidly when faced with lack of consistency of assessment type. That is, variation of assessment mode for assessing given criteria can inhibit the development of student judgement. An explanation for this is perhaps that students are overloaded with the task of understanding how to present their work in a new mode, but can focus their attention more fully on the subject matter they are representing in their work. This is not an argument however for removing variation in assessment method—a range of different approaches is still needed to address different learning outcomes—but it does suggest that a wide variety in assessment methods may not in itself be such a good practice. It may therefore be that a balance is needed in the variety of tasks, with a small repertoire of appropriate assessment methods employed to foster learning outcomes that allow opportunity for students to become familiar with the expectations of each mode.

Some caution should be expressed about accepting this analysis too readily. As the use of the self-assessment process was entirely voluntary, the data produced while representing a very large number of students does not reflect the whole population of students. Students who did not complete the self-assessment activities might be postulated to be less engaged in the program and have a higher proportion of lower-achieving students. If this is the case, then the findings with regard to lower achieving students showing little improvement is even more worrying as those included in the data may well be the more diligent representatives of this category of student. Also, we should be wary of implying that it is the formalising of self-assessment that causes the effects identified here. Other factors might be expected to be influential: comments received from staff, discussions with peers and indeed students' own aspirations. Nevertheless, without tracking studies of the kind illustrated here, we are unlikely to be able to observe whether the goal of students being able to effectively judge their own work is being achieved.

Conclusion

In concluding we note that this study raises interesting questions about the embedding of self-assessment and feedback processes as part of a normal engagement with criteria and standards for both tutors and students. Subject documentation may describe intended learning outcomes but they are often not explicit in assessment criteria and their related attribute developments remain unclear. This study suggests that

significant impact on students making judgements about their work is not through continuity of content (sequenced units of study), but consistent assessment criteria relating to attributes that are intended for development through particular types of assessment task. We suggest that engaging students in self-assessment practices provides opportunities to further develop their judgement and that this should include not only opportunities for such self-assessment but tutor feedback on the students' self-assessment as part of a dialogue about improving self-regulation skills

The significant improvement in the judgements of mid-range ability students raises questions about the lack of improvement for both low and high ability students. The motivation of students to self-assess may relate to these questions and interventions that reward self-assessment activity either summatively or formatively may form the basis for further studies.

References

Boud, D. (1990). Assessment and the promotion of academic values, *Studies in Higher Education*, 15(1), 101-111.

Boud, D. (1995). *Enhancing Learning through Self Assessment*. London: Routledge.

Boud, D. (2000). Sustainable assessment: rethinking assessment for the learning society. *Studies in Continuing Education*, 22(2), 151-167.

Boud, D. and Falchikov, N. (2007). Developing assessment for informing judgement. In Boud, D. & Falchikov, N. (Eds.) *Rethinking Assessment for Higher Education: Learning for the Longer Term*. London: Routledge, 181-197.

Boud, D., Lawson, R. and Thompson, D. (2013). Does student engagement in self-assessment calibrate their judgement over time? *Assessment and Evaluation in Higher Education*, doi:10.1080/02602938.2013.769198

Boud, D. and Molloy, E. (2013). Rethinking models of feedback for learning: the challenge of design.*Assessment and Evaluation in Higher Education,* 38(6), 698-712.

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

Ericsson, K. A., Krampe, R. T. & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*. 100(3), 363-406

Falchikov, N. & Boud, D. (1989) Student self assessment in higher

education: a meta-analysis. *Review of Educational Research*, 59(4), 395–430.

Fazey, J.A. & Marton, F. (2002). Understanding the space of experiential variation. *Active Learning in Higher Education,* 3(3), 234–250.

Knapper, C. K. and Cropley, A. J. (2000). *Lifelong Learning in Higher Education*. 3rd ed. London: Kogan Page.

Lawson, R. (2011). Constructively aligned teaching methods and students' approaches to learning and motivational orientations. *Global Journal of Human Social Sciences*, 11(8), 59-68.

Nicol, D.J. & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*. 31(2), 199–218.

O'Donovan, B., Price, M. & Rust, C. (2008) Developing student understanding of assessment standards: a nested hierarchy of approaches, *Teaching in Higher Education*, 13(2), 205-217

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, 28(1), 4–13.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119–144.

van Merrienboer, J. and Kirschner, P. (2007) *Ten Steps to Complex Learning: A Systematic Approach to Four Component Instructional Design*, New Jersey: Lawrence Erlbaum Associates.