Does student engagement in self-assessment calibrate their judgement over time?

Article *in* Assessment & Evaluation in Higher Education · December 2013

DOI: 10.1080/02602938.2013.769198

CITATIONS
193

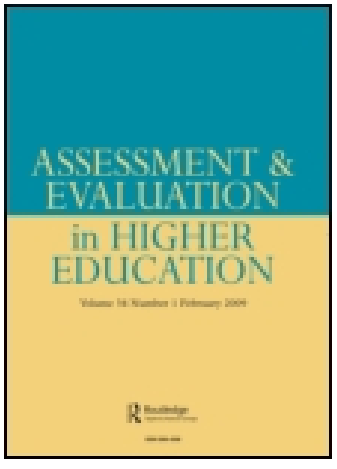
READS
918

3 authors:

David Boud
Deakin University
312 PUBLICATIONS **35,299** CITATIONS
SEE PROFILE

Lawson Romy
Flinders University
24 PUBLICATIONS **521** CITATIONS
SEE PROFILE

Darrall Thompson
University of Technology Sydney
19 PUBLICATIONS **469** CITATIONS
SEE PROFILE



Assessment & Evaluation in Higher Education

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/caeh20

Does student engagement in self-assessment calibrate their judgement over time?

David Boud[a], Romy Lawson[b] & Darrall G. Thompson[c]

[a] Faculty of Arts and Social Sciences, University of Technology, Sydney, Australia.

[b] Faculty of Law, Business and Creative Arts, James Cook University, Sydney, Australia.

[c] Faculty of Design, Architecture and Building, University of Technology, Sydney, Australia.

Published online: 20 Feb 2013.

To cite this article: David Boud, Romy Lawson & Darrall G. Thompson (2013) Does student engagement in self-assessment calibrate their judgement over time?, Assessment & Evaluation in Higher Education, 38:8, 941-956, DOI: 10.1080/02602938.2013.769198

To link to this article: http://dx.doi.org/10.1080/02602938.2013.769198

Assessment & Evaluation in Higher Education, 2013
Vol. 38, No. 8, 941–956, http://dx.doi.org/10.1080/02602938.2013.769198

Does student engagement in self-assessment calibrate their judgement over time?

David Boud[a]*, Romy Lawson[b] and Darrall G. Thompson[c]

[a]*Faculty of Arts and Social Sciences, University of Technology, Sydney, Australia;* [b]*Faculty of Law, Business and Creative Arts, James Cook University, Sydney, Australia;* [c]*Faculty of Design, Architecture and Building, University of Technology, Sydney, Australia*

One of the implicit aims of higher education is to enable students to become better judges of their own work. This paper examines whether students who voluntarily engage in self-assessment improve in their capacity to make those judgements. The study utilises data from a web-based marking system that provides students with the opportunity to assess themselves on each criterion for each assessment task throughout a programme of study. Student marks were compared with those from tutors to plot changes over time. The findings suggest that overall students' judgements do converge with those of tutors, but that there is considerable variation across achievement levels, with weaker students showing little improvement. Whilst the study is limited by the exigencies of voluntary participation and thus consequential gaps in the data set, it shows how judgement over time can be demonstrated and points to the potential for more systematic interventions to improve students' judgements. It also illustrates the use of the web-based marking and feedback software (ReView) that has considerable utility in aiding self-assessment research.

Keywords: self-assessment; judgement; student participation; learning; assessment software

Introduction

One of the core purposes of education is to develop the capacity for students to make judgements about their own work (Boud and Falchikov 2007). Such self-evaluation is needed both to enable effective study, so that students can focus on the most important aspects of their work they need to improve, and to build the skills that they will need in any area of work following graduation. If a graduate is not able to make their own judgements about the quality of their work, they will be ill equipped for most professional or even non-professional roles. The development of the capacity to make self-judgements about performance tends to be an assumed outcome of higher education. That is, it is taken to be a part of any course without the need for specific practice. This is possibly an act of faith, as it is rarely evident in curricula through learning activities or assessment processes (O'Donovan, Price, and Rust 2008).

In contrast to this, research on student self-assessment has suggested that explicit opportunities need to be included for the skill of self-assessing to be developed (e.g. Boud 1995). Building the capacity to make judgements needs to be

*Corresponding author. Email: david.boud@uts.edu.au

an overt part of any curriculum and one that needs to be fostered (Boud and Falchikov 2007). If this is the case, then the following questions arise. How might such capacity for judgement be encouraged? Does engagement in making such judgements over time improve capacity for doing so? The extensive literature on self-assessment in higher education addresses the first of these questions and suggests that self-assessment activities are beneficial. It has been known for many years that under appropriate conditions, students can judge their performance on common assessment tasks (Boud and Falchikov 1989; Dochy, Segers, and Sluijsmans 1999). Additionally, students in later years of their course are better able to judge their performance than in earlier years (Falchikov and Boud 1989). What is less apparent is students' performance in criteria-based assessment contexts and the circumstances in which their judgement can improve (Ward, Gruppen, and Regehr 2002; Galbraith, Hawkins, and Holmboe 2008).

This paper addresses the second of these questions to determine whether more extensive opportunities are than offered by the typical within-module self-assessment intervention lead to students improving their capacity to make judgements about their own work. The study uses data from a web-based assessment system that enabled students to make self-assessments against descriptive assessment criteria. It examines the development of student judgement across course modules to explore whether students' judgements improve over time and whether any effects, in terms of students' grades, differ across a cohort.

Data from an online marking system were gathered as a result of student's voluntary use of self-assessment. Students allocated criteria-based grades to their submitted work prior to knowledge of the criteria-based grades they were given by tutors. Lecturers or tutors graded on the same scales without knowledge of the student's judgement. These data enabled us to track students in their self-assessment performance across subjects (units of study/courses) and across semesters. The data provided an opportunistic experiment (as it was not originally collected for this purpose) to study the disaggregated grade judgements of tutors and students across a range of subjects and semesters of study. As students were neither required to self-assess, nor were they rewarded for doing so, there are some data gaps that will be discussed.

Conceptual background

Developing judgement and how courses limit it

Graduates who are able to be effective practitioners in any area need to have the capacity to make judgements about their own work. Once students move beyond the protected environment of a course, they need to be able to do this for themselves, in conjunction with others; drawing upon whatever resources they have available to them. A person who has the capacity to make good judgements about their work will be able to know why and how their work can be improved (Black and Wiliam 1998). They will also be aware of the scope of their practice and when they will need to refer to and involve others, as well as recognising areas for further development.

The capacity to make judgements is not well represented in many current assessment practices. Assessment items are often strongly knowledge-based, with criteria unilaterally set by teachers. The role of students tends to be to offer themselves to be assessed by others. This can create dependency on the authority of the teacher, rather

than other sources of judgement, and can give rise to the implication that judgements are necessarily made by others. This is in contrast to the learner being positioned as an active agent in assessment decisions, as is advocated by many assessment theorists (e.g. Nicol and Macfarlane-Dick 2006; Nicol 2009).

The making of judgements is often an informal and personal act that may or may not occur as students prepare themselves to be assessed by others. Many formal acts of assessment, particularly those used for summative purposes, are positioned to effectively deny the possibility of students taking such responsibility.

Development over time

Developing judgement involves more than making self-assessments, and it is not necessarily strongly promoted by the addition of simple self-assessment interventions (e.g. Boud 1995). It certainly involves more than self-testing. Engagement with criteria and the standards to which they are to be applied is quite central to judgement. Sadler argues that self-evaluative skills need to be developed 'by providing direct authentic evaluative experience for students' (1989, 119), i.e. making specific judgements about particular work.

It is unlikely that one-off examples of self-assessment will build capacity for judgement, and it is even more unlikely that such examples are able to do so beyond the immediate knowledge domain of the particular case. Such capacity needs to be promoted systematically throughout courses, as it is reasonable to assume that, like any expertise, it is related to each knowledge domain encountered (e.g. Dreyfus and Dreyfus 2005).

We assume that the key feature of the development of judgement, like any other kind of expertise, is that it requires consistent engagement over time (e.g. Ericsson, Krampe, and Tesch-Romer 1993). Standards for the quality of work need to be assessed and interpreted, and these need to be applied in the work of the student. Different standards for different kinds of work are needed, and students need considerable practice in working out how to identify what is appropriate in any given situation and how they can see their own work with sufficient distance to be able to apply such standards.

Calibrating judgement and the role of feedback

What is required for students to learn how to make judgements? Sadler suggests that students develop skills in evaluating the quality of their own work through moving beyond 'teacher-supplied feedback to learner self-monitoring', and that the instructional system in which they operate [the course] needs to 'make explicit provision for students themselves to acquire evaluative expertise' (1989, 143). We posit that students learn by consistently making evaluations and relating these to the evaluations of others: reflecting if their judgements were accurate or not, looking for reasons behind poor judgements and for ways to improve future judgements, wondering what they have missed in making their judgements that others have seen. Such activities cannot be done in isolation. It needs the development of evaluative expertise and the input of others. In particular, it needs input from those who can tell if appropriate judgements about the quality of work are being made. These may be teachers, practitioners, or, for some aspects, students' peers. As Sadler describes, 'providing guided but direct and authentic evaluative experience for students enables them to

develop their evaluative knowledge, thereby bringing them within the guild of people who are able to determine quality using multiple criteria' (1989, 135).

However, Sadler, following Ramaprasad (1983), identified that the possession of evaluative expertise is a necessary (but not sufficient) condition for improvement. He identified 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, 138). None of these are simple processes. Knowledge of standards requires information about what counts as good work in any particular area and the identification of appropriate criteria that indicate these standards. Comparing these standards to one's own work needs the ability to operationalise or ground the standards in relation to the particular kind of product being judged. This might require the use of models or exemplars of what a standard might mean. Finally, taking action to close the gap requires opportunities for such an activity to reoccur. When courses are forever moving on to new material, occasions for continuing practice might be difficult to find.

What might also be needed to aid comparisons is for judgements to be calibrated against those who might be regarded as experienced judges of the kind of work being considered. Judgements need to be made in the light of those of appropriate others, information gained about discrepancies between the judgement of the novice and that of the experienced judges, and judgements refined. Feedback on such discrepancies is probably more important than feedback on any other matter, because if a misperception of judgement occurs, then the learner may not know that they need to take any useful action to remedy their work and perform better subsequently.

There is an inherent tension, however, in providing guidance to students on their own judgements to help improve them, and students becoming more able to exercise evaluative judgement independently of teachers. Providing information to students to assist them calibrate their judgement is only one part of a more complex process of them developing their own expertise. Students need also to learn when not to trust the judgements of others.

Grade judgements

One measure of students' ability to judge their own work is the grades they give to assignments. A student who is a good judge of their own work is likely to rate their work in a similar way to an experienced judge of the same assignment, assuming they share the same criteria for assessment. These experienced judges should be viewed with caution, however, as research on the reliability of tutor marking over the past 85 years or so (since Hartog and Rhodes 1935, 1936) would suggest that there can be considerable error and inconsistencies in tutor judgements and also variation across tutors, depending on the nature of the task assessed. Notwithstanding this, the most readily available surrogate for an expert judge is the person who marks assignments and allocates grades. Whilst over many assignments this person changes, and their marks are subject to normal variation, without the intrusion of other measures, this, with the moderation processes that take place, is as close as one can get to expert judgement in a normal teaching environment. It should also be noted that there might be a difference between measurement (that is marks) and judgement (what is acceptable or not). Yorke (2007) discusses that when judgement is used rather than measurement then marker reliability is far higher.

There have been considerable studies over a long period of time comparing students' marks with those of teachers (Boud and Falchikov 1989; Falchikov and Boud 1989; Dochy, Segers, and Sluijsmans 1999). These and subsequent studies show that students are reasonable judges of their own grades, but that the accuracy of judgement varies according to the expertise of the student and the level of course: stronger students are more likely to underestimate grades, weaker students over-estimate; students in advanced courses are more likely to underestimate, students in introductory courses over-estimate. It follows from this that it would not be surprising to find a tendency that when students encounter new subjects or academic areas, their ability to make good judgements of their work declines.

There are many limitations of such studies as Ward, Gruppen, and Regehr (2002) have pointed out. Not least of these is reliance on experienced raters of performance as the gold standard in self-assessment. They suggest the need for multiple experienced raters, better use of scales and using an intra-individual, as opposed to inter-individual, comparison process, that is, comparing individuals with their own performances over time. They warn though that:

> studies that make use of the traditional designs to study self-assessment without accounting for the potential methodological flaws inherent in these approaches will not be able to contribute meaningfully to the self-assessment literature in the future. (80)

The study

Our study involved the tracking of assessments students made of their own work against that of the marker across assignments in an undergraduate design programme in an Australian university. The assessments were facilitated by a criteria-based assessment system (ReView). The degree programme used the web-based system to publish criteria that referred to the specifics of each task rather than 'fixed sets of criteria' (Sadler 2009, 159). These criteria make explicit the aspects of learning valued in the assessment of the subject taught. Instead of usual grade codes (F, P, C, D, HD) being used in benchmarking with tutors and discussions with students, the more descriptive terms passable, creditable, distinctive and highly distinctive were used. These terms were intended to replace the usual reference to percentage marks to encourage the development of a judgement culture (Sadler 2005, 190).

The completion of student self-assessment revealed visually the variation between their own grading and the tutor's grading for each criterion. During the period of the study, students engaged in voluntary self-assessment and graded their work on a sliding scale against criteria prior to the tutors entering their grading into the web-based system. Whilst the percentage marks were not shown to students in their ReView feedback, these were recorded in the database and have been used as the data for statistical comparison in this study.

Data used

Students were enrolled in a four-year undergraduate design programme, the Bachelor of Design (Honours). The degree has four discipline strands – Industrial Design, Visual Communication Design, Fashion and Textile Design and Interior Design – with a total of approximately 1400 students.

The data used in the study were students' individual self-assessment grades for up to four tasks in each subject taken per semester. Each of the tasks had descriptive criteria against which percentage marks were gathered as well as the total percentage mark for each task. Tutors' judgements of student performance were gathered for comparison with students' self-assessments, which the system also stored as percentage marks. Each subject was taught by varying numbers of tutors according to subject size, but in this study between one and six. In the Design programme, there has been a tradition of moderation across tutors to reduce inter-tutor variation.

The ReView software is a web-based marking aid developed by Darrall G. Thompson, an academic in the Visual Communication Design programme. Due to the convenience of online marking and improved efficiencies in the management of tutors, the system usage in the design degree programme has gradually increased during the period from which data were collected, from 56 subjects (units of study or course modules) in 2008 to 127 subjects in 2010. However, this increase has largely been through word of mouth with no official training or guidance for academics, although approximately 20% were personally supported in the use of the system or attended demonstrations of the software.

ReView was designed to give criteria-based feedback and comments. It provides various options that can be enabled (including self-assessment). Students can track their progress over time through a visual representation of grades by category of criteria. The student self-assessment option was available to academic staff when setting up their tasks or assignments for marking, but unless staff attended one of the demonstrations mentioned they would be unaware of the option and its educational value. The data used in this study are from academic subjects that had this feature enabled during the data collection period. These included the following:

- 13 in Industrial Design
- 7 in Multidisciplinary Research
- 10 in Interior Design
- 20 in Visual Communication Design

These subjects had between a minimum of two and a maximum of four summative assessment tasks for students to complete.

The tutors in these subjects were professional practising designers, and the subject coordinators were full-time or fractional academics with professional design experience. The study data are drawn from a broad range of assessment tasks, they included the following:

- individual and group projects, research reports and oral presentations,
- critical and reflective essays,
- portfolio presentations of individual exhibitions of work.

There are no examinations used in the design degree.

The assessment criteria for these tasks varied in clarity and level of specificity. Subject coordinators classified them into groupings of graduate attributes and edited them to be relevant to the task.

To illustrate what students might see, the following criteria are from one Industrial Design task weighted at 50% of the total assessment in the subject:

(1) Professional approach to the explanation of consumer benefits and other factors affecting the appropriateness of solutions.
(2) Appropriate use of convergence tools, eg. linkograph, taught in this subject.
(3) Depth of consideration regarding the evaluation and validation of your proposal.
(4) Level of innovation and or creativity evident in your proposal.
(5) Quality of synthesis of ideas generated through creative thinking processes.
(6) Appropriate outcome based on the context outlined in your restated brief.

Another example is from an interim presentation task in the Visual Communication Major Project subject, weighted at 30%:

(1) Level of innovation and appropriateness of design response to findings drawn from archive.
(2) Clarity of project presentation communicated by the design of the three A3 panels.
(3) Quality of understanding of positioning of designed response in relation to viewer engagement/response and potential social implications (including ethical considerations where appropriate).
(4) Quality of presentations to audience of peers, tutors and industry advisor.
(5) Appropriateness of project proposition evidenced in the research underpinning the project.
(6) Depth and range of visual (contextual, experiential and/or generative) investigation undertaken.
(7) Clarity and relevance of insights/significant findings drawn from the archive.

There was a visual grading scale next to the criterion that corresponded to the following bands:

- 0–49% F (Fail)
- 50–64% P (Passable)
- 65–74% C (Creditable)
- 75–84% D (Distinctive)
- 85–100% HD (Highly Distinctive)

The process of tutors entering ratings against criteria happens through the use of visual ‘data sliders’ as shown in Figure 1.

Once the staff member has completed marking for a student’s work and clicked ‘Save’, if the student has self-assessed, blue triangle sliders appear. This immediately flags to the marker that there were sometimes disparities between their marks and the students grading judgements. Some staff used this disparity to subsequently guide comments to students typed into a comment box. This allows for feedback to focus on areas where students have a lack of accurate perception of their performance on a criterion, and is, therefore, part of their work that needs extra feedback. There is also a moderation mechanism within the system where subject coordinators can benchmark marking across the teaching team.

Students use a similar interface to enter their self-assessment ratings of their own work. After the subject coordinator publishes the tutors’ marking, students can see their self-assessment triangles compared to the tutors’ assessments. In the

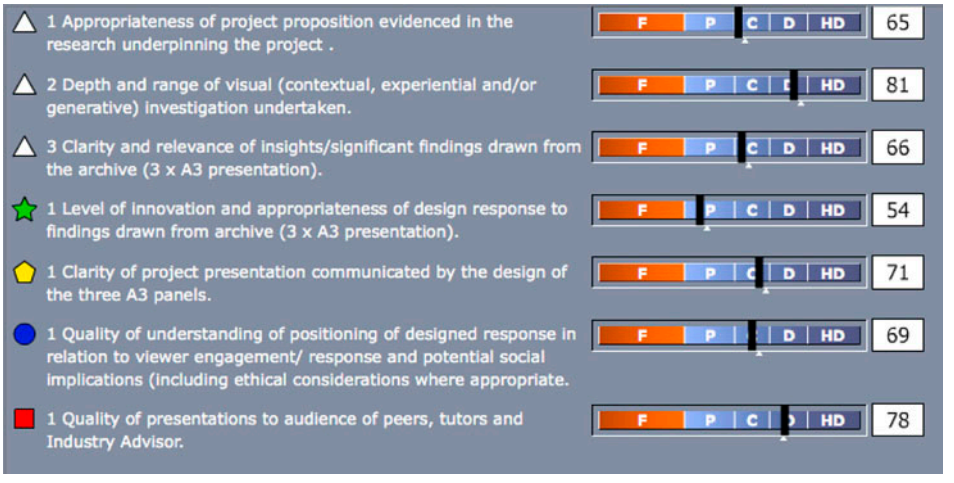


Figure 1. Screenshot from the ReView software showing data sliders for tutors marking against criteria. As the black bar is dragged the percentage numbers appear alongside the slider.

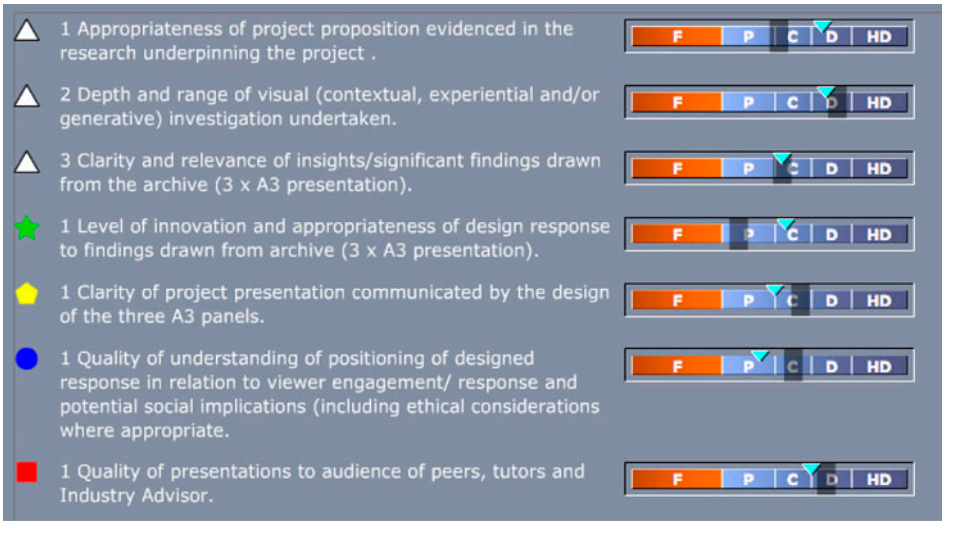


Figure 2. Screenshot of students interface showing the tutor's assessment (grey bars) in relation to their own self-assessments (triangles).

student interface, percentage numbers are not displayed, and the tutors' black grading bar was spread (as can be seen in Figure 2) in order to reduce the focus on marks and foreground the pattern of feedback against criteria. It should be noted that ReView is not the official repository for marks and does not replace the record-keeping system administered by the university.

Motivation to engage in self-assessment

No student in this study was *required* to self-assess against the criteria or rewarded for doing so. There are a number of reasons why students may have decided to

engage in self-assessment. The novelty factor or encouragement from a keen academic tutor may have encouraged initial participation, but the students in this study self-assessed over at least two semesters. So perhaps, they were using self-assessment as a way of understanding the criteria used for assessing a task or developing a value for the visual comparison between their own estimates and teachers' grades. Given the potential educational benefits of this reflective process, investigating how to improve the intrinsic motivational factors could be an interesting further study. Missing data arise both from those subjects in which coordinators had not enabled the self-assessment features and from students who chose not to avail themselves of the facility when it was enabled.

Response rates

This research was not conceived until after this assessment system (ReView) had been used for several years by academics teaching in the degree. Given that the use of ReView was voluntary, we selected to use data from students who had used ReView to self-assess over a minimum of two semesters. This meant that even though we had self-assessment scores matched with tutors' marks for over 13,000 criteria, we only used 2196 self-assessments from 182 students. As the study was focusing on the effects of self-assessment over time, we also conducted analyses on students who had used ReView over three semesters (66 students) and four semesters (24 students). The data for students completing more than four semesters of self-analysis were limited and so has not been included in this study. It should also be noted that numbers for third and fourth tasks are reduced, as a number of subjects required only two summative assessments tasks within their subject.

The data are clearly limited, and the results of the analysis reported here are merely suggestive. We do not know what the effects would be of including students who chose not to undertake the exercise. We suspect that, as the missing students are more likely to be the less conscientious and perhaps the less able, the effects may be weaker than we identify here. The act of expecting students to undertake self-assessments though is not predictable. It may be that a formal requirement could act as a scaffolding effect to support the very students who need most assistance in developing their judgement. However, unless students enter into the process with serious and committed intent, then any intervention is likely to be ineffective.

Questions addressed

The data available enabled us to address the following questions:

(1) Do students' marks agree with tutors' within a subject?
(2) Do differences between tutors and students decrease with each subject undertaken?
(3) Does students' overall performance affect their ability to agree with tutor marks?
(4) Does students' ability to calibrate lead to improved performance?

1. Do students' marks agree with tutors' within a subject?

A series of paired *t* tests were conducted on overall differences between tutor and student scores on tasks within a single subject. There was a significant difference

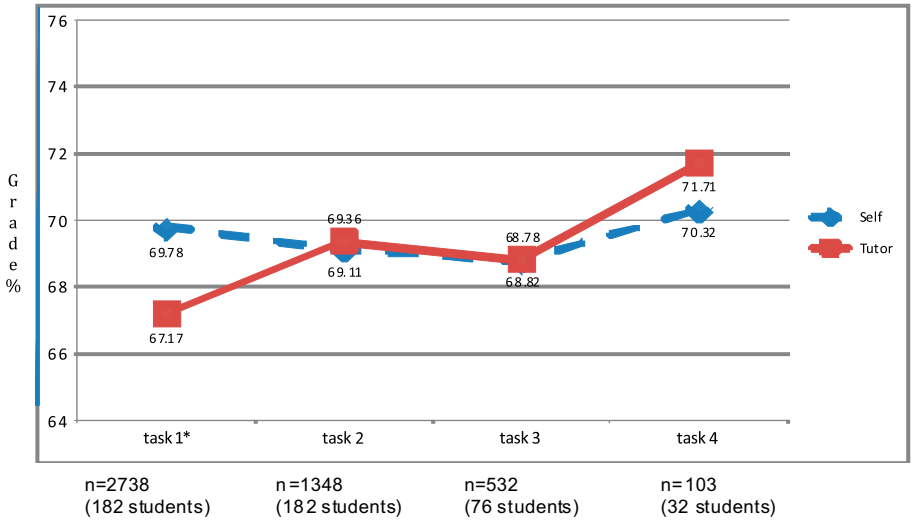


Figure 3. Comparison of tutor and student scores on assessment tasks within a subject.

found between the student and the tutor at the first task, with students rating themselves higher than the tutors ($t(1, 2827) = 9.279$; $p < 0.00$). By the second task, this significant difference was no longer evident ($t(1, 1347) = -0.654$; $p = 0.513$), and remained non-significant for third ($t(1, 553) = -0.068$; $p = 0.945$) and fourth ($t(1, 102) = -1.482$; $p = 0.141$) tasks where present.

This suggests that, although students may initially struggle to accurately self-assess, with time and benchmark scores from their tutor they appear to get more accurate. However, the data show greater divergence on the fourth task, but the sample size for this is diminished. It should be noted that, whilst most subjects had three assessment tasks per semester, some had four, so in Figure 3, the third task is the final task for the semester for the majority.

2. Do differences between tutors and students decrease with each subject undertaken?

A series of paired *t* tests were undertaken to examine if the difference between students and tutors marks at task 1 decreased with practice over semesters. It was found that students marked 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 ($t(1, 216) = 1.956$; $p > 0.05$).

This suggests that, for the first task in a new subject, students appear to refine their judgement over time. There is convergence between student and tutor marks for each first task. However, the sample size for semester 4 is considerably lower than for the other semesters. When the effect size for semester 4 is considered in the light of this smaller sample size, it is found to be 0.05, which indicates a large overlap between the tutor and student marks and so reinforces the *t* test finding.

3. Does students' overall performance affect their ability to agree with tutor marks

In order to investigate differences between students with differing achievement levels, the sample was divided into three achievement level groups: low, mid and high. This was derived from the tutors' marks on completion of the assignments. The low achievement group consisted of students who scored less than 60% in their tasks, the high achievement group was students who were marked at higher than 75%, and the mid group were the students who fell between these two percentages. When students were examined according to their achievement level, it was found that the low achievement students significantly over-estimated their performance at both the first ($t(1, 249) = 22.461$; $p < 0.00$) and the last ($t(1, 81) = 12.724$; $p < 0.00$) examples of assessments. The high achievers significantly underestimated their performance at the initial ($t(1, 778) = -15.398$; $p < 0.00$) and final ($t(1, 173) = -8.297$; $p < 0.00$) stages of assessment. The mid group however were significantly higher than the tutors at the beginning task ($t(1, 1698) = 13.427$; $p < 0.00$), but by the end task, there was no significant difference between themselves and the tutors ($t(1, 65) = 1.466$; $p = 0.148$).

This suggests that it is the mid achievement group who were the most able of the three groups in developing self-assessment skills in this context.

4. Does students' ability to calibrate lead to improved performance?

The students were categorised as being over-estimators, under-estimators or accurate estimators in order to address this final question. The groups were divided by calculating the difference between the student and tutor mark on each task on completion of the assessments. Those students who were within 3% (above or below) of the tutor's score were deemed to be accurate estimators, whereas students who were more than 3% below the tutor were classed as under-estimators, and those more than three per cent above the tutor's mark were assigned as over-estimators. A one-way ANOVA was conducted to look at differences in performance scores in relation to ability to calibrate: that is, over-estimators, under-estimators and accurate estimators. In the first semester, the over-estimators showed significantly higher scores on each subsequent task compared to the first ($F(3, 606) = 12.607$; $p < 0.00$). Again, in the first semester, the under-estimators showed a significantly higher score in the second task than the first task ($F(3463) = 3.489$; $p < 0.05$), but did not show any significant differences in following tasks or semesters. The accurate estimators, however, showed a significant increase in scores across all the tasks in the first ($F(3700) = 10.099$; $p < 0.00$), second ($F(3688) = 6.171$; $p < 0.00$) and third ($F(3222) = 5.064$; $p < 0.01$) semesters.

These data suggest that students who are both accurate estimators (mid-range achievers), and, to a degree, those who tend to underestimate their performance (high achievers) improve their performance over successive tasks. However, over-estimators, who tend to be poor achievers, do not appear to learn how to improve their performance over time.

Discussion

The finding (Figure 3) that students' self-assessment marks converge with tutors over the length of a semester is not unexpected and supports previous work on self-assessment (Lew, Alwis, and Schmidt 2010; Lawson et al 2012). It is, however,

encouraging to note that, within a subject, students' understanding of the criteria and standards expected in that particular subject develops.

Figure 4 tells a more detailed story, indicating that, although within a subject, students' ability to accurately self-assess increases when they begin a new subject, the difference between self-assessment mark and tutor mark is again evident. This may be due to having to understand a new set of criteria and standards for each subject and so would suggest that this increase in accurate self-assessment is not immediately transferrable. It was not until students had experienced three semesters of self-assessment that they showed the ability to adapt to a new subject and more accurately self-assess from the first task. It should be noted that completing three semesters involves a change in academic year for the student and so the potential for facing higher standards, for example, the difference from first year to second year.

The breakdown of students into groupings of high, mid and low achievement reveals quite strong contrasts between the groups. The data in Figure 5 confirm the findings from many other studies of self-assessment (Falchikov and Boud 1989) that high achievers tend to underestimate performance and low achievers over-estimate performance. However, Figure 6 provides some intriguing hints about more detailed differences in these groups. The accurate estimators tend to increase their performance as the semester progresses on each task, with the under-estimators also showing some improvement in performance, but the over-estimators do not show any progress in performance over time. It may be that this group is content to merely pass each task and has no desire to invest the effort to do better, or they may not have the capability to improve without additional educational interventions. This study does not provide an explanation of why this might occur.

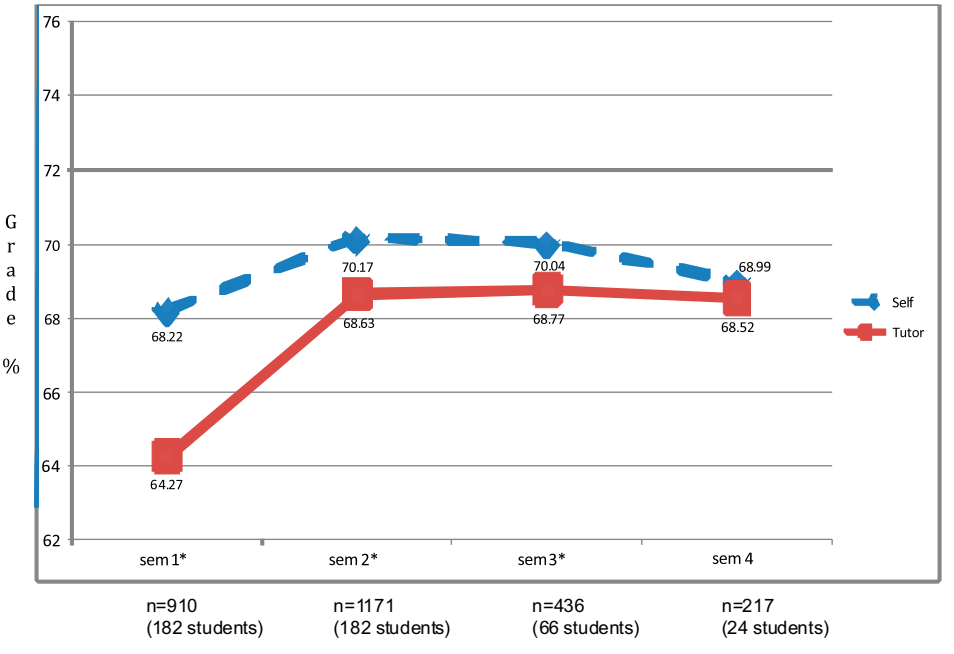


Figure 4. Differences between student and tutor marks on the first task in each subject over four semesters.

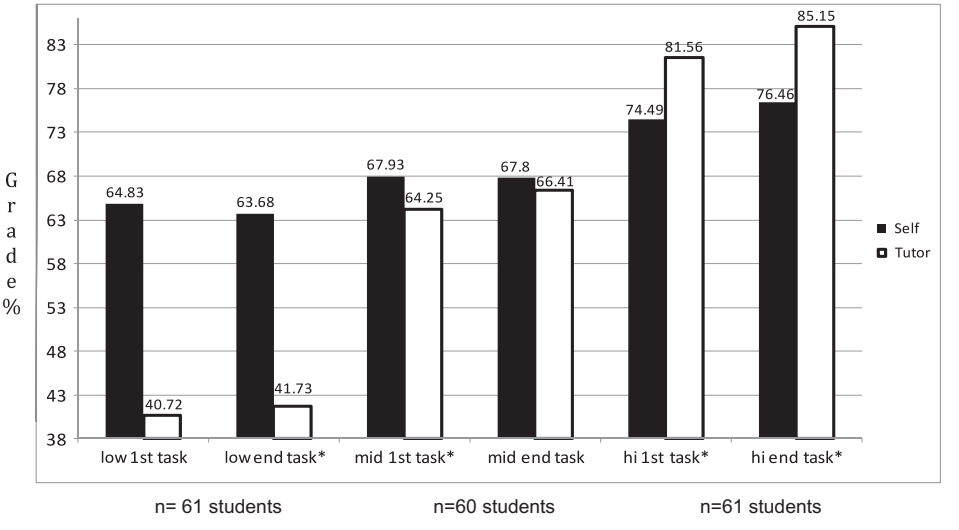


Figure 5. Comparisons between first and last assessment tasks in a given semester by student achievement level.

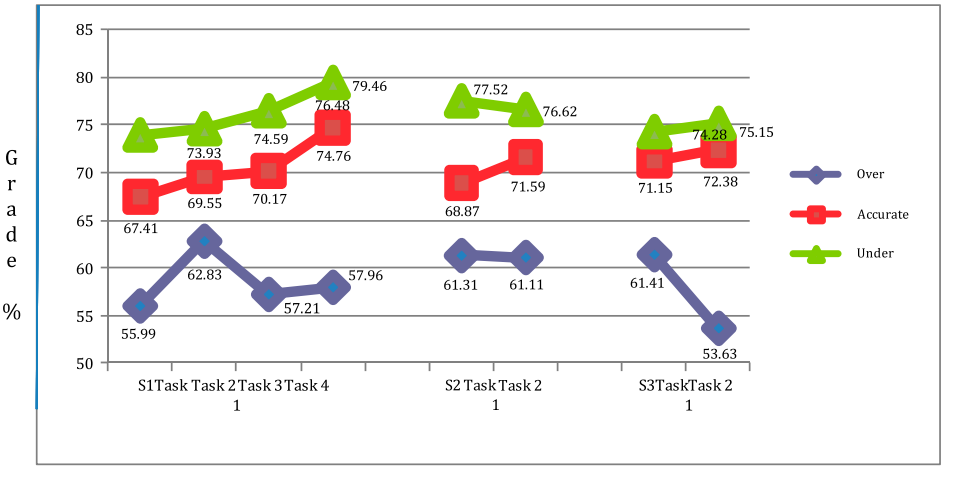


Figure 6. Difference in performance scores by ability in calibrating judgement.

Implications

Whilst this is an initial study with the various limitations noted, it does suggest that students can become better judges both within a subject and across a range of subjects over time. However, when confronted with new subject matter, their judgement declines somewhat, suggesting that the ability to make judgements may be domain-specific. Judgement improves again with further application in more subjects. This provides some support for a practice effect.

The study gives support to the idea that students can improve their grades and become more effective judges of their own work through self-assessment practice: that is, with knowledge of standards and comparison of standards to their own work and 'direct authentic evaluative experience' (Sadler 1989, 119). But it does so only

with students who volunteer to undertake the process, and we have no knowledge of the extent of particular action taken by students to close the gap between standards and their own work. The study also suggests that there may be a regression effect in that, when students are confronted with a new subject, there is a tendency for them initially to be less effective in judging their performance than in a subject with which they have had prior practice. There is also a suggestion that students become more effective in criteria-based judgement over semesters.

Improvement in ability to make judgements is of interest. If the improvement was due to the repetition of being given the opportunity to self-assess, then one would expect judgements to continually improve over time over all units of study. This was not the case. It would, therefore, suggest that this improvement within a subject would be due to students gaining a better understanding of the expectations of the assignments. This may be through the experience of completing the work or by receiving feedback on it from the experienced marker. The improvement in judgement over time would suggest that students also have to learn to adapt their understanding of these expectations to each new unit of study, a skill not mastered automatically.

In order to explore this further, a fuller data set that includes a greater sample of students over a longer timescale would be necessary. A full population of students rather than just volunteers would also be desirable. Further studies need to be considered in other disciplines, as our findings may arise as an effect of courses in design or of assessment types used in the courses studied. It would also be valuable to consider groups of course modules in which one builds on another to see if the apparent learning effect is greater within particular subject matter. Of greater importance would be studies that examined whether students who improved their self-assessment accuracy increased their grades vis-a-vis other students. If there were greater security of findings, there would be important implications for the design and structure of courses and the involvement of students in assessment decisions. More deliberate interventions might have a greater impact than the rather passive measures explored in this study. Strategies that might be considered are the provision of detailed feedback information from tutors on the quality of students' self-assessments, and the engagement of students in exercises working with standards and criteria to appreciate how they can apply them to their own work. In the present study, the assumed motivational effects of having tutors' marks and feedback revealed immediately on completion of the self-assessment task meant that tutors could not have knowledge of student ratings before writing their initial comments. If tutors could provide information to help students' focus their judgements this might have greater impact on the development of self-assessment over time. Such a strategy, if undertaken after the initial set of tutor comments, would open up possibilities for the kinds of dialogic feedback proposed in recent publications (e.g. Carless, Salter, Yang, and Lam 2011; Boud and Molloy 2012).

Conclusion

Notwithstanding that this is an initial study with incomplete data biassed towards students enthusiastic in seeking to judge their own performance, there are interesting pointers to phenomena that if confirmed would have quite substantial pedagogic implications. The study addressed the question of whether student engagement in self-assessment over an extended period of time in a standards-based context could

help calibrate their judgement and make them more effective judges of their own work. The tutor data and the student self-assessments show no significant differences by the second assessment task in a unit of study which suggest that students' judgements converge with those of tutors. It was also found that this convergence was not evident when students began a new unit of study and that this more accurate judgement did not occur in an initial task in a new unit until students had had opportunities to practice self-assessment over three semesters.

This outcome is potentially important as it supports the notion that under appropriate conditions most students can improve their judgement skills. However, when students are categorised by achievement level, differences are found in students' ability to develop accurate judgement. High achievers are found to underestimate their ability whilst low achievers over-estimated. When the groupings were examined for performance over time, it was the accurate estimators (the mid achieving group) showing the highest level of improvement, with the under-estimators (high achievers) demonstrating some increased performance, but improvements were not evident in under-achievers. This has important implications for both educators and learners in appreciating the role of criteria and standards in assessment, and how understanding these elements needs to be fostered to develop students' judgement in order to support optimum performance.

Further research needs to be undertaken to explore the improvement in judgement skills in other settings. Such studies could usefully address conditions that more actively promote the development of self-assessment skills, and the kinds of intervention needed for lower-achieving students to show similar improvements to the majority of their cohort.

Notes on contributors

David Boud is a professor of Adult Education in the Faculty of Arts and Social Sciences, University of Technology, Sydney.

Romy Lawson is an associate dean (Teaching & Learning) in the Faculty of Law, Business and the Creative Arts at James Cook University. Her research areas include constructive alignment, assessment and assurance of learning. She has just completed a National Office for Learning and Teaching funded project to explore strategies for curriculum mapping and data collection for assuring learning.

Darrall G. Thompson is the director of Teaching and Learning and senior lecturer in the School of Design at the University of Technology, Sydney. He is currently completing PhD titled 'Design Thinking for Educational Change'.

References

Black, P., and D. Wiliam. 1998. "Assessment and Classroom Learning." *Assessment in Education* 5: 7–74.

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

Boud, D., and N. Falchikov. 1989. "Quantitative Studies of Student Self Assessment in Higher Education: A Critical Analysis of Findings." *Higher Education* 18: 529–549.

Boud, D. and N. Falchikov. 2007. "Developing Assessment for Informing Judgement." In *Rethinking Assessment for Higher Education: Learning for the Longer Term*, edited by D. Boud and N. Falchikov, 181–197. London: Routledge.

Boud, D., and E. Molloy. 2012. "Rethinking Models of Feedback for Learning: The Challenge of Design." *Assessment & Evaluation in Higher Education*, doi: 10.1080/02602938.2012.691462.

Carless, D., D. Salter, M. Yang, and J. Lam. 2011. "Developing Sustainable Feedback Practices." *Studies in Higher Education* 36 (5): 395–407.
Dochy, F., M. Segers, and D. M. A. Sluijsmans. 1999. "The Use of Self-, Peer and Co-Assessment in Higher Education: A Review." *Studies in Educational Evaluation* 24 (3): 331–350.
Dreyfus, H. L., and S. E. Dreyfus. 2005. "Expertise in Real World Contexts." *Organization Studies* 26 (5): 779–792.
Ericsson, K. A., R. T. Krampe, and C. Tesch-Romer. 1993. "The Role of Deliberate Practice in the Acquisition of Expert Performance." *Psychological Review* 100 (3): 363–406.
Falchikov, N., and D. Boud. 1989. "Student Self Assessment in Higher Education: A Meta-Analysis." *Review of Educational Research* 59 (4): 395–430.
Galbraith, R. M., R. E. Hawkins, and E. S. Holmboe. 2008. "Making Self-Assessment More Effective." *Journal of Continuing Education in the Health Professions* 28 (1): 20–24.
Hartog, P., and E. C. Rhodes. 1935. *An Examination of Examinations*. London: Macmillan.
Hartog, P., and E. C. Rhodes. 1936. *The Marks of Examiners*. London: Macmillan.
Lawson, R., T. Taylor, D. G. Thompson, L. Simpson, M. Freeman, L. Treleaven, and F. Rohde. 2012. "Engaging with Graduate Attributes through Encouraging Accurate Student Self-Assessment." *Asian Social Science* 8 (4): 3–12.
Lew, M. D. N., W. A. M. Alwis, and H. G. Schmidt. 2010. "Accuracy of Students' Self-Assessment and Their Beliefs about its Utility." *Assessment & Evaluation in Higher Education* 35 (2): 135–156.
Nicol, D. 2009. "Assessment for Learner Self-Regulation: Enhancing Achievement in the First Year using Learning Technologies." *Assessment & Evaluation in Higher Education* 34 (3): 335–352.
Nicol, D. J., and D. Macfarlane-Dick. 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., M. Price, and C. Rust. 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 (1): 119–144.
Sadler, D. R. 2005. "Interpretations of Criteria-Based Assessment and Grading in Higher Education." *Assessment & Evaluation in Higher Education* 30: 175–194.
Sadler, D. R. 2009. "Indeterminacy in the Use of Preset Criteria for Assessment and Grading in Higher Education." *Assessment & Evaluation in Higher Education* 34: 159–179.
Ward, M., L. Gruppen, and G. Regehr. 2002. "Measuring Self-Assessment: Current State of the Art." *Advances in Health Sciences Education* 7: 63–80.
Yorke, M. 2007. *Grading Student Achievement in Higher Education: Signals and Shortcomings*. London: Routledge.