International Journal of Learning, Teaching and Educational Research
Vol. 10, No. 1, pp. 142-156, January 2015

Oral Assessment and Grade Integrity: A Case Study

Vidar Gynnild
Norwegian University of Science and Technology (NTNU)
Trondheim, Norway

Abstract. The construct of "grade integrity" refers to the degree to which awarded grades accurately represent the breadth and depth of academic achievement (Sadler, 2009b). This case study examines issues of grade integrity at university level by studying a class of 26 students in science and engineering. While the use of two sets of oral examinations, a mid-term and a final examination, appears to be well justified, grade integrity is challenged by the allocation of scores acquired at the two assessment points, and by issues associated with assigning scores at the mid-term exam. The introduction of an assessment rubric increased consistency and inter-rater reliability at the final examination, but challenges still remain in enabling greater transparency in assessment practices. A key task is to incorporate institutional policies and assessment frameworks in grading, thereby empowering students by making them familiar with standards and criteria.

Keywords: oral assessment; grade integrity

Introduction

Written end-of-term examinations are still by far the most widespread mode of assessment at the subject university, yet oral assessments are also part of the assessment repertoire, notably in re-sit examinations and in courses with few students. In many instances, oral assessments are combined with traditional written examinations, portfolios and take-home examinations. These changes have largely taken place over the last decade and the major driving force has been the quest for improved learning and more authentic assessment practices. The efficacy of frequent examinations has been assumed, and such practices have been encouraged without examining potential, inherent flaws. Less effort has been invested in ensuring that practices are rooted in sound assessment principles, and that the representation of achievement is valid and reliable. While new assessment practices may have marginalised instances of "selective negligence" made by students, other concerns have emerged in particular related to the link between student achievement and symbolic expressions as seen in grade transcripts. In this study, an apparent concern is related to the provision of a robust assessment framework. It seems vital to ensure whether or not such guidelines exist as well as to examine the nature and cognitive depth of such guidelines when they do exist. We will examine potential use of such guidelines and explore in further depth existing grading practices.

This article draws on data collected from a case study of a science/engineering course at a research intensive institution, here called the subject university. Our study was conducted during a period of reform when professors were urged to adopt a broader range of assessment formats to promote learning and to solidify the evidence base of grades. In our course, the final examination had been replaced by two oral exams, a mid-term and a final, both counting towards the final grade. The underlying concern that informed our study did not pertain specifically to the quest for improved learning, but to the extent to which grades actually represented individual achievement, and not something else.

This study draws extensively on work of Royce Sadler (1989; 2005; 2009a; 2009b; 2010). Sadler was the first researcher to introduce the construct of grade integrity stipulating that each grade awarded should be "…strictly commensurate with the quality, breadth and depth of a student's performance" (Sadler, 2009b, p. 807). Sadler's contribution has been to address the value of grades from a conceptual perspective, which also motivated this study. In addressing some of those issues, we first draw a distinction between the intrinsic value of a grade, and the grading practice. The intrinsic value, or merit, is what grade integrity is about. In essence, this implies that a student's work should be graded according to its value without a view to the candidate's previous achievements, other students' performances or issues such as gender, age or ethnicity. Grade integrity is concerned with implications of criterion-based assessment, consistency and fairness as well as maintaining same value of grades within and across educational programs and institutions (Sadler, 2005).

Theoretical background

The centrality of assessment in education is well documented (Brown & Knight, 1994; Gynnild, 2001; Rust, 2002). Whilst historically the quality of teaching attracted the bulk of interest at the subject university, this has changed over the last two decades. In particular, different assessment formats and types of tasks have attracted increased interest as vehicles for improving student learning. The new slogan around the turn of the century was that of formative assessment, or assessment for learning as opposed to summative assessment. This linguistic distinction and the call for improved learning stimulated the introduction of a range of innovative assessment formats. Typical examples are continuous assessment formats which exclusively, or in combination with an end-of-term examination, counted towards the final grade. This wave of reform was justified by the quest for improved learning, while potentially troubled areas were ignored. The scholarship of assessment later made significant steps forward and helps us elucidate the remaining problematic. The provision of new concepts represents a valuable contribution to the scholarship of assessment as does the literature addressing the links between assessment theory and practice.

Three requirements are proposed for the aspiration of grade integrity to be realised: "assessment evidence [should be] of a logically legitimate type;

evidence [should be] of sufficient scope and soundness to allow for a strong inference to be drawn; and a grading principle [should be used] that is theoretically appropriate for coding the level of a student's performance (Sadler, 2009b, p. 807). Whilst there are convincing arguments for these propositions, demands for improved learning may overshadow issues such as the application of appropriate reference frameworks, and the need for sufficient sampling of tasks to make strong inferences about students' achievements.

Oral assessment is used in a variety of settings; however, relatively little research has been conducted on this format (Pearce & Lee, 2009). Most of what we know about oral assessment is from the examiners' perspective, and apart from anxiety studies, little is known about students' experiences (Joughin, 1998). Generally, oral assessment represents a flexible assessment format with a reputation for authenticity and content validity that may be hard to achieve in written assessments, particularly when communication or problem-solving skills are deficient (Joughin, 1998). Although interactivity generally carries positive connotations, it is not unproblematic in oral assessment. Some degree of unpredictability is always inherent in such sessions because of the potential interaction range between the presentation pole and the dialogue pole (Joughin, 1998, p. 371). The former resembles written examinations in that students respond largely uninterrupted, to set tasks. In the latter case, interactivity is pervasive, at the potential cost to the student of being able to address only a fraction of the pre-scheduled items. The uneven distribution of power between the student and the examiner may harm the progression of the questioning schedule and negatively affect the assigned grade. In our case, questions varied at both sets of exams. Exercises 1-3 had been assigned to the mid-term examination and exercises 4-10 had been assigned to the final examination. Believers in this format still argue that the pros far outweigh the cons: "If we want to truly know what students know about, what they can do in and how they are disposed towards their chosen field, at some point we must get them to talk to us!" (Joughin, 2011, p. 3).

Context and research questions

Oceanography is an elective in the third/fourth year of the Master of Science programme, with an estimated weighing of 7.5 credits from a total of 60 credits annually. The aim of the course was to help students understand the physical phenomena involved in the interaction between the earth's atmosphere and oceans. Learning objectives were created for the mid-term and the final examination, but the 20 minute sessions addressed only a selection of those objectives. Unlike written examinations, questions varied according to selections of content area, based on a draw made on entering the examination room, one taken from exercises 1-3 at the mid-term and one taken from exercises 4-10 at the final examination. Assigned scores accounted for 30 (mid-term) and 70 per cent (final) respectively towards the overall course grade. The syllabus was divided into three parts: wind, waves, and currents, each of which was fully covered in the textbook, as well as in lectures and three hours of weekly exercises. Students in engineering programmes typically have to complete such weekly tasks but in this course, approval of submitted work was not mandatory. The 26 persons on the course included Norwegian and international students. The latter group was questioned in English, whilst all native students were allowed to use their first language at both sets of examinations.

Letter grades (A, B, C, D, E and F) were used to represent achievement. The introduction of alphabetic grades was originally intended to allow translation from one set of grading symbols to another. However, the Ministry decided to adopt this grading scale for all higher education. An official memorandum from 2004 stated that grading should be based on descriptors, and that additional guidance would be provided for the application of this framework. There should be "no predefined distribution of grades …" (Johansson & Kjellemo, 2004, p. 1), thus indicating the adoption of a criterion based grading principle. Although the application of grade descriptors is a mandatory requirement, it was never used during the course of this study. Letter grades are purely nominal, but were converted into numerals for research purposes (A=5; B=4; C=3; D=2; E=1; F=0). Students' grades were allocated by assigning all aggregates according to institutionally predefined numerical ranges, as illustrated in Table 1:

Table 1: Aggregate score ranges used at the Subject University

A	B	C	D	E	F
100 - 90	89 - 80	79 - 60	59 - 50	49 - 40	39 - 0

The adoption of aggregate score ranges simplifies data management, "giving the impression of openness, objectivity and comparability across courses" (Sadler, 2009b, p. 815). However, the uneven score ranges still raise concerns related to the construct of grade integrity that will be discussed later in this article. The following research questions are explored: Is the grading principle appropriate for coding the level of a student's performance? (1) To what extent are assessment and grading practices compatible with grade integrity? (2) How could grade integrity possibly be improved in the course under scrutiny?(3)

Methodology and scoring procedures

The application of a case study "…is not a methodological choice, but a choice about what is to be studied" (Smith Macklin, 2007, p. 211). This approach is claimed to be appropriate for asking "how" and "why" questions (Yin, 1994) in the systematic examination of one or more events. The purpose of our study is to gain a deeper understanding of appraisal and grading in the selected course, and to identify issues of grade integrity to be addressed in the future. The focus is therefore on one particular element of the case being investigated. Hence, the choice of methodology is determined by our research questions, which also called for an intervention to promote grade integrity. We found principles of action research appropriate for our work as it involves learning about learning with a view to improved practice (Zuber-Skerritt, 2002).

Observational evidence of achievement was obtained through two sets of oral examinations, in which each candidate was assigned scores ranging from 0-100 by an external examiner and the professor teaching the course. The latter conducted the interrogation of all candidates while the external examiner observed and took notes, and only occasionally asked questions of clarification. Slightly different procedures were used for the allocation of scores at the two assessment points. Prior to the mid-term examination, no deliberations took place on the selection of potential themes and questions. The professor had assumed this responsibility for years and so could draw on his experience, without any explicit and shared assessment framework having been elaborated. First, both examiners assigned scores individually for each candidate, followed by a brief discussion between the two. Calibration of standards took place progressively. First, both graders assigned scores based on experience and tacit knowledge. Second, candidates' performances were compared, especially those who received near-identical scores. The professor set the standard for the allocation of scores, while the external examiner routinely adjusted his scores in events of a disparity of 10 percentage points or more. On completion of the examination, an average score was assigned based on the previously adjusted scores assigned by the external examiner and the professor.

Due to conflicting commitments, the external examiner serving at the mid-term exam was replaced by a younger post doc candidate at the final examination. This time the author served as an external consultant to assist with the implementation of a criterion-based grading framework, which in this course was an innovation. With the mid-term examination in mind, the purpose was now to introduce assessment rubrics, focused in content areas and transparent in terms of academic requirements. In theory, this would promote greater fairness in grading, and serve as a frame of reference for those who might later seek an explanation of their score. Content areas and weightings were negotiated before, rather than during, the examination. Scores were assigned out of a maximum potential score for each selected topic. Total scores were allocated as an average of assigned raw scores with no scope for progressive moderation.

Content areas addressed in the weekly exercises (4-10) served as the guiding principle for the rubrics. Selected content areas were covered in the textbook as well as in the set of exercises; however, the oral examinations featured larger emphasis on conceptual understanding and less weight on doing calculations. Table 2 portrays the rubric related to one of the selected content areas (waves). Similar, but not identical rubrics guided the questioning of candidates based on their individual draw of content area on entering the examination room.

Table 2: Excerpt of assessment rubric used at the final examination

To be Assessed (The "What")	**Expected Knowledge** (The "How Well")	**Scores and Comments on Performance**	
		Max Score	Assigned Score
Topic 1 Stokes waves, 2nd order • Properties compared with linear waves	$\zeta_A \sim \underbrace{a\cos(kx-\omega t)}_{linear} + \sim a^2\cos\left[2\left(kx-\omega t\right)\right]$ Free surface elevation; make a sketch $\omega^2 = gk\tanh kh$ From 3rd order: ω also depends on a • Higher crest, lower trough - crest increases as h decreases - trough decreases as h decreases • water particle motion; not closed particles as for linear waves	60	
Topic 2 • Solitary waves versus linear waves	• phase velocity $c_w = \sqrt{g(\zeta_A + h)}$ • sketch → - high and narrow crest - low and wide crest	15	
Topic 3 • Ursell number	$U_r = \frac{k\zeta_A}{(kh)^3} = \frac{nonlinearity}{dispersivity}$	10	
Topic 4 • Breaking criteria	• crest angle at max steepness 120° • $u = c_w$ • vertical acceleration downwards = $g/2$ • wave steepness $s = H/\lambda = 1/7$ Strictly properties of highest and steepest Stokes wave	15	
Total		**100**	

The quest for grade integrity has comprehensive implications, such as the application of an appropriate reference framework and the selection of grading standards and criteria. Attention has to be paid to the selection of content, ways of assigning scores, weighting of assessment items and allocation of grades. While institutional guidelines provided some basic information, such as an outline of the grading scale and suggested grade descriptors, the assessment panel enjoyed great freedom in adopting operational procedures for the oral. This format offers challenges due to the dialogic nature of the questioning, the uneven power distribution between the parties and the non-identical tasks to be addressed for the various candidates.

Analysis

Given the current procedure for determining mid-term scores, we examine how assigned scores depend on scores suggested by the professor and the external examiner. At the mid-term examination scores were assigned progressively by comparing and negotiating suggested scores suggested by the external examiner and the professor. The more experienced professor assumed a hegemonic role over the external examiner in deliberations taking place in the aftermath of each candidate's performance. In effect, this meant that the assessors contributed unequally to the agreed mid-term score. A regression analysis is used to explore the relationship between the independent variables (individually assigned scores) and the dependent variable (final mid-term score). This was estimated by a regression equation in which M is the final mid-term score; M1 is the professor's mid-score and M2 is the score of the external examiner: $M = 0.675M1 + 0.328M2$. The professor's score accounts for 67.5 % of the final mid-term score, while the external examiner accounts for 32.8 %. On average, the external examiner assigned higher scores than the professor, meaning that the average score at the mid-term would have been higher only if the average of the two sets of scores had been used. The same score pattern emerged at the final exam as the professor's average score was 72.88, with standard deviation (Sd) 31.75 while that of the external examiner's was 74.96 (Sd31.36). This time both examiners contributed almost equally towards each individual score.

As indicated, the sets of exams contributed unequally towards the final score, as indicated by the following equation (R = final total score; M = score at the mid-term; E = score at the final exam): $R = 0.3M + 0.7E$. Our goal is now to examine how final scores depend on initially suggested scores, at the mid-term as well as the final examination. This is achieved by conducting a regression analysis based on our data, and the regression equation is: $R = 0.201M1 + 0.0979M2 + 0.339E1 + 0.365E2$. R is the final score based on the mid-term and final examinations; M1 and M2 denote assigned scores by the professor and the external examiner at the mid-term examination, while E1 and E2 are scores assigned by the professor and the new external examiner respectively at the final examination.

It appears that the professor's scores account for about 54% towards the final scores, while the external examiner's score accounts for about 46%. The introduction of a grading rubric enabled a more consistent allocation of scores than those allocated at the mid-term examination. The rubric featured a set of agreed themes to be addressed, including specified questions and their respective weightings towards a composite score. This time, the assessors shared their interpretations of expectations and reached consensus before, rather than during, the examination. This resulted in increased inter-rater reliability and higher correlation of scores at the final exam compared with the mid-term; (E1, E2) = 0.99; (M1, M2) = 0.95.

Finally, we compare correlations of scores for three selected samples of students at the mid-term and the final exam respectively: Minimum score range [0-40], medium score range [40-80] and maximum score range [80-100]. Table 3 features average scores for the three samples at the mid-term (M1, M2) and the final examination (E1, E2) and final result (R).

Table 3: Means and correlations for three groups of students

	Minimum	Medium	Maximum
Mean (M)	27.00	57.56	92.83
Correlation (M1, M2)	0.53	0,59	0,82
Mean (E)	38.40	71.56	90.75
Correlation (E1, E2)	0.99	0.98	0.97
Mean (R)	35.20	67.49	91.41

While correlations of scores assigned at the final exam correlate well for the three categories of students, this is not the case at the mid-term. This time, the correlation is high for the sample of academically successful students (0.82) while the two remaining categories exhibit much poorer correlations (0.53 and 0.59) indicating that the negotiated criteria utilized at the final exam provided a more robust framework compared to the mid-term scoring system. On average, students' scores were poorer at mid-term compared with the final, implying that grades would have been better without the mid-term scores included.

The mid-term exam had been introduced to comply with requirements announced by the Ministry (KUF, 2001, p. 31-32) suggesting that teaching, assessment and grading should be combined to avoid cramming and facilitate progression. Later, the European Qualifications Framework (EQF) gave rise to concern over cumulative assessment practices by requiring time-framed learning objectives. Knowledge acquisition was to be demonstrated at the end of, not during the course. In this study, the formula for final scores marginalised the effects of the mid-term, which accounted for only 30% of the final grade. The impact of cumulative assessment may still affect the final grade. The sketch in Figure 1 illustrates issues of three hypothetical students in the same course.

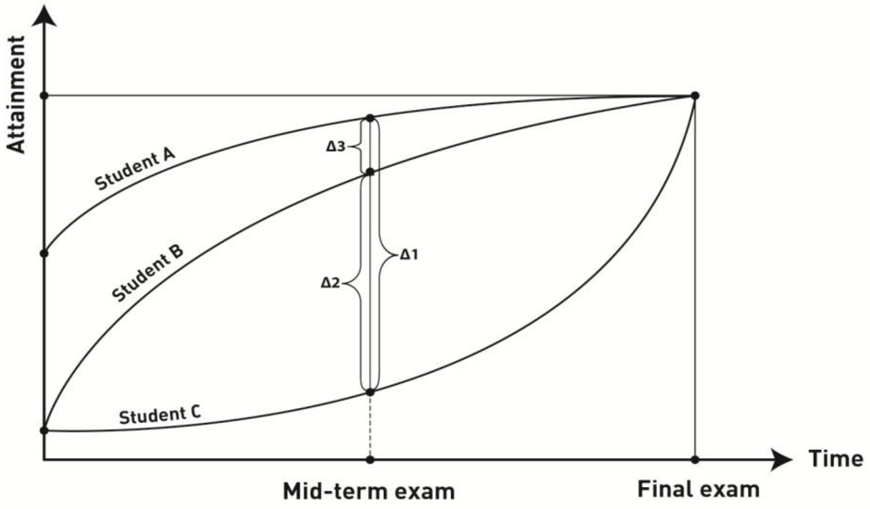

Figure 1: Attainment paths and impacts caused by the mid-term examination

Student A entered with higher qualifications than B and C, who had identical entrance qualifications, but pursued different learning paths. A, B and C achieved identical scores at the final examination, but due to scores assigned at the mid-term may not get the same grade. Differences in scores at the mid-term exam are marked as Δ1, Δ2 and Δ3. The final grade is of course affected by the weighting of any assessment point introduced throughout the semester.

We then explored effects of different weightings of the mid-term examination of the course. Table 4 illustrates the impact caused by differential weightings of the mid-term examination. As expected, the greater the weighting of the mid-term examination, the greater the distribution of the spectrum of grades. Two students lose A grades, and two others lose B grades due to the mid-term score, contradicting the popular notion that, as a rule, mid-term examinations benefit students by allowing them to focus on smaller areas of content at a time.

Table 4: Grade distributions as a function of differential weightings of mid-term scores

	A	B	C	D	E	F
Mid-term accounting for 0 % of final grade	11	5	6	0	0	4
Mid-term accounting for 15 % of final grade	10	5	7	0	0	4
Mid-term accounting for 30 % of final grade	9	3	10	0	0	4
Mid-term accounting for 50 % of final grade	9	3	8	2	0	4
Mid-term accounting for 100 % of final grade	10	2	3	4	3	4

Often, however, learning does not adhere to a linear path, but occurs in stages, sometimes towards the end of the course, as seen in recent research on threshold concepts (Land, Cousin, Meyer, & Davies, 2005). Deep learning takes time and normally requires a greater thematic focus and contextual overview than can be attained in a few weeks, during which there are time constraints due to other courses, part time jobs and other commitments.

Examination without tight adherence to set items compromises the rigour of oral assessment. In this study, examination questions as well as the order of topics were constructed in advance along with weightings of each assessment point (see Table 2). This added rigour to the examination in the sense that students were asked the same questions and assigned scores according to agreed criteria and standards. If students were unable to respond adequately, the professor intervened by posing scaffolding questions to help them back on track again. Students could neither select items, nor their order as seen in this quote:

"Obviously, I need to know the premises of the exam, e. g. that the duration is 20 minutes and that there are four questions to be posed. By comparison, you don't get access to questions one at a time in any written exam … There should be no other surprises other than the exam questions themselves. … This runs counter to the interests of the candidate unless the questions to be answered are exhibited to the student at the outset of the oral exam, or when there are questions that remain unanswered during the interrogation" (Student NN).

The chart in Table 5 exhibits the weighting of assessment items associated with exercises 4–10. The different thematic areas feature three or four tasks, but the weighting of items is not transparent to the student. While the order of tasks was decided by contextual logic, as seen in Table 4, task 1 in Ex4 and Ex5 carries a lot of weight, while this is different in Ex6 – Ex10.

Table 5: Weighting of tasks associated with content areas at the final exam

	Ex 4	Ex 5	Ex 6	Ex 7	Ex 8	Ex 9	Ex 10
Task 1	60	60	20	40	30	20	40
Task 2	15	15	10	30	40	60	40
Task 3	10	10	30	30	30	20	10
Task 4	15	15	40				10

General access to the entire set of questions to be posed at the exam, including their relative weighting, would empower students and increase reliability in grading by enabling them to respond to the entire spectrum of themes, rather than having decisions made on their behalf by the examiner. In this study, the problem is not only the inconsistency of questions asked to each student, but the examiners' assumed authority in deciding the order of questions, timing and progression. One argument in favour of the current practice is that the order of questions may be set by factors inherent in their subject matter. In such instances it could potentially add to the student's confusion and increase problems.

The algorithm for combining scores is also of interest to this study: "Aggregation is the process of combining a series of module scores, or degree classifications derived from such scores, into a final unique degree classification" (Morrison, Cowan, & Harte, 1997). Methods belong to one of two broad groupings, "namely, 'grade combination' or 'mark aggregation' methods, where the former involves a summation of grades, and the latter a summation of marks" (Morrison et al., 1997). Because the former method takes no account of students' raw marks, grade combination algorithms may be uninformative. This was not an issue in this study because global scores incorporated raw scores from both exams. However, the inclusion of scores assigned at the mid-term examination compromises the integrity of the grade, and final grades would have been more accurate without the inclusion of the mid-term scores.

Unlike random error, bias is systematic and is present in all forms of assessment, probably more so in oral assessments than in written work, which can be further checked at a later date. Although we have no evidence of any bias caused by ethnic, language or gender partiality, our data indicate that the external examiners were slightly more lenient in their grading than the professor (see Figure 2). This runs counter to one of the assumed responsibilities of the external examiner, namely to ensure academic standards and avoid grade inflation.

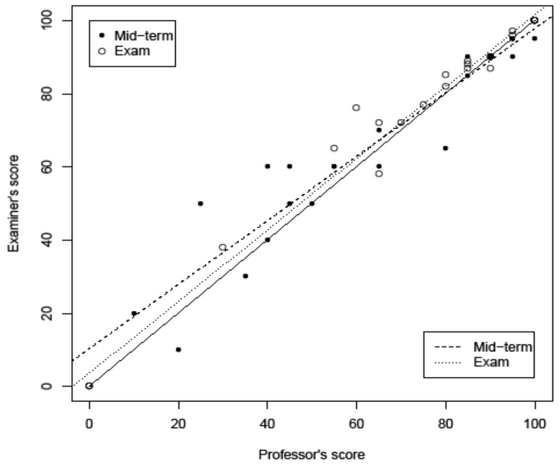

Figure 2: Professor's and examiner's scores at the mid-term and the final, including "best fit". Professor's scores on the horizontal line & Examiner's scores on the vertical line

However, this cannot be deemed as bias because there was no "official" rating of performance. The use of a scoring rubric at the final examination reduced the need for an external examiner compared with the holistic assessment at the mid-term examination. In the latter case, the calibration of scoring practices evolved progressively through reference to prior performances, while at the final this process was largely settled before the examination. This contributed to greater consistency in scoring and thereby to the integrity of the grade. While rubrics do not by themselves guarantee high levels of reliability in grading, they may be helpful in identifying and making explicit standards and criteria to be applied.

Yet another issue is to secure a match between individual achievements with accurate representation of these achievements. Misrepresentations can occur in a number of ways, e.g. in methods used to collect raw scores (norm-based or standards-based assessment), or in the conversion of scores into grades. In this study, grades were assigned according to score ranges laid out in institutional policies. In our study, the cut-off ranges have been adjusted to show a normal distribution curve. The wide score range for "C" (79-60) increases the likelihood of a bell-shaped curve, but compromises grade integrity. Students therefore appear to be graded unfairly at policy level. If the university were to retain its principle of standards-based grading, the implementation of equal score ranges for all categories of grades would increase grade integrity. In Table 6, this scenario is explored based on scores assigned in our course:

Table 6: Grade distributions according to two different scenarios of cut-off scores

Letter grades	A	B	C	D	E	F
Score ranges, scenario 1	100 - 90	89 - 80	79 - 60	59 - 50	49 - 40	39 - 0
Grade distribution 1	9	3	10	0	0	4
Score ranges, scenario 2	100 - 88	87 - 75	74 - 62	61- 49	48 - 36	35 - 0
Grade distribution 2	11	4	7	0	0	4

At first glance the application of score ranges gives the impression of accuracy and objectivity. However, these are not standardised units, and distributions ranging from 1-100 percentage points therefore do not tell anything about achievement in an absolute sense. The source of error is with selected score ranges as well as in judgments made by examiners. In this study, identical score ranges for all passing grades would have yielded more students receiving better grades. The use of cut-off scores facilitates the management of initially scores as the conversion into other grade distributions is purely a technical operation.

Discussion

With what certainty can we claim that the current assessment and grading design constitutes a sound and legitimate way to judge achievement? Some strengths are obvious, especially that the oral format enables comprehensive interactivity in the examination room. Oral examinations also effectively prevent common issues of academic integrity, such as plagiarism and unintended peer collaboration. Knowledge and reasoning may also be probed in greater depth than would have otherwise been possible. The setting allows the examiner to gauge skills and competencies progressively; however, the random selection of items of varying levels of difficulty may raise concerns. While assessment items in written exams are identical for all examinees, the asynchronous nature of oral assessment requires non-identical assessment items. This does not represent an issue by necessity as tasks may vary, yet still be of the same academic standard.

The brevity of the interrogation sessions may be more of as serious matter. With what certainty can we argue that our assessment evidence is of “sufficient scope and soundness to allow for strong inferences to be drawn in terms of scores and grades”? (Sadler, 2009a, p. 2). Given the theoretical emphasis of the curriculum, oral assessment offers opportunities that are non-existing in written formats. Candidates were allowed to reason and make drawings and calculations on the blackboard, and the professor helped clarifying potential misinterpretation of questions posed. A mixed design combining diverse assessment formats would have extended the evidence base grades, but would have been more resource intensive for the assessment panel, and for the students.

Examinations provoke anxiety and disadvantage students to varying degrees, but the rarity of oral examinations and the short time span in an unpredictable setting may have added to tensions before and during the interrogation session that may have misrepresented achievement. The same applies to language proficiency. English was the spoken language for international students, though not their first language, whereas Norwegian students were examined in their native tongue. It requires integrity of the examiners to differentiate between the “the student's command of the medium itself, i.e. the student's oral communication skills in general or language skills in particular; and the student's command of content as demonstrated through the oral medium” (Joughin, 1998, p. 367). The purpose of examinations is not to measure oral ability, but to test “cognitive knowledge, understanding, thinking processes, and capacity to communicate in relation to these” (Joughin, 1998, p. 368).

At the final examination, the assessors were largely in agreement for the entire range of the scale (0-100), while inter-rater agreement increased towards the upper end (70-100) at the mid-term examination This may be evidence of the effects of the shared interpretations of the rubric, enabling more accurate judgments of performance levels for the entire scale, not only towards the positive end. While in theory deficient language skills may impact graders (and grades), there is no evidence of such potential flaws in this study. An indicator of the robustness of assigned grades (and the integrity of the graders) is shown in Table 6. While policy decisions on score ranges by implication impact grade distributions (Table 5), data presented in Table 7 shows a relatively high degree of consistency in score allocation. The underlying attainment appears to be well represented in the assigned grades; however, as seen in Table 6, differences in grading still occur because assessment is not and can never be an exact science. Table 7 indicates that the professor contributes to upholding academic standards by consistently assigning poorer grades as compared with the external examiner.

Table 4: Grade distributions if assigned either by the professor or external examiner

Letter grades	A	B	C	D	E	F
Score ranges	100 - 90	89 - 80	79 - 60	59 - 50	49 - 40	39 - 0
Only professor	7	4	10	1	0	4
Only external examiner	8	5	9	0	1	3

How may grade integrity be improved in the selected course? Assessment and grading are socially constructed activities with a long history of discourse and collaborative inquiry. Grading panels share authentic challenges contributing to skills development and shared understandings within a community of practice (Wenger, 2000). While shared knowledge generally remain implicit in the act of completing tasks, they become operational in recognition of similar, if not identical tasks (Polanyi, 1958). "The shared repertoire in a community of practice refers to a pool of resources that members not only share but also to which they contribute on an ongoing basis" (Smith Macklin, 2007, p. 206). An apparent risk is that communities develop cultures featuring weak ties to institutional policies and the repertoire of theoretically based reference frameworks. Who knows if the integrity of grades is well taken care of across the range of such communities of practice? Neither students nor employers can be sure of the meaning of grades since so little is known about their theoretical underpinnings and the locally grown cultures that maintain and solidify established practices.

The beliefs, skills and knowledge acquired in appraisal of student achievement may be viewed as a social and cultural capital formed out of comprehensive experience and peer socialisation. This is the collective habitus of a community of practice (Jawitz, 2009, p. 604) which is the shared repertoire of beliefs, skills and knowledge associated with appraisal of work. The situation for a newcomer into the community has been described as a process of harmonisation "arising out of the interaction between the agency of new academics and these key contextual aspects within the workplace" (Jawitz, 2009, p. 605). It has been assumed that newcomers to communities of practice initially adopt a peripheral participation after which they progressively adopt shared beliefs and practices from the collective habitus. The uneven power distribution between a newcomer and the community of practice typically sets the direction of the harmonisation process, but this was reversed in this case.

The new external examiner was a post-doctoral student and completely new to the assessment and grading community. He assisted in the implementation of the scoring rubric, and negotiated interpretations of questions and weightings towards the final score. The increased consistency in scoring was clearly a contribution to the integrity of the grades. In this case; however, the construct of the harmonisation between the newcomer's individual agency and collective habitus does not quite fit in. The absence of the long-standing external examiner altered the conditions of the grading panel by making it more susceptible to new practices. The new examiner's cultural capital greatly supported the intervention because it was aligned with practices with which the new examiner was already familiar. The change was enabled by the abandonment of the existing collective habitus and guidelines suggested by Smith Macklin (2007, p. 212), which may also be helpful to other scholars in similar, if not identical, educational settings:

"Creating situations where the group needs to develop interpretations of their experiences (1); Structuring interactions so that these interpretations can be tested, assessed, extended, refined, rejected, or revised for a specific purpose (2); Providing tools (artefacts, symbols, and language) to facilitate the construction of ideas and models (3); Using formative feedback and consensus-building to develop and improve thinking (4)" (Smith Macklin, 2007, p. 212).

Undertaking this clearly presupposes professional external support, mutual trust and respect to degrees that cannot be expected to happen overnight. In the fortunate event that these qualities do exist, successful interventions may set examples for the dissemination of new practices.

Concluding remarks

Practices associated with assessment for learning have been widely embraced in higher education while issues of concern related to this approach are often left unattended. While administrative rules and regulations are provided by the university centrally, in practical settings examiners are often left alone to deal with issues of theoretical as well as practical nature in this area. Dealing with the quest for *grade integrity* is just but one example. This study has demonstrated some benefits of sharing own practices with and external expert, and the value of structuring interactions in order for new practices to be tested and revised. In this study, the impact of the external examiners was minimal, which supports the notion of the teaching professor as the prime carrier of assessment criteria and academic standards. The application of a rubric served as a useful frame of reference for the graders to ensure shared understandings of selected content areas, to harmonise expectations and reduce the risk of bias an error in assigning grades. Such efforts are likely to be of great significance to new assessors, and would help experienced examiners to become more reflective in their practice.

Acknowledgements

Thanks to Professor Dag Myrhaug, NTNU, for letting the author reproduce parts of his assessment rubric, as seen in Table 2.

Thanks also to Eirin Tangen Ostgaard for helping with the statistical analysis.

References

Brown, S., & Knight, P. (1994). *Assessing learners in higher education*. London: Kogan Page.

Gynnild, V. (2001). *Læringsorientert eller eksamensfokusert? Nærstudier av pedagogisk utviklingsarbeid i sivilingeniørstudiet.* Dr. philos, Fakultet for samfunnsvitenskap og teknologiledelse, Pedagogisk institutt, Norges teknisk-naturvitenskapelige universitet, Trondheim.

Jawitz, J. (2009). Learning in the academic workplace: the harmonization of the collective and individual habitus. *Studies in Higher Education, 34*(6), 601-614.

Johansson, T., & Kjellemo, B. T. (2004). Retningslinjer for bruk av det nasjonale karaktersystemet, from http://matematikkradet.no/dokumenter/2004-05-10_bokstavkar.pdf

Joughin, G. (1998). Dimensions of Oral Assessment. *Assessment & Evaluation in Higher Education, 23*(4), 367 - 378.

Joughin, G. (2011). "Speaking of which ...": The Intriguing Case of the Spoken Word in Assessment. *HERDSA news, 32*(8).

KUF. (2001). *St.meld. nr. 27 (2000-2001) Gjør din plikt - Krev din rett* Oslo: Retrieved from http://www.regjeringen.no/Rpub/STM/20002001/027/PDFA/STM200020010027000DDDPDFA.pdf.

Land, R., Cousin, G., Meyer, J. H. F., & Davies, P. (2005). Threshold concepts and troublesome knowledge (3)*: implications for course design and evaluation. In C. Rust (Ed.), *Improving Student Learning Diversity and Inclusivity*. Oxford: Oxford Centre for Staff and Learning Development.

Morrison, H., Cowan, P., & Harte, S. (1997). The Impact of Modular Aggregation on the Reliability of Final Degrees and the Transparency of European Credit Transfer. *Assessment & Evaluation in Higher Education, 22*(4), 405-417.

Polanyi, M. (1958). *Personal knowledge*. London: Routledge and Kegan Paul.

Rust, C. (2002). The impact of assessment on student learning. *Active Learning in Higher Education, 3*(3), 145-158.

Sadler, D. R. (1989). Formative Assessment and the Design of Instructional Systems. *Instructional Science, 18*(2), 119-144.

Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education, 30*(2), 175 - 194.

Sadler, D. R. (2009a). Fidelity as a precondition for integrity in grading academic achievement. *Assessment & Evaluation in Higher Education*.

Sadler, D. R. (2009b). Grade integrity and the representation of academic achievement. *Studies in Higher Education, 34*(7), 807 - 826.

Sadler, D. R. (2010). Assessment in higher education. In M. Baker, B. & Peterson, P. (Ed.), *International encyclopedia of education*. Oxford: Elsevier.

Smith Macklin, A. (2007). Communities of Practice. In G. M. Bodner & M. Orgill (Eds.), *Theoretical Frameworks for Research in Chemistry/Science Education* (pp. 204-227). New Jersey: Pearson Education.

Wenger, E. (1998). *Communities of practice: learning, meaning, and identity*. Cambridge: Cambridge University Press.

Wenger, E. (2000). Communities of Practice and Social Learning Systems. *Organization, 7*(2), 225-246.

Yin, R. K. (1994). *Case study research: design and methods*. Thousand Oaks, Calif.: Sage.

Zuber-Skerritt, O. (2002). The concept of action learning. *The Learning Organization, 9*(3), 114-124.