Exams as learning arena: A criterion-based system for justified marking, student feedback, and enhanced constructive alignment

C. Jørgensen, A. Goksøyr, K. L. Hjelle, and H. Linge, *University of Bergen, Norway*

ABSTRACT: A constructively aligned course will provide learning benefits for students. Armed with this insight, teachers often revise teaching/learning activities, only to realize that assessment and its backwash are the main obstacles to course alignment. Delving into the literature, teachers will be bewildered by terms such as measurement versus standards model, summative versus formative assessment, norm-based versus criteria-based grading (Biggs and Tang 2011), and disappointingly retort to traditional written exams graded using the well-established method of gut instinct. We therefore introduce a step-by-step method for criterion-based assessment, which extends grading rubrics with non-overlapping skill sets and draws on the SOLO taxonomy. Implementing the method can achieve desirable outcomes for student learning without having to know all the assessment theory. As side products, the method helps refine learning outcomes and produces individual student feedback. The first step is to define a set of 3-5 assessment criteria that cover all learning outcomes without overlap, and are made clear to the students before the exam. These may relate to: i) factual course knowledge; ii) clear language and argumentative style; iii) application of theory, concepts, analysis; and iv) ability to draw broader perspectives. For each assessment criterion, indicators of achieved learning outcomes are defined, with potential scores. For each of these, four levels of one-sentence comments are formulated: a) praise for great achievement (full score); b) a laudable attempt pointing to weakness or inconsistency (half score); c) mild critique explaining what the student should improve (no score); and d) criticism of direct mistake (no score). Marking then amounts to selecting appropriate verbal labels that describe the student's performance, the associated scores count towards the grade, and the descriptions are provided as written feedback to the students, making the exam part of the learning arena. We present a concrete example which is automated in a spreadsheet, and experiences obtained when transferring the method to courses in other disciplines. Because each point needs to be justified with an appropriate verbal description, grading becomes fairer. The method also requires the teacher to specify good student performance, which helps focus the course on its stated learning outcomes, or revise these, which in turn makes it easier to identify appropriate active learning components.

1 INTRODUCTION

In a traditional university course, students spend a lot of time preparing for a final exam, which will be graded by an expert in what can be the teacher's main time investment if the course is big. Typically, this is the only time the professional teacher sees the academic performance of the individual student, but despite big efforts from both students and teachers the consequence is often a single-letter grade with no explanation. For the student, the learning effect of the exam is likely marginal, and from a systems perspective it is clear that the huge efforts spent by both students and teachers could provide more learning were exams arranged differently.

Ideally this is not so in a constructively aligned course, where expected learning outcomes, teaching activities, and exams all line up to provide continual learning benefits for students (Biggs and Tang 2011). Armed with this insight, teachers often revise teaching and learning activities, only to realize that assessment and its backwash are the main obstacles to course alignment. Delving into the literature on assessments and exams, teachers will be bewildered by concept-pairs such as measurement versus standards model, summative versus formative assessment, norm-based versus criteria-based grading (reviewed for example in Biggs and Tang 2011, Chapter 10). Often these dichotomies are caricatured, exaggerated, and difficult to relate to one's own teaching situation, which in the end may lead many teachers to disappointingly retort to traditional written exams graded using

the well-established method of gut instinct, with uninformative single-letter grades provided to students.

Here we present a method for grading exams that moves beyond these obstacles and facilitates constructive alignment, based on rigorous use of grading rubrics. Grading rubrics are tables whereby several assessment criteria are broken down into descriptors of high, average, and low student performance, and are intended to operationalize the difficult task of selecting the appropriate grade based on transparent criteria (Reddy and Andrade 2010). Designing a good rubric is challenging (Popham 1997); consider a typical rubric provided by Biggs and Tang (2011, p. 240): Here assessment of an argumentative essay is broken down to separate scores for "Introduction", "Argument", "Summary and conclusions", and "References". A grade of B is suggested if the "Argument" has:

> *"Most/all relevant points drawn from mainstream literature; uses appropriate structure to resolve issues in convincing argument"*

What if the essay has a convincing argument but omits parts of the relevant literature, or vice versa? How should one then balance reading broadly and thinking well when setting the score? To acknowledge the problems, just think of the bright student who writes and thinks well but hasn't followed lectures or read – typically the essay can be described by elements spread all over the grading rubric. Which grade should then be assigned? How may the student learn from a description that does not describe exam performance? And why shouldn't students argue if part of their performance fulfils the description of a better grade? In many ways, rubrics such as this only forces one to apply gut feeling repeatedly, first for "Introduction", then "Argument", and so on. Similar critique can be formulated for the grading rubric in Reddy and Andrade (2010, their table 1).

Note that the description cited above (from the grading rubric in Biggs and Tang, 2011) compounds factual knowledge, logical structure, and clarity of language, and getting a score here will not help students realize which of these elements they mastered and which they should train more. It is a classroom reality that academic proficiency involves multiple skills, and that a student's performance needs not be correlated across the skill set. It then becomes even more important that assessment and feedback can provide the student with directions for which skill to focus on to efficiently improve overall academic performance. For both the examples of grading rubrics mentioned in the previous paragraph, poor writing skills would affect many of the points lost (Wellington and Osborne 2001), but it would not necessarily become clear to the student that it is writing that is the problem.

Our goal is to present a method for an elaborated grading rubric that makes grading more transparent, preparations more goal-directed, and provides students with feedback for improved learning.

2 METHODS

Briefly, our approach consists of specifying clear evaluation criteria that are communicated to students. For each criterion, the evaluator chooses among pre-defined sentences that best describe the student's performance using a large grading rubric in a spreadsheet. The positive descriptions are associated with points that are summed up towards the student's grade. Points and descriptions, both positive and negative, are then sent to each student as feedback.

2.1 Defining Non-Overlapping Evaluation Criteria

In a constructively aligned course the expected learning outcomes typically encompass skills beyond the encyclopaedic competence required to just regurgitate the curriculum. The learning outcomes can then be grouped according to the type of academic skill they pertain to (Table 1). For transparent grading and feedback, it is preferable that 3-4 competencies are evaluated. For each competence one evaluation criterion is specified with as little overlap as possible, and a maximum number of points assigned to attribute weighting towards the total grade.

2.2 Designing the Grading Rubric

Each selected evaluation criterion is then broken down into smaller attributes of good student performance and arranged in a rubric. Table 2 gives an example of part of a rubric for communicative competence. Based on experience, four levels of descriptions are desirable for each desirable attribute, ranging from outright praise (full score), via laudable attempt (half score) and mild critique of inconsistency/inaccuracy (no score) to direct critique of mistakes (no score). Each description should

Competence	Example Evaluation Criterion
Encyclopaedic	Know or repeat facts from the curriculum
Empirical	Collect observations or find and organize relevant data sets
Numerical	Perform computations, produce and interpret graphs and tables
Analytical	Apply learnt theories and logic to analyse and argue
Communicative	Communicate with unambiguous language and precise use of the discipline's jargon
Reading	Find, use, and reference relevant academic literature
Generalizing	See connections and perspectives across the curriculum and beyond

Table 1. Examples of academic skills or competencies that evaluation criteria could target, preferably with as little overlap as possible.

first describe the performance in a way that the student can recognize, then point to what the student should focus on for future learning (or make clear what the student has actually achieved if the performance is excellent).

Typically, the sum of potential points for an evaluation criterion could be 150-200 % of the maximum score for that evaluation criterion, i.e. if 10 points is max for communicative skills then attributes totalling 15 to 20 points should be described, with more potential points if the evaluation criterion is open such that students can excel in many different ways.

Attribute	Clear sentences	Using scientific terms	Defining scientific terms
Max points	2	2	2
Praise for outstanding performance (full score)	Sentences are clear and unambiguous, allowing you to demonstrate your thinking.	You use scientific terms with precision and where appropriate, which characterizes efficient scientific communication.	You define key terms and concepts so that it is easy to follow your reasoning.
Approval of laudable attempt, pointing to weakness or inconsistency (half score);	Sentences are for the most part clear but allow in some places for ambiguity, which may make it hard to demonstrate your thinking convincingly.	You mention some scientific terms and concepts but could use more of the scientific language you have learned, as it would allow you to communicate more precisely and efficiently.	You define some of the scientific terms and concepts you use, but you could demonstrate your reasoning more efficiently if you presented definitions more often.
Mild critique explaining what the student should improve (no score)	Some sentences are unclear, resulting in opaque and ambiguous language – you should reread every sentence critically to detect whether alternative interpretations may open for misunderstandings.	You rarely use concepts and definitions we have learned in the course, your communication would be more efficient and precise if you used scientific terms more actively.	You only define a few of the scientific terms you use – if you had defined more of them whoever reads your text can ascertain whether you have correctly understood key concepts.
Criticism of direct mistake (no score)	Your sentences can often be interpreted in multiple ways, including ways that are positively wrong – you should critically reread every sentence in isolation to check whether it can be misunderstood.	You sometimes use scientific terms imprecisely or wrong – remember that many words have a technical definition.	You sometimes misuse scientific terms – if you had defined more of the terms and concepts you use you would probably see when you apply them wrongly.

Table 2. Example of a small part of a grading rubric pertaining to written communicative skills. The blue text describes the student's writing in a way the student should recognize, while the red part puts the performance in context and should ideally be helpful for directing the student's future learning, either by maintaining good habits or acquiring new ones.

2.3 Assigning Descriptions to Student Performance

Evaluation now amounts to selecting the verbal descriptions that appropriately match the student's performance. For key attributes one could force oneself to choose one of the four descriptions, but for open assignments it may be that many columns will be left blank. It is more helpful for the student and the evaluator if some mild critique points to the weakest parts, even if those parts are relatively good. As critical descriptions do not come with negative points, one can always add critique if one thinks it would help the student to focus future learning. This does not go the other way: if performance is weak one should not add positive comments, as they would come with points that count towards the grade.

2.4 Automated Grading and Feedback

If the rubric is implemented in a spreadsheet, visual aids can guide the evaluator for example by highlighting the selected comments for any student, points can be summed, and the verbal descriptions automatically assembled into individual feedback that can be sent to each student using mail merge. By getting access to exactly the same qualitative descriptions the evaluator picked to grade the exam, students can compare the expert's comments with their own evaluation and learn from that. Although it remains secret exactly how those comments quantitatively relate to points or grade, it allows a qualitative alignment between the expert evaluator and the students' self-evaluation.

2.5 Rubric Design Concepts

Dawson (2015) lists fourteen design concepts one should think through when introducing grading rubrics. Our approach primarily deals with evaluative criteria, accompanying feedback information, quality levels, quality definitions, scoring strategy, and secrecy.

3 EXPERIENCES

3.1 Development for an Introductory Course in Biology

The method was first developed for a first-semester large enrolment course *BIO100 Introduction to ecology and evolution*, 10 ECTS. With growing student numbers (>200 students) and four exams during the course, grading became overwhelming. The majority of the students met higher education for the first time, and it quickly became clear that many approached university studies with inadequate strategies for academic learning as well as for demonstrating their skills. During the first two years of the course, individual feedback was written by the evaluator, but because many sentences kept repeating themselves, it was evident that a more systematic approach could be partly automated, save time, and increase course alignment and transparency towards students.

3.2 Transfer to an Introductory Course in Geology

The approach was adopted when redesigning an introductory-level course in Earth science, *Earth system history and geobiology*, 10 ECTS with ca. 60 students. The course consists of lectures, seminars, lab practicals, and a four-day field excursion with report. Assessment will be based on assignments and reports from seminars, practicals, and the excursion, supplemented with a final written exam.

3.3 Transfer to an Existing Bachelor-Level Course in Biology

The grading rubric from BIO100 was modified for an on-going course, *BIO260 Nordic cultural landscapes*, 10 ECTS. The course typically has 10–15 students and a written exam with traditional grading. The descriptive objectives, content, and learning outcomes were transformed into emphasizing four skills: i) course knowledge; ii) language and argumentative style; iii) application of theory, concepts, and analysis; and iv) ability to draw broader perspectives. The course consists of a series of lectures (32 h), two days field excursion, an essay assignment (ca. 3000 words), and oral presentation/discussion of the essays. Except for the lectures, all activities are mandatory. A grading rubric similar to Table 2 was developed and used for the essay; it was presented to the students prior to submission of essays and sent to each student as feedback. The goal was to prepare the student for future learning (if the performance was not excellent) and for the final exam.

3.4 Mistakes We Learnt From

In a first attempt in BIO100, the grading rubric was formulated with many atomic categories but with only two levels of descriptions. It quickly became apparent that none of the performances fit the descriptions of dream-like excellence or harsh critique that were prepared beforehand. This was rectified by toning down positive and negative aspects of the descriptions so they better fit with typical student performance. Still, a problem of qualitative resolution remained, which was resolved in the version presented here with four qualitative levels of descriptions for each category.

4 DISCUSSION

Through systemizing and automating grading rubrics used in assessment of student performance, our approach produces written feedback to individual students, thus making the exam with all associated efforts by students and evaluators part of the learning arena, and in a way that makes each individual student seen by the expert evaluator.

For the evaluator, each point counting towards the final grade needs to be justified with an appropriate verbal description, using standardized descriptors which are shared with the student. This implies that grading becomes fairer because descriptions have to match standardized criteria (commonly achieved with many applications of grading rubrics), but here transparency is also improved, in that the student can compare the expert's descriptions with their own performance and learn from that.

Especially when courses are designed around threshold concepts (Meyer and Land 2003), it becomes vital that students receive thorough feedback rapidly to quickly correct misconceptions and guide learning (Brown et al. 1997; Black and Wiliam 1998; Gibbs and Simpson 2004). The automated system here can speed up feedback, and provide directed learning tips for each student provided the descriptions are formulated as in Table 2.

The method also requires the teacher to specify good student performance prior to the exam, which helps focus the course on its stated learning outcomes, or revise these, which in turn makes it easier to identify appropriate active learning components and for students to focus learning.

REFERENCES

Biggs J, Tang C. 2011. Learning for quality learning at university, 4th ed. Maidenhead, UK: Open University Press.

Black P, Wiliam D. 1998. Assessment and classroom learning. *Assessment in Education*, **5**: 7-74.

Brown G, Bull J, Pendlebury M. 1997. *Assessing student learning in higher education*. London, UK: Routledge.

Dawson P. 2015. Assessment rubrics: towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education* **42**: 347-360.

Gibbs G, Simpson C. 2004. Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education*, **1**: 3-31.

Jonsson A, Svingby G. 2007. The use of scoring rubrics: reliability, validity and educational consequences. *Educational Research Review* **2**: 130–144.

Meyer J, Land R. 2003. Threshold concepts and troublesome knowledge: linkages to ways of thinking and practising within the disciplines. *ETL Occasional Report* **4**. Last accessed 28.01.2017 from http://www.etl.tla.ed.ac.uk/docs/ETLreport4.pdf

Popham J. 1997. What's wrong - and what's right - with rubrics. *Educational Leadership* **55**(2): 72–75.

Reddy YM, Andrade H. 2010. A Review of Rubric Use in Higher Education. *Assessment & Evaluation in Higher Education* **35**: 435–448.

Wellington J, Osborne J. 2001. *Language and literacy in science education*. Buckingham, UK: Open University Press.