Grading-- research and philosophy

There have been lots of questions in the department about online grading and testing. This has come up in teaching workshops I have run, so I had written up some material on it. In advance of the departmental discussion on this topic, I am offering a slightly modified version of that material for those who might find it useful. There are lots of issues involved, so it is kind of long, but the key points are underlined for quick scanning. The first part discusses the relevant research on exams and grading in science courses at the college level, and the second provides my own philosophy and practices for grading, guided by this research. I believe it minimizes the problems with cheating.

Research on grading and testing

The first thing to remember in thinking about grades is that grades at the university level have no objective meaning. They are completely arbitrary and defined by the individual faculty member. There is no process or standard that allows any comparison between what a B grade means in Prof X's course to what it means in Prof Y's course, even if those happen to be the same course offered in subsequent years, let alone different courses at different institutions. In the North American university system, great importance is placed on the right of individual Profs to be able to set grades however they choose. One implication is that, while faculty and students set great store in grades, they have little meaning beyond that perception.

There is a classic (and very long and detailed) review paper by Gibbs and Simpson on the research on college assessments and how they do or do not support student learning. I have done a two-page summary at http://www.cwsei.ubc.ca/resources/files/Assessment_That_Support_Learning.pdf in the cwsei instructor resources. They conclude that:

- Exam scores correlate very weakly with post graduate performance. ("Exam" being where students are sequestered in room without access to other resources and with a time limit.) For example, when researchers look at how good practicing physicians are at diagnosing patients, there is no correlation with their scores on the standard tests that play such an essential role in their getting into medical school.
- Scores on marked assignments are better predictors than exams of long-term learning retention.
- When assignments are a significant fraction of the course grade, the failure rates are 1/3 what they are when the course grade is based solely on exam scores.
- Students study and learn in more naïve ways when the grade is based solely on exams.

There are several reasons for this lack of correlation, the most important being that performing a task like solving an authentic physics research problem or diagnosing a complex patient disease, requires a certain set of cognitive skills that involve making a set of complex decisions. Learning to make these decisions, and hence the associated cognitive skills, is very much a learned activity, and that learning is quite specific to the cognitive skills involved. There is little overlap between the specific cognitive skills required by the typical exam question (primarily rapidly recognizing and applying a specific procedure) and those needed for authentic professional tasks. Authentic problem solving involves such skills as: recognizing and justifying appropriate simplifications or approximations, recognizing what information is needed and how to seek out that information, extensive planning and testing of the solution process, and when a possible solution is obtained, many specialized tests of its validity. In many contexts, the ability to collaborate effectively with others is also important (plus 19 other identifiable skills). The

typical exam problem requires the use of a very small set of these authentic problem-solving skills, and in many cases penalizes a student for utilizing such skills.

A second reason for the lack of correlation is that typical exam questions and how they are graded are quite idiosyncratic. Students frequently report they get better grades when they focus more on the instructor's idiosyncrasies in exam creation and grading, rather than on mastering the material. After I sampled a dozen or so final exams and saw how bad they were, I gained a better appreciation for this. Faculty almost never get any training on making a good exam and no feedback on the quality of their exams. This also explains how some students may justify cheating—if doing well on the exams is a meaningless barrier set by the peculiarities of the instructor and has nothing to do with actual learning, then any method of getting over this barrier is justified. In one research study, multiple faculty members from different institutions were asked to grade the same intro physics exam solutions, and their scores varied by nearly 100%. In the work of my own group, we have reviewed the grading of physics 41 finals and those of the comparable course at UBC, and we find that, using the standard rubric the TA s were to use, nearly half the exams are mis-graded.

A third factor affecting the correlation is that performance on high-stakes exams is affected by stress. Both the level of stress and how a student handles it have been shown to be influenced by social-psychological factors unrelated to mastery of the material, such as stereotype threat, and affect their exam performance. Given these three factors, it is not surprising that exam performance correlates well with performance on other exams, but not much else. Gibbs and Simpson find that, in general, the closer the assessment task is to authentic problem solving in the discipline, the better the correlation between assessment scores and post graduate performance. For example, grades on complex course projects that take a long time to complete correlate best with later job performance.

What the research shows is that, in terms of student learning (and usually happiness), frequent low stakes exams are better than a few high stakes exams. The low stakes exams are less stressful and provide much more useful ongoing feedback to the students so they can evaluate their level of mastery better and see what they need to do to improve. Although I know of no research testing this, such tests also likely reduce the tendency to cheat relative to high stakes exams. I discussed what is known about the educational negatives of grading on a curve in an early note sent to the departmental faculty, so I will not repeat here. There I also discussed how it is surprisingly easy to create open book exams that are appropriately difficult.

A second important thing to remember about grading is that, by the time students have been admitted into a good university, they have been conditioned to believe that grades define what is important, and so an activity is seen to be important for learning *only* if it is associated with a grade. While unfortunate, it is important to remember this it is not a moral failing of the students, rather it is something they have been taught and selected for by the educational system. If they did not thoroughly embrace this belief, they would never have been admitted to Stanford.

The main points relevant to deciding on how to handle testing and grading either in-person or online are: 1) grading is quite arbitrary, 2) traditional high stakes exams measure little of long term value, and 3) students study efforts are driven by grading policies.

Wieman grading method and avoiding cheating

Recognizing that grading is entirely a matter of individual philosophy and preferences, here are my own. My approach has the advantage that it works just as well for online as for in-person teaching, and I have seen very little indication of cheating. I base grades on a point system with many different components of the course contributing, essentially rewarding students for everything they might do that will contribute to their learning, including preparing for class, attending and participating actively in class, homework, exams, follow-up review of their mistakes, etc. Grades determined in this way are likely better predictors of future student success than grades on any exam I create. The points they receive from each contribution depends somewhat on the effort required, but, contrary to expectations, students are not very discriminating as to the point value versus the time required. If there are points attached they do it. Many of these items are not graded for quality, only that they are done.

I do include exams, with emphasis on their educational value, which is primarily from the studying students do and the feedback they receive from the results. The exams are relatively short, frequent. and low stakes, and they are open everything (book, notes, ...). To ensure students carefully review what they got right and wrong and learn from it, they can earn back a substantial fraction (has varied between 30-50%) of the points they missed on a problem by explaining how they were thinking incorrectly in doing the problem, and how they could avoid making that error in the future. I also tell them exactly what will be on the exams, in that I give them a detailed set of learning goals (~60) that lists what they will be expected to do in operational terms upon successful completion of the course. Those goals are clearly reflected in the homework problems and in every exam question. The exam questions also require them to make their thinking explicit, rather than just executing a procedure "Explain in your own words why this equation ... has a 1/v^2 in it." "Give three examples of how a classical interpretation of the light induced emission of electrons from metal surfaces fails to match observations." My students' exam scores (typically averaging in the mid to upper 70%) suggest that if they are cheating, they are not doing a very good job of it, but also, having to put something in their own words makes it quite easy to detect cheating. I did have one case where two students had very similar (incorrect) wording on an answer, and when confronted they confessed, but that is out of hundreds of students. More importantly, I believe that there is a combination of factors that reduces student inclination to cheat: the exams are relatively low stakes; they perceive them as valid measures of their learning and doing well is not dependent on guessing what is in instructors mind or memorizing some easily looked up piece of information; and most importantly, they are providing a clear educational benefit that the students recognize. If, in spite of all this, there are some students still want to cheat, I am just not going to worry about it, it is clearly not a very high percentage and not that important.

As a note added for the special situation of Stanford. It seems like the discussion about cheating on online exams is a bit bizarre, in that faculty, and perhaps students as well, are saying there will be widespread cheating, but no monitoring is allowed because of the honor code. That makes no sense. There is either an honor code set by the students which is followed and so there is little cheating, or there is not an honor code. In the latter case, one can just use one of the well-established monitoring procedures for online exams. But this should be considered a problem for the students to deal with, not the faculty, as it is their honor code setting the rules, and ultimately, students are the ones that will suffer from cheating, not the faculty.