Grading More Accurately

Mark Carl Rom

Georgetown University
Grades matter. Students care about them, for reasons both objective and subjective. Objectively, grades influence a wide variety of resources such as access to scholarships, awards, advanced education and employment. Grades are also often seen as the personal judgment of the professor upon the student: we all, perhaps, remember the warm glow of a “Well done! Fine work!” as well as the sting of a “Not up to expectations.” Moreover, a substantial body of research demonstrates that how students learn, and what they learn, is profoundly affected by how they are assessed (see for example Carnoy and Loeb 2002; Rust 2002; Brown and Knight 1994; Brown 1997; Ramsden 1992; Gibbs 1992). Professors, recognizing the personal and professional consequences of the marks they award, undoubtedly also care about grades. Although few professors likely enjoy grading, most professors take it seriously, and seek to assign the ‘right’ grade to each student.

But what is the right grade? I suspect that most professors develop their grading schemes through an individual trial-and-error process designed to satisfy their own preferences and minimize student complaints. In some ways this is understandable: in our graduate education we typically develop research skills and substantive knowledge, but we rarely obtain systematic training about how to teach, with grading one of teaching’s central tasks. What knowledge we do obtain about grading tends to be ad hoc, to say the least. This is understandable; so far as I know, at the collegiate level there is no professional recognition of, and precious few incentives for, developing and using sound student assessment systems. But what is understandable is not exactly defensible. This is especially true for those of us privileged to teach political science at the undergraduate
and graduate levels, where we often strive to provide careful analytical guidance to our students on how to conduct high quality empirical research. In the classroom, however, if we do not apply the same efforts to our own grading systems, then we do not practice what we preach.

This essay focuses on one factor that contributes to high quality grading systems: grading accuracy. I proceed in several steps. First, I discuss the elements of accurate grading. Next, I present analytical results indicating how often our grading schemes are likely to be inaccurate, and I also discuss the causes for grading inaccuracy. The following sections describe ways to make our grading systems more accurate. Finally, I offer data from my own ‘natural’ experiment in grading and offer some concluding comments.

GRADING ACCURACY

Student assessment systems can seek to accomplish many goals. Although there is little consensus on the purposes of grading (as to whether grades can and should be used to motivate students, for example, or even whether grades should be used at all), it should not be controversial to claim that, if grades are used, they should be as accurately tied to performance as possible.

Grades represent estimates of student ‘achievement’ – an unknown parameter. In statistical terms, an ‘estimator’ is the function used to produce the estimates of this parameter. Good estimators are, in general, accurate (technically, the term of art is ‘efficient’), unbiased, and consistent (see for example Kennedy 2001: 11-20). An estimator is more accurate than another if it produces estimates of the parameter that have smaller variances than an alternative estimator; that is, the better estimator yields estimates that are less variable. An estimator is unbiased if it yields estimates that, on
average, do not deviate systematically from the parameter; that is, the expected value of the estimate equals the parameter. An estimator is consistent if, as the sample sizes upon which the estimates are based grow large, the estimates approach the parameter.

Quantitative researchers devote substantial effort to ensure that the estimators they use are accurate, unbiased and consistent. If researchers neglected to use high quality estimators, their research would surely be discredited. It is thus ironic that professors do not always expend as much effort in making sure that the estimators, i.e., grading systems, that they use to estimate student performance are similarly accurate, unbiased, and consistent.

In this paper, I focus on the first key element of high quality estimators: accuracy. I do so because the third element, consistency, will generally exist if the estimators are accurate and unbiased. Moreover, although grading systems are fraught with bias, substantial research has been conducted on the sources of grading bias and potential remedies for it (e.g., Daly and Dickson-Markham 1982; Fleming 1999; Hales and Tokar 1975; Hughes, Keeling and Tuck 1980; Spear 1984; Sweedler-Brown 1992); I also consider the sources and remedies for bias in a separate paper (Anonymous 2010). Some scholars (e.g., Ebel 1969; Cresswell 1986) have examined the closely related terms reliability and unreliability, but they have not fully identified the sources of or remedies for unreliability. At any rate, these articles are unlikely to be commonly accessed by the political science community, as they tend to appear in general education journals.

**Accuracy**

Good grading systems assign scores that accurately reflect a student’s performance. For our purposes, let us assume that there is a ‘true’ grade on a given assignment that
accurately reflects this performance. Presumably, professors should want to assign grades that are true. But unless professors take steps to ensure this, it is likely that they will dispense false scores.

Consider the following scenario. Let us assume a professor has two students whose ‘true’ academic performance differs by 4 points, with Student A-’s ‘true’ score being, say, 92 and Student B+’s ‘true’ score being 88.² Student A- thus is ranked higher than student B+. Let us also specify that both students’ true grades are approximately at the midpoint of the scale for their grades, so that for example a B+ is awarded to those with scores between 86 and 89.99, while an A- is assigned for those scoring between 90 and 93.99.

Now, let us also make the assumption that each student’s actual performance on any given assignment fluctuates randomly around the student’s true mean, with the random fluctuations having a standard deviation of 2 points.³ Finally, let us assume that the professor on average scores the actual assignments correctly, but the professor’s scores also randomly fluctuate around the true mean, with a standard deviation of 2 points.⁴ These two sources of random variation are additive, so the overall standard deviation in scores would be 4 points.

Given these assumptions, two questions can be asked. What is the probability that each will receive the correct letter grade? Whatever their grades, what is the probability that Student A- will in fact receive a higher score than Student B+?

The answers are troublesome. Each student would receive the ‘correct’ grade on any given assignment only about 40 percent of the time; in 30 percent of the assignments they would obtain a grade lower than their true grade, and 30 percent of the time they would
get a higher grade. Moreover, Student B+ would receive a higher score than Student A-
approximately 23 percent of the time.

The likelihood that a student would receive an incorrect grade is a function of the
position of the student’s true score on the grading scale, the amount of random
fluctuation in these scores, and the number of scores the student receives. Figure 1 shows
the probability of an incorrect grade on a single assignment for grades subject to various
amounts of random fluctuation. The horizontal axis shows how far the student is from
the next grade, with the distances ranging from a student ‘on the edge’, only 0.1 point
away from the next grade, to one ‘in the middle’ of a four point range. Each bar
represents a student with differing amounts of random fluctuations in their scoring,
ranging from a standard deviation of one point to four points.

Two implications are worth highlighting. Most obviously, for students ‘on the edge’ the
likelihood of receiving an incorrect grade is quite high, ranging from 46 to 75 percent;
that is, most students close to the cut point have less -- sometimes much less -- than a
coin flip’s chance of receiving the correct grade. More importantly, the critical element
in whether or not scores are correctly assigned concerns the size of the random
fluctuations. For students having scores with a standard deviation of three points or
more, the probability that they will receive an incorrect grade is always more than 50
percent. In short: when random fluctuations are even moderately high, the likelihood of
incorrect grades is enormous.

The probability of awarding an incorrect grade for the course as a whole also depends on
the number of scores the student receives and the random fluctuation in scores. Figure 2
shows these probabilities for a student with true scores in the middle of a grade scale,
given various amounts of random fluctuation in the scores assigned. The key conclusions are that the likelihood of awarding a correct score increases with the number of scores assigned, but decreases with the size of random fluctuations. If the fluctuations are small then four assignments or more will reduce the probability of awarding an incorrect grade to less than ten percent; for ten assignments, the probability of awarding an incorrect score is negligible. If the random fluctuations are higher, however, the probability of awarding an incorrect score remain fairly large: if the random fluctuations are more than four points, even scoring ten assignments will lead to incorrect grades more than twenty percent of the time.

**Plausibility of Assumptions**

If the assumptions specified above are more pessimistic than warranted, then the problems in assigning correct scores might not be too severe. But, if anything, these assumptions perhaps understate the frequency of grading errors.

The first assumption is that the differences in true scores across students are relatively small. In the student population as a whole, it is probably true that there are rather large differences across students. But at many individual schools and classes the differences across students are likely to be modest. At selective universities, or in upper-division classes, student performance is likely to be rather compressed. At Georgetown University, for example, only about 20 percent of applicants for undergraduate education are admitted. The admitted students, on average, are in the top six percent of their high school classes and have GPA averages of 3.9. Within Georgetown College average SAT scores are about 1400, placing these students in the top six percent of those who take the
SAT (Inside Admissions n.d.; CollegeBoard.com n.d.). Virtually all Georgetown are, quite literally, above average.

But the problem of closely-matched students is not unique to these circumstances. It is commonly believed that student capacities and performances are normally distributed—that is, that they follow a bell-shaped curve. If this is true, and grades are assigned ‘on the curve’, this guarantees that many students will very close to the grading cut-points. Such students, understandably, are likely to be quite concerned about their grades given the emotional and consequential differences between, say, a D+ and C- or B+ and A-.

The second assumption concerns the size of the random fluctuations in scores. Regarding the students, we do have some evidence regarding the minimal amount of variability that is likely to exist. Random fluctuations in student scores have been studied on common, well-designed, well-explained national tests such as the SAT, GRE and GMAT. One study noted that, for students who take the GMAT twice with no additional learning between tests, scores fluctuate by about 0.5 percent (Rudner 2005: 8) The GMAT is explicitly designed to produce accurate scores: it has many questions and the format is well known, so that students can learn with fair precision what they will be expected to do and how they will be scored.9 It seems unlikely that any professor’s grading scheme would produce such little intrapersonal variation in student performance. Still, the amount of random variation in a single student’s GMAT scores serves as a benchmark for what a well-designed grading scheme can aspire to in terms of accuracy.

No systematic studies on the variability in intra-professorial scoring apparently exist. An adequate test of the variation in assigned scores would be difficult to construct, as it would require data from professors who grade the same assignments twice without
realizing that they are doing so. Anecdotally, in my own classes I have occasionally and inadvertently graded the same papers twice, and my scores seem to vary by a couple of points, but further study is needed on this matter.

It is clear that professors often score only a very small number of assignments. I surveyed the 38 versions of Department of Government syllabi posted on the Georgetown University website in the spring of 2006 (Georgetown University, n.d.). On average, the typical syllabus listed under four graded assignments (mean = 3.87, standard deviation = 1.22). Fully seventeen of the courses had three or fewer graded assignments. When such a small number of assignments are graded, the chances for incorrect grades are quite high.

To supplement this small, non-random sample of syllabi, I collected a larger random sample of 60 undergraduate course syllabi from a random sample of six top-ranked political science departments. On average, each course had just over five graded assignments (mean = 5.37, standard deviation = 2.40); exactly half of the syllabi reported four or fewer graded assignments. Exams counted for a majority of the overall grades (mean = 52 percent) but a minority of the assignments (mean = 1.97). At least among the highest-ranked political departments, the typical number of graded assignments is modest, leaving the substantial possibility that incorrect grades will be given.

The specifics will vary from student to student, professor to professor, class to class, and university to university, but it seems reasonable to infer that, on average, a student in the midrange of a four point grade scale, graded on 5 assignments over the course of the semester, would have something like a 30 percent chance of receiving an incorrect grade. Whether or not this is unacceptably high is, of course, a normative matter, but I doubt
that many organizations would consider a 30 percent failure rate on a critical task worthy of an ‘A grade.

**The Causes of Random Fluctuations in Scores**

Before considering the ways to reduce incorrect scoring, it is worth paying more careful attention to the sources of the random fluctuations in the students’ and professors’ scores. For both groups, it seems that the fluctuations can be divided into two components: those having to do with the individual, and those having to do with the grading system itself.

For the students, fluctuations around the ‘true’ grade will involve the normal vicissitudes of college life: the ability to devote time, attention, and talents to the task. Students will occasionally score above their true average when they devote exceptional amounts of these attributes to the work, for example; other times, they will score lower. A good grading system – or more broadly, a sound educational strategy -- will perhaps have only a modest impact on these features, although one might imagine that a sufficiently motivational grading system could suppress the variation in scores, primarily by reducing the amount of ‘less than true’ grades due to these individual characteristics.

More problematic are the fluctuations caused by the grading system itself. One aspect of a poor grading system is that students are left puzzled by what constitutes good performance. If students are baffled, they are left ‘shooting at an uncertain target’. Sometimes they might get lucky, guess right, and score higher than their true competence; other times, they’ll score lower. Such random variation is conceptually independent of student attributes although, perhaps, one definition of a ‘good’ student is one who is better at guessing what the professor values.
Professors can suffer the same individual and systemic fluctuations. Professors themselves might not always devote the same time, attention, and talents to grading a given set of work. For example, student papers that are read too quickly, or under too much fatigue, might lead the professor to mark some too generously and others too stingily.

The variation in professorial scoring that can be attributed to the grading system itself is more troublesome. Here, the professor does not know precisely what the grading standards are: reading a paper, the professor is left guessing how to evaluate it in assigning it a grade. I cannot estimate how big a problem this is in general, but I do realize that I have sometimes given assignments not knowing exactly what I expected, and consequently I was left guessing how to score the work. I suspect that the random variation in scoring in such cases is unacceptably high.

It does appear that professors rarely articulate what their grading standards are, at least not in their syllabi. Of the Georgetown University Department of Government syllabi surveyed, all provided extensive reading lists, but only one of the syllabi provided explicit standards regarding what scores constituted which grades, and 29 of the 38 offered no guidance whatsoever regarding performance standards. The nine syllabi that did offer some advice did so only in a quite limited way, usually concerning class participation. No syllabus offered specific counsel on expectations for performance across the required assignments. It may be the case that the professors distributed other guidance regarding assessment standards, but it seems odd how little advice was incorporated in the courses’ key reference document. Future research might fruitfully examine the extent and quality of the performance standards typically distributed by
political science faculty; my sense is that they are likely to be less common and more *ad hoc* than ideal.

The randomness of grading can have several negative consequences for the students. It can reduce the motivation to work. (“If grades are largely random, then why should I devote my efforts to get high grades? This makes as much sense as working hard to pick the right lottery numbers.”) It can leave them puzzled as to what constitutes high-quality performance. (“I wrote two papers of similar quality, but one professor gave me an A and the other a C. What gives?”) It can lead the student to suspect that bias (rather than randomness) is at work. (“I usually get an A for work like this, but that professor gave me a C. That professor must be biased against me.”) Such perceptions seem likely to be widespread:

I frequently [ask] educators to raise their hands if they have ever received a grade that was a ‘flagrantly inaccurate representation of their achievement in a course of study.’ Virtually all of the thousands of teachers to whom I have posed this question have raised their hands. I then ask ‘How many believe that the grades you received in school were not an accurate representation of your scholarly achievement?’ Sometimes as many as 50 percent of the educators in my workshops respond affirmatively. I find this an amazing commentary on our system of grading – even those within education have little confidence in the current system’s validity (Marzano 2000: 8)
IMPROVING ACCURACY

Within the grading scheme, the main factors that influence the frequency of incorrect grades are the number of scores assigned and the random variation in scores on each assignment. Fewer scores and more variation lead to more incorrect scores, and vice versa. The solutions are obvious, though not always easy: increase the number of scores, and reduce the random fluctuations in scoring.

*Increasing the Number of Scores: More Work Assessed*

Given random fluctuations in grading, more graded assignments will produce more ‘true’ grades than will fewer scores. Professors may nonetheless be reluctant to require additional assignments, for reasons both principled and pragmatic.

A case could be made that requiring only a small number of ‘high stakes’ assignments is educationally better because it allows students to focus their efforts on big tasks rather than be distracted by multiple smaller assignments. This rationale is questionable. After all, even big projects – say, a semester long research paper – can be decomposed into smaller, discrete elements. Preparation for a comprehensive final exam could include smaller tests given over the course of the semester. Another possible basis for having a few big assignments is that this trains students for work in the ‘real world.’ Most jobs involve frequent, smaller tasks, however. No lawyer would argue before the Supreme Court without preparing many draft briefs, and having these briefs subjected to critical scrutiny. All teachers create many lectures. Even less compelling is the argument that few assignments are assigned because, well, few assignments are customary: that’s the way we have always taught courses. For any teacher in the sciences to relay on tradition as a guiding principle is, at least, paradoxical.
Principled rationales aside, the main reason that professors are probably disinclined to require more assignments is that doing so requires more work, both for the students and for the professors. Neither group has much incentive to seek this. Moral arguments for professors to ‘do the right thing’ by giving more assignments can be expected to have limited value, and it is unlikely that students themselves will request additional work.

Empirical arguments may possibly be more persuasive. If professors can be convinced that a small number of assignments poses severe threats to posting ‘true’ grades, then it should be more difficult for responsible professors to resist requiring them.

Fortunately, more work can be assessed with minimal increases in effort. Two possibilities might be suggested. First, when possible and appropriate, professors can rely on increasingly available computer-graded assignments. For example, in my “Introduction to the U.S. Political System” course the students take sixteen on-line multiple-choice quizzes, of which the twelve highest scores are counted towards their overall grade. These quizzes come straight from the textbook databank; they are automatically graded; the students can immediately see which questions they answered correctly and incorrectly; the scores are instantly entered into their Blackboard grade book. These exams allow me to test a broad array of factual knowledge with minimal extra effort, while greatly increasing the total number of assignments scored.14 A second possibility is to require more ‘interim’ assignments contributing to a final project. For instance, in my undergraduate “Scope and Methods in Political Science”, the main task is to write a final research paper. Along the way, the students must produce three preliminary research reports, as well as four statistical problem sets.15 As a result, both of
these classes have nine or more graded elements, which substantially reduces grading errors relative to a class with five assignments, *ceteris paribus*.

**Increasing the Number of Scores: More Scores Generated**

Requiring more assignments is not the only way to produce more accurate grades, however. One alternative is to increase the number of scores produced for existing assignments. Assume, for instance, that numerous graders are available and that each grader assigns grades correctly on average, with the random fluctuations having a standard deviation of three percent. With a single grader, 95 percent of the time we would expect the correct score to be assigned, plus or minus 6 points. If two graders average their scores, in 95 percent of the cases the correct score would be assigned, plus or minus 4.25 points. With three graders, the margin of error is 3.5 points; with ten graders, the margin of error is 1.9 points; with 20 graders the margin of error falls to 1.3 points. Thus, with even a modest number of additional raters, errors due to random fluctuations could be substantially reduced.

It is most unlikely that a professor will have twenty teaching assistants, with each of them grading all the assignments in order to generate more accurate scores. But there is another way to increase the number of assessments of individual assignments: engage the students in the assessment process by having them dispense scores.

There is a substantial literature on the benefits and risks of student scoring (i.e., ‘peer assessment’; for one review, see Topping, 1998). As a conceptual matter, peer assessment can improve grading accuracy to the extent that the peer scores are themselves accurate and unbiased. One might expect that peer scores would have greater random fluctuations than the professor’s, however, mitigating the benefits of using
additional scorers. This risk can be reduced if the professor is sufficiently explicit in presenting the standards to be used in scoring which, to justify the professor’s own grades, the professor should be. As one scholar puts it:

Peer assessments become more valid as they are based on a larger number of observations and a greater number of dimensions of skill. They are also most helpful when standards are clear and more than one peer provides an assessment. Peer assessment exercises are also enhanced if instructors communicate the purpose of the exercises clearly, articulate the dimensions of judgment clearly, provide training when necessary, and monitor students’ evaluations, intervening when they are too harsh or too lenient (Norcini, 2003).

**Increasing the Clarity of Performance Criteria**

The impact of random fluctuations on grading accuracy can be reduced by increasing the number of assignments scored. But increasing the number of scores alone will have only a modest effect on accuracy if the random fluctuations in the scores are large. To ensure greater grading accuracy, these random fluctuations must be reduced.

Let us briefly recall the importance of reducing random fluctuations (Figure 2). Given the assumptions listed above, if the random fluctuation is three points the probability of an incorrect grade declines from 28 percent to 16 percent when the number of scores is increased from four to six, and to 10 percent when eight scores are used. On the other hand, if four assignments are scored, the probability of an incorrect score declines from 28 percent to 14 percent of the random fluctuations are reduced from three points to two points, and then to three percent if the random fluctuation is a single point. Reducing
random fluctuations thus has a more powerful impact on improving accuracy than simply increasing the number of assignments.

Perhaps the best way to reduce random fluctuations in scoring is to provide the students clear and specific guidance as to what constitutes high quality performance. Providing such guidance contains three main elements. First, the professor must indicate what criteria are relevant in assessment. Second, the students should be given a ‘performance rubric’ which indicates how the different criteria should be scored. These rubrics “must seek a balance between detail and practicality” (Guskey and Bailey 2001: 148; see also Linn and Gronlund 2000: 377-404). A high quality rubric has the advantages of reducing random variation within students, by reducing their need to guess what they must do to achieve a high score on the various assignments, and across students, by producing greater consistency in grading any specific assignment. Finally, the professor must provide some training to ensure the students are clear about the criteria used and the scores appropriate for different levels of performance. These three elements are essential whether the professor is grading the assignments alone or whether she is using peer assessments as well. Political science professors, unfortunately, appear rarely to give clear and specific guidance regarding performance standards, at least in their syllabi.

Objections to Providing Clear Standards

Professors might resist giving clear guidance for a couple reasons. They require additional work. They fail to recognize the inherent subjectivity in grading. They reduce professorial discretion. Each objection is legitimate, but none are convincing.
Developing clear standards does require extra work, at least at first. However, once the professor develops clear standards they can be used over and again with modest additional effort. Moreover, the work required to develop clear standards does not seem to be any more onerous than the effort need to develop a new course or update an existing one. Given the potential to improve grading accuracy, it seems time well spent.

A second potential objection to providing explicit standards is that grading is inherently subjective – and that such standards thus cannot be developed. But without explicit standards, the professor and student are playing out this scenario:

Professor: Your task is to shoot at the target. Your score depends on how close you get to the bull’s eye.

Student: Where is the bull’s eye?

Professor: I’ll tell you after you shoot.

Even when subjective judgments must be used, performance criteria can be outlined in advance. The evolution of scoring in figure skating is instructive on this point.

Traditionally, judges had almost complete discretion in scoring skaters, and awarded them two scores based on two broad categories: technical proficiency and artistic merit. Though both scores were based on the informed opinion of experts, these opinions clearly varied from judge to judge. To provide summary scores that were accurate and fair, given the inherent subjectivity and potential bias in scoring, panels of judges were typically used; the highest and lowest scores were dropped, and the rest of the scores averaged to provide the summary score. But controversies still abounded, as judges were often seen as being biased by politics or, in at least one highly publicized case, money.
To remedy these flaws, both in reality and perception, the International Skating Union (ISU) in 2004 changed the scoring rules to provide greater consistency and greater accuracy. Rather than relying on purely subjective scores for the technical merit component, specific standards were set for the various elements, with explicit points attached to each element. Artistic merit continued to be scored subjectively, although the ISU does provide various criteria for assessing it.\textsuperscript{17} The total score weights technical and artistic merit about equally.\textsuperscript{18}

The new scoring system has clear benefits over the old one.\textsuperscript{19} It helps the judges, by reducing perceptions that they are biased; it allows them to focus their discretion on matters that actually require it; it reduces the likely variability of their scores. The system also improves the confidence of the viewing public. More importantly, it benefits the skaters, in at least two ways. By giving them more guidance about what they must do to earn high scores, it gives them clear signals about how to succeed. It also increases their conviction that the scores they receive are those that are actually merited.\textsuperscript{20} Surely professors should treat their students with the same respect.

A final objection to setting performance standards is that reduces professorial discretion. Undoubtedly the subject matter influences how much discretion might be used in assessing students. A general principle might be that the greater the technical objectivity of the material, the less the discretion; the greater the ‘artistry’, the greater the discretion. For some subjects, the distinctions will be difficult to draw. For example, some great works of literature (e.g., James Joyce’s \textit{Ulysses}) would receive low scores if assessed on conventional standards of grammar; its rejection of the conventional standards is indeed a central element of its mastery. The danger is that professors, to the extent that they value
discretion, may have incentives to proclaim that their subject matter is primarily ‘artistic’. While sometimes reasonable, such justifications are potentially self-serving, as they send a signal to students that “I can’t tell you what good work is, but I know it when I see it.” To say the least, this provides students precious little guidance about how they can improve the quality of their work.

EXPERIENCE WITH STUDENT ASSESSMENTS AND CLEAR CRITERIA

In principle, clear performance criteria can reduce the random fluctuations in scoring because it gives both the students and the professor guidance as to what kinds of performance are associated with what kind of scores. In principle, using peer assessments can provide additional information for assigning accurate scores. In practice, how useful are performance criteria and peer assessments in improving the accuracy of grading?

To answer these questions, I have used data from a natural experiment involving the course “Ethics and Values in Public Policy” I routinely teach. This is a required course for the Masters in Public Policy at Georgetown University, and the class typically enrolls between 15 and 20 graduate students. In recent years I have tinkered, in an ad hoc way, with various assessment schemes. These were not true experiments and therefore cannot always answer the questions worth asking, but they can reveal some information on the merits of clearer standards and peer assessment.

In this course I graded the students on the following six assignments: two policy memos, one oral debate presentation, a final oral presentation, a final research paper, and participation. I used peer assessments for two assignments (the oral presentations). I collected and assessed grading data over six semesters. For the first three semesters I
provided no formal performance criteria, although I did offer it informally and verbally over the course of the semester. These three semesters are denoted ‘BC’ (before criteria). For the next three semesters I provided a written rubric for every assignment except the participation element; these rubrics indicated how to score the various elements of each assignment. For example, the students were to score the debate presentations on a 1-5 scale across five dimensions: recommendation, analysis, rebuttal, style and ‘overall’. The rubrics were distributed at the beginning of the semester and also discussed in class. The scores for the three semesters after rubrics were used are denoted ‘AC’ (after criteria).

These data allow for a pretest-posttest comparison, with one control (the participation element, which remains in the pretest state throughout). The expectation is that the variability in the scores that students receive will be substantially lower in the posttest period. Any reductions in variability can be attributed to smaller random fluctuations in student performance, grader performance, or both, although it is not possible with these data to determine which factor dominates.

Figure 3 provides the summary statistics. The six courses are arrayed on the X-axis; the standard deviations in the scores on the Y-axis; the three bars represent the variability in the professor’s scores for all assignments except participation, the peer assessments of the oral presentations, and the professor’s participation scores. The bars are not directly comparable, as the professor’s bars represent the standard deviation in scores for the entire class while the peer assessments indicate the standard deviations for each individual student, averaged for the entire class. Still, the bars all depict the trends in variability across the six semesters.
The patterns are clear: the professor’s scores and the peer assessments were much less variable after formal performance criteria were introduced. The standard deviation in the professor’s scores declined from 2.91 to 1.64 in the shift from BC to AC: a decline of about 40 percent. The decline in the variability of peer assessments is even more striking, as the standard deviations fell from 8.67 BC to 3.69 AC, a 57 percent decline. Furthermore, for professor and students the standard deviations are lower in every semester AC than BC.

Compare these pretest-posttest scores to those of the participation element, for which I provided no real guidance in either period. The most relevant comparisons are professor-to-professor. Even in the pretest period, the participation scores were more variable than for the other assignments (4.52 v. 2.91), and the gap between the participation and other assignments grew larger in the posttest (3.96 v. 1.64). The participation scores declined somewhat in the posttest period, but this is entirely due to one exceptionally high pretest score. These results suggest that establishing specific performance criteria greatly reduces the fluctuations in grading for both the professor as well as the students. A final critically important question remains: How much have the performance criteria and peer assessments improved the overall accuracy of my grades? It is impossible to say with certainty, as the ‘true’ grades are not observed. But the analysis is suggestive. Figure 2 implies that, for six graded assignments, reducing the standard deviation by one point, from 3 to 2, can reduce incorrect grades by well more than half, to much less than 10 percent. Even though the variability in peer assessments remains fairly high – just under 4 points – the large number of peer assessments also reduces the probability of
incorrect grades to less than 10 percent. Increasing the number of scores, and reducing their variability, should thus be able to reduce grading errors to tolerable levels.

CONCLUSIONS

Some might object to my concerns about the potential impact of random fluctuations on grading by arguing: “Don’t worry. Over the long run, these mistakes average out.” This argument has merit over the course of a student’s career and, if there are a large number of assignments, within an individual class. Given that a student will typically take 48 courses to receive a BA degree, the student’s overall GPA is likely to be a fair measure of the student’s true academic performance, as the grades that are ‘too high’ are balanced out by those that are ‘too low’.

I find this argument unpersuasive. Most professors, I suspect, would consider what they do in the class more important than what occurs on the playing fields at their schools. But I doubt that many professors would be willing to publicly state that it makes no difference when umpires blow calls because, after all, over a season the mistakes even out: any game lost by a bad judgment will counterbalanced by a game won by one. Surely the public would not, and does not, accept such sloppiness. We do not expect umpires to call games correctly ‘on average’ but each and every time, on each and every play. We require umpires to take extensive training so that they will make good calls. We institute instant-replay to guard against bad calls. We fervently hope that inept umpires will be removed from the field. Surely, if professors actually do believe that academics are as important as sports, they should be willing to adopt for themselves high standards for making the correct call by assigning grades accurately.
REFERENCES


Georgetown University. N.d. Available at http://explore.georgetown.edu/courses/index.cfm?Action=List&ProgramID=27&AcademicYear=2005


INside Admissions. N.d. Available at http://www.myfootpath.com/insider/IASample.php?S=1


Wikipedia. N.d. “Figure Skating” Accessed on March 27, 2006 and available at http://en.wikipedia.org/wiki/Figure_skating
Figure 1: The Probability of an Incorrect Grade on a Single Assignment as Determined by the Random Fluctuations in Scoring (SD = standard deviation) and the Distance from the Grading Cut-Point.
Figure 2: The Probability of an Incorrect Grade on Multiple Assignments as Determined by the Random Fluctuations in Scoring (SD = standard deviation) and the Number of Scores.

![Bar chart showing the probability of an incorrect grade with different SD values and number of scores.](chart_image)
Figure 3: The Standard Deviation in Scores Assigned on Debates by the Professor and by the Peer Assessors, and the Standard Deviation in Participation Scores, by Course, Before and After Specific Grading Criteria.
1 The literature on the purposes of assessments is too vast and varied to review here (for one discussion, see Marzano 2000:14-15) For our purposes, ‘grading systems’ involves the awarding of scores; ‘assessment systems’ involve scoring and feedback.

2 This assumes, of course, that such a thing as a ‘true’ score exists. Some might object to this notion, arguing that a student’s grade is nothing more than a professor’s (subjective) assessment of the student’s work: a student’s grade is merely what the professor says that it is. There is no need to resolve this dispute here if we simply assume that, if the professor accurately assigns the grade, that score is the ‘true’ score.

3 Let us assume that these fluctuations are normally distributed around the mean. I consider the possible reasons for these fluctuations below.

4 Again, I assume the fluctuations are normally distributed, and explain the reason for such random fluctuations below.

5 For example, the probability that Student B+ would receive a score of 90 or greater is defined by comparing the relevant z-score (90-88)/4 to the normal distribution table. The exact probability that Student B+ will receive a score of 90 or higher is 0.3085.

6 This probability was derived by generating large numbers of random scores for both students given the assumptions about their mean scores and random fluctuations, and then determining the proportion of the paired scores in which B+ scored higher.

7 The probabilities were calculated from the normal distribution table by calculating z-scores under the various assumptions as described in note 5.

8 The probabilities were calculated from the t-distribution.

9 Note that such standardized tests have no variability in the professors’ scores, as there is no professorial discretion in the scoring.

10 This is clearly a non-random sample, as not every course had a syllabi posted, though I have no particular reason to think the posted syllabi are either better or worse than the typical ones. When multiple syllabi were posted by a single professor for a single course, I examined the most recent one.

11 This does not include the vague ‘participation’ category, which is usually included.

12 The universities are the University of California-Berkeley, Columbia University, the University of Florida, the University of Texas-Austin, the University of California-San Diego, and the University of North Carolina at Chapel Hill. Ten syllabi were obtained from each department. Details of the sampling strategy are available from the author on request.

13 In a simple example, the students are instructed to take a multiple choice exam, but are given no guidance as to what constitutes a correct score. In frustration, each student randomly guesses at the answers. With random guesses, some students will score higher than others not because they are better students but because, on that particular assignment, luck favored them. If another test is given, scores will again vary randomly, with some higher and some lower, but with no correlation with the scores on the first exam.
Using online multiple choice exams raises a couple issues. Most importantly, professors should consider how much they are willing to rely on ‘objective’ (and potentially trivial) testing. In addition, there are potential problems with cheating, although proper precautions can mitigate this.

In addition to providing additional scores, these preliminary projects allow me to give additional feedback to help the students refine and improve their research.

For example, in oral presentations ‘style’ might be a relevant criterion. But does style include ‘professional appearance’ (i.e., dressing as if interviewing for a white collar job)? Whether or not the professor deems appearance important, the students should be informed so that they prepare appropriately.

Within the broad categories of technical proficiency and artistic merit multiple factors are considered, and each factor has multiple components (see ISU 2004 for details). For example, in assessing the “interpretation of music” (one of the 5 components of artistic merit) judges are advised to consider “effortless movement in time to the music (timing); expression of the music’s style, character and rhythm; use of ‘finesse’ to reflect the nuances of the music; and [the] relationship between the partners reflecting the character of the music.” Finesse is defined as “the skater's refined, artful manipulation of nuances. Nuances are the personal artistic ways of bringing subtle variations to the intensity, tempo, and dynamics of the music made by the composer and/or musicians” (ISU 2004: 37).

Though there is no objective reason for this weighting, it does reflect the professional judgment of the skating authorities about the merit of each category.

Similar assessment schemes have been adopted for gymnastics (see Fédération Internationale de Gymnastique 2006). Competitive diving is still judged subjectively, with scores averaged across a panel of judges (USADiving.org n.d.)

The new scoring system is not without problems and controversy, however. In the 2006 Olympics scores for technical proficiency varied by about 20 percent for any given skater (though averaging scores across judges reduced the variation to about 2 percent). ISU standards call for the scores of 9 judges to be randomly selected from 14 anonymous judges, with the high and low marks dropped. The U.S. Figure Skating Association, in contrast, uses the scores of all 9 named judges (Wikipedia n.d.)

It is possible, though it seems less likely, that some professors will do just the opposite, perhaps to avoid grade appeals.

This implies the N for each semester was greater than 75 (i.e., 5 assignments times 15 students).

This implies the N for each semester was approximately 400 (14 students rating 14 other students over the two presentations).

The rubric is available from the author upon request.

Note that the variability in participation scores, in declining order, is BC, AC, AC, BC, BC, AC.