

## Why Our Grades Are (Often) Inaccurate and How We Can Make Them More Accurate

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: access to scholarships and awards, entrée to advanced education and employment, and so forth. 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.

Professors, recognizing the personal and professional consequences of the marks they award, undoubtedly also care about grades. Though 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 a central task). What knowledge we do obtain about grading tends to be *ad hoc* to say the least.

This is understandable; so far as I know, there is no professional recognition of, and precious few incentives for, developing and using sound 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 strive to provide careful analytical guidance for conducting empirical research. In the classroom, however, if we do not apply the same efforts to our own grading ‘policies’ we do not practice what we preach.

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

### GRADING ACCURACY

Student assessment systems can seek to accomplish many goals.<sup>1</sup> Although there is little consensus on the purposes of grading (for example, as to whether grades can and should be used for motivate students, 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 accurate as possible.

Grades represent estimates of student ‘achievement’ – an unknown parameter. In statistical terms, an ‘estimator’ is the function used to produce the estimate of this parameter. Good estimators are (in general) efficient, unbiased, and consistent. An estimator is more efficient than another if it produces estimates of the parameter that have smaller variances than the alternative estimator (that is, the better estimator yields estimates that on average are closer to the parameter). 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 time – as well they should – to ensure that the estimators they use are efficient, unbiased and consistent. If a researcher neglected to use high quality estimators, their research would surely be discredited. It is thus ironic that professors (apparently) expend so little effort in making sure that the estimators (grading schemes) they use to estimate student performance are similarly efficient, unbiased, and consistent.

In this section, I focus on the first key element of high quality estimators: efficiency. I do so because the third element (consistency) will exist if the estimators are efficient and unbiased. Moreover, although grading systems are fraught with bias (and I touch on this topic briefly below), substantial research has been conducted on the sources of grading bias and potential remedies for it.<sup>2</sup>

### *Efficiency (Accuracy)*

Good grading systems will assign scores that accurately reflect a student’s performance (that is, the estimators they use to estimate grades will be efficient). For our purposes, let us call this the student’s ‘true’ grade. Presumably, professors should want to assign grades that are true. But unless professors takes 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.<sup>3</sup> Student A- thus is ranked higher than student B+. Let us also specify that Student both students true grades are approximately at the midpoint of the scale for their grades (e.g., B+ are awarded to those with scores between 86 and 89.9).

Now, let 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 (i.e., sometimes the student’s work is better than the ‘true’ score, and sometimes it is worse).<sup>4</sup> Finally, let us also assume that the professor on average scores the actual assignments correctly (e.g., if the student actually produces an 88 paper, the professor awards it an 88), but the professor’s scores also randomly fluctuate around the actual mean, with a standard deviation of 2 points.<sup>5</sup>

Given these assumptions – their reasonableness is discussed further below -- two questions can be asked. What is the probability that each will receive the correct 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.<sup>6</sup> Moreover, Student B+ would receive a higher score than Student A- approximately 23 percent of the time.<sup>7</sup>

The likelihood that the students receive an incorrect grade is a function of how many scores the student receives, the position of the student's true score on the grading scale, and the amount of random fluctuation in these scores. Figure 1 shows the probability of an incorrect grade on a single assignment for grades subject to various amounts of random fluctuation.<sup>8</sup> The horizontal axis shows how far the student is from the next (higher or lower) 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' (say, a student with a score of 88 and a B+ awarded for scores of 90 and a B for scores of 86). Each bar represents a student with differing amounts of random fluctuations in their scoring, ranging from a standard deviation of one point to six points.

Two implications are worth highlighting. Most obviously, for students 'on the edge' the likelihood of receiving an incorrect score are 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 scores. 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 scores 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 (not on a single assignment, but for the course as a whole) also depends on the number of scores the student receives. Figure 2 shows these probabilities for a student with true scores in the middle of a grade scale (e.g., a score of 88 where a B+ is given for scores between 86 and 89.9), given various amounts of random fluctuation in the scores assigned.<sup>9</sup> The key conclusions from this figure are that the likelihood of awarding a correct score increases with the number of scores assigned, but decreases with the amount of random fluctuations. If the fluctuations are small (e.g., standard deviations of two points or less) 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 high, however, the probability of awarding an incorrect score remain fairly large: if the random fluctuations are grading 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 too pessimistic, then the problems in assigning correct scores might be less 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 relative small (4 points on a 100 point scale). 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 elite universities, or in upper-division classes, student performance is likely to be rather compressed. At Georgetown, 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 (on a 4 point scale). In Georgetown College (the liberal arts division of the university) 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.): quite literally, all students there are 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.

We do have some evidence regarding the minimal amount of random fluctuations that are likely in student performance. 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 (unaffected by random variation) scores: it has many questions and the format is well-known (so students can learn with fair precision what they will be expected to do and how they will be scored). It seems unlikely to me that any professor’s grading scheme would produce such little intrapersonal variation in scoring.

There are undoubtedly studies with data on the variation on instructors’ scoring, but thus far I have not been able to locate them. An adequate test of this would be difficult to construct, as it would require data from professors who grade the same assignments twice, close together in time, without realizing that they are doing so. In my own classes, occasionally I have inadvertently graded the same papers twice, and my scores seem to vary by a couple of points.

There is ample evidence that instructors award different grades when assessing the same assignments. Marzano (1988; cited by Marzano 2000: 5-6) evaluated seven studies examining classes that were team-taught, with each instructor independently grading the identical body of work.<sup>10</sup> Across the seven studies, the instructors awarded the same letter grade an average of 59 percent of the time.<sup>11</sup> Twenty-five percent of the grades differed by one letter; twelve percent by two letters; and four percent by three letters.<sup>12</sup> This implies that if a student received an A from Professor Y, there is about a 40 percent chance that the same work would yield a B, C, or even (rarely) a D from Professor X.

It also appears clear that professors – well, political science professors at Georgetown, anyway -- often score only a very small number of assignments. I surveyed the 38 most recent versions of Department of Government syllabi posted on Georgetown University website in the spring of 2006 (Georgetown University, n.d.).<sup>13</sup> On average, the typical

syllabus listed under four graded assignments (mean = 3.87, standard deviation = 1.22).<sup>14</sup> Fully seventeen of the courses had three or fewer graded assignments. As we saw above, when such a small number of assignments are graded, the chances for incorrect grades is quite high.

### *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; 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 a 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.<sup>15</sup> Such random variation is conceptually independent of student attributes (though, perhaps, one definition of a 'good' student is one who is better at guessing that 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 herself does not know precisely what the grading standards are: reading a paper, the professor does not know how to evaluate it in assigning it a grade. I do not know 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 was left guessing what score to award the students. 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 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 syllabi offered specific counsel on expectations for performance across the required assignments.

It is course possible that the professors distributed other guidance regarding assessment standards, but it seems odd (at least to me) how little advice was incorporated in the courses' key reference document.

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.) Marzano concludes his remarks with an anecdote:

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)

It is not clear from Marzano's work whether the reported grading discrepancies should be attributed to the instructors using different quality standards or weighting schemes. In either case, however, the remedy for producing more consistent grades is clear: make the weighting and the performance criteria more explicit and transparent.

## IMPROVING EFFICIENCY

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 clear, 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 fluctuation in grading, in theory more graded assignments will produce more 'true' grades than will fewer scores. In practice, professors may be reluctant to require additional assignments that they have to grade, however.

A couple principled reasons for requiring only a few (or even one) 'high stakes' assignments can be offered. One is that it is better to require students to focus their efforts on a small number of big tasks rather than distract them with multiple smaller assignments. This rationale is questionable. After all, even big projects – say, a semester long research paper – can (and, I believe, should be) be decomposed into smaller, discrete elements (e.g. hypotheses, literature review, research design, analysis, and so forth). Preparation for a comprehensive final exams 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’ – but, of course, most (professional) jobs involve frequent, multiple tasks. 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 (whether natural or social) 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, as I hope they are, 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. No thorough researcher should be content, after all, if their research contained such (potentially) extensive errors.

In my own classes, I have increased the number of graded assignments in a couple ways. First, when possible, I have increased my reliance on computer-graded assignments. For example, in my “Introduction to the U.S. Political System” course (a required entry level course for majors) 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 automatically entered into their Blackboard grade book. These quizzes account for 30 percent of their grade. 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.<sup>16</sup> Second, I require more ‘interim’ assignments contributing to a final project. For instance, in my undergraduate “Scope and Methods in Political Science” (a research methods course), 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).<sup>17</sup> Currently, I require seventeen graded elements in my introductory class, nine for the research methods class, and five for my ethics class (although the ethics class has many other scores, as discussed below). I believe, and the analytical evidence presented above suggests, that I am scoring enough projects to award incorrect grades to only a small fraction of the students.

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

Requiring more assignments is not the only way to produce more (and hopefully 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 (on a 100 point scale). With a single grader, 95 percent of the time we would expect the correct score to be assigned, plus or minus 6 points (e.g., if the ‘true’ score is 90, then 95 percent of the time the assigned score would

be between 84 and 96). 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’).<sup>18</sup> As a conceptual matter, peer assessment can improve grading accuracy to the extent that the peer scores are themselves efficient 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.<sup>19</sup> 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 if the random fluctuations are reduced from three points to two points, and then to three percent if the random fluctuation is a single point. In other words: reducing random fluctuations can have a powerful, positive impact on grading accuracy.

The most obvious 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 (or irrelevant) in assessment.<sup>20</sup> Second, the students should be given a ‘performance rubric’ which indicates how the different criteria should be scored.<sup>21</sup>

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 (if the Georgetown syllabi mentioned above are representative).

### *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; and 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 over 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 the 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.<sup>22</sup>

To remedy these flaws (in reality and perception), in 2004 the International Skating Union (ISU) 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.<sup>23</sup> The total score weights technical and artistic merit about equally.<sup>24</sup>

There are clear benefits for the new scoring system over the old one.<sup>25</sup> It helps the judges, by reducing perceptions that they are biased; it also allows them to focus their discretion on matters that actually require it. 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.<sup>26</sup> Surely professor 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., *Proust*) 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’.<sup>27</sup> While sometimes reasonable, such justifications are potentially self-serving, as they send 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

This section presents data from student assessments from the “Ethics and Values in Public Policy” course I routinely teach.<sup>28</sup> This is a required course for the Masters in Public Policy at Georgetown University. The class typically enrolls between 15 and 20 graduate students. Over the past several years I have experimented, in the typical *ad hoc* way, with various assessment schemes. These are not true experiments (i.e., carefully designed with clear hypotheses, adequate controls, and so forth) and therefore cannot always answers the questions worth asking, but they can reveal some information on the merits of student assessments and clearer standards.

Table 1 provides data on the variability in the grades I have given in this course the past six times I have taught it, as well as the scores from the peer assessments.<sup>29</sup> For my scores, I do not have data on intra-student variability on a single assignment (as, obviously, I only assign a single score) but I do have data on the variability across the students. The first column of data contains the mean size of the standard deviation across all course assignments except ‘class participation’; the second column contains the participation scores I assigned. The third column contains the peer scores. For the first three semesters (summer 2005 through spring 2006) I did not provide specific performance standards; the last three assignments, I did except for the participation score.

The hypothesis I want to examine is whether clear standards lead to reduced variability (as students have less need to ‘guess’ what I want) across the students. While I have not yet conducted formal hypothesis tests, the pattern seems clear. When I switched to

clear(er) standards in the summer of 2006, the mean standard deviations of my scores declined from about three points to about 1.7 points – a decline of about 40 percent. The participation scores, which have substantially higher standard deviations, remained high and stable. The standard deviation of peer assessments are much higher, but they also declined markedly after I provided clearer scoring guidance. These results suggest that establishing specific performance criteria greatly reduces the fluctuations in grading for both the professor as well as the students.

Table 2 shows the results of OLS regressions used to predict overall scores for the debates from the ethics class during the past three semesters.<sup>30</sup> For each class the students were given five criteria for scoring (argument, analysis, rebuttal, style and overall score) as well as specific guidance about how to score the debates (see Appendix 2). Scores on each criteria were on a 100 point scale (except for the summer class, which was on a 10 point scale). Sex was coded as 0 for male, 1 for female; Language was coded as 0 for a non-native English speaker and 1 for native English speaker. The N was determined by the total number of assessments submitted, which equals the number of debate performances multiplied by the number of raters.

Several features are especially noteworthy. First, there was no evidence of bias in the scoring. In each class there was no difference between the scores awarded by men and women or the scores received by men or women. The language of the speakers and the raters had no impact on overall scores, with the exception that native English speakers gave slightly higher scores (a little more than half a point) than non-native speakers in one class. This speaks well for the ability of the raters to assess the presentations unaffected by considerations of gender or language.<sup>31</sup>

Second, it is clear that the assessment criteria are quite useful in predicting overall scores. The  $R^2$  reveal that together the scores on the separate criteria explain about 90 percent of the variation on the overall scores. The standard error of the estimates (S.E.E.) indicate that these criteria allow us to predict the overall score with a margin of error of about one percentage point: a very close prediction.

Third, the individual criteria are all uniformly important in predicting the overall scores: these criteria are always significant at less than  $p < .01$ . This means that the students do in fact see these as separate performance dimensions that each contribute to the overall score, and that they are able to score each one independently of the others. Yet the consistency of the impact of each criterion on the overall scores is amazingly consistent across classes. In general, each additional point on a single criterion increases the overall score by about 0.23 points (with the estimates ranging from 0.14 to 0.27).

In short: the data from the student assessments presented here provide substantial evidence that students can assess other students fairly, that they are able to use separate criteria in making these assessments, and that the overall scores are closely linked to the used to score the performance. These data increase my confidence in using student assessments.

## 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 enough 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 take the position (publicly) 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 the high standards the public demands from umpires.

FIGURES AND TABLES

Figure 1: Probability of Incorrect Grades by Random Fluctuations

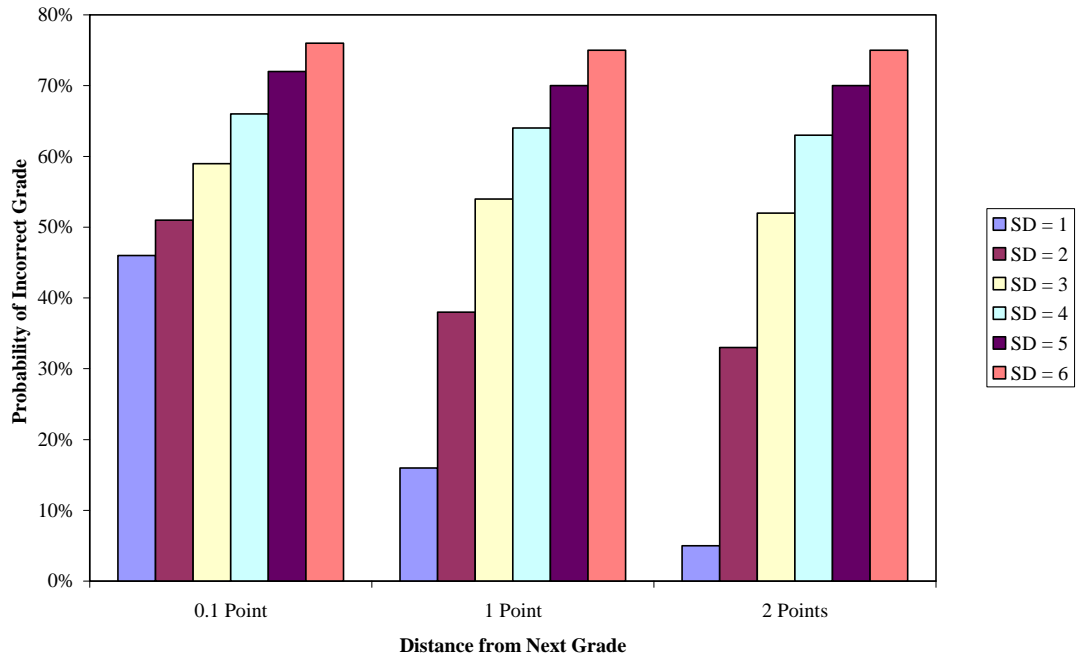


Figure 2: Probability of Receiving Incorrect Grade by Number of Assignments

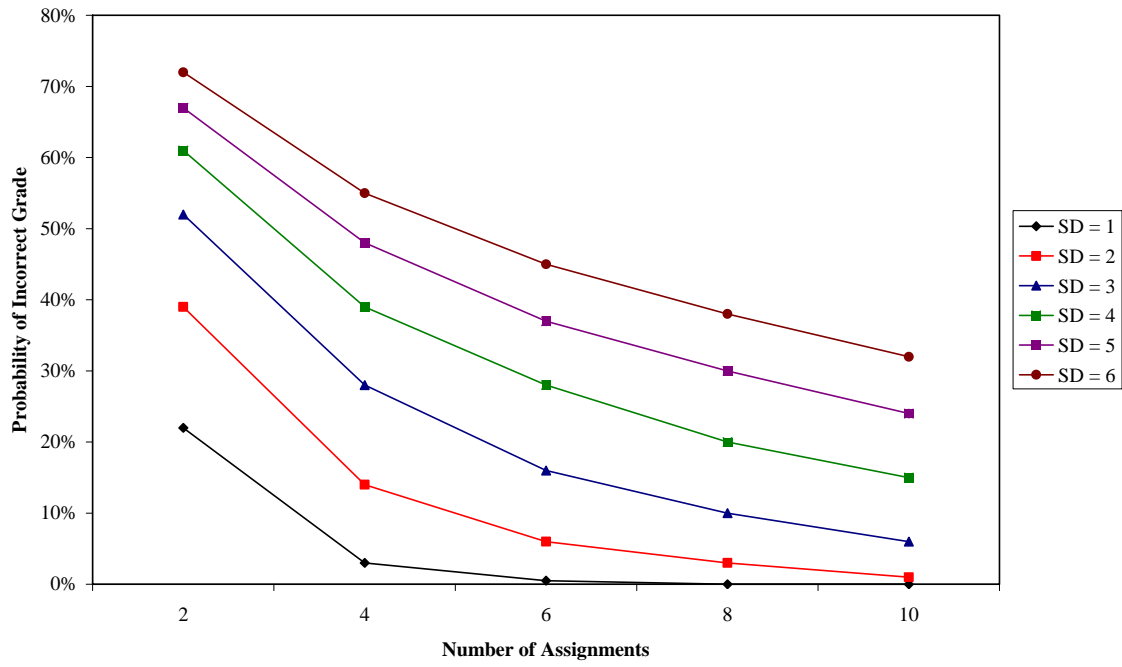


Table 1: Mean Size of Standard Deviations on Student Assignments, Ethics Class			
	Assignments	Participation	Class Scores
Summer 2005	2.96	3.59	NA
Fall 2005	2.66	5.92	9.39
Spring 2006	3.11	4.04	7.94
Summer 2006	1.71	3.24	4.69
Fall 2006 Morning	1.47	4.30	3.09
Fall 2006 Evening	1.73	4.34	3.28

Table 2: The Determinants of Overall Debate Scores			
	Summer 2006	Fall 2006 Morning	Fall 2006 Evening
Debater's Sex	0.07 (0.74)	-0.01 (0.12)	-0.05 (0.23)
Rater's Sex	-0.12 (1.27)	-0.05 (0.42)	0.03 (0.13)
Debater's Language	0.17 (1.08)	0.08 (0.39)	0.54 (1.92)
Rater's Language	-0.15 (1.07)	0.17 (0.87)	0.58* (2.29)
Argument	0.20** (8.34)	0.18** (7.53)	0.14** (3.51)
Analysis	0.20** (7.86)	0.28** (11.38)	0.24** (5.91)
Rebuttal	0.27** (16.69)	0.27** (14.67)	0.25** (10.32)
Style	0.27** (15.22)	0.25** (13.09)	0.21** (8.37)
R2	.91	.91	.85
F	452.19**	312.24**	91.31**
S.E.E.	0.85	0.88	1.0
N	355	263	139

Notes: The table presents unstandardized regression coefficients and (t-statistics). \* indicates significance at .05; \*\* at .01

## REFERENCES

<sup>1</sup> 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.

<sup>2</sup> Nonetheless, I do offer a summary of some of the causes of and remedies for bias in Appendix 1.

<sup>3</sup> 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.

<sup>4</sup> Let us assume that these fluctuations are normally distributed around the mean. I consider the possible reasons for these fluctuations below.

<sup>5</sup> Again, I assume the fluctuations are normally distributed, and explain the reason for such random fluctuations below.

<sup>6</sup> 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.

<sup>7</sup> 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.

<sup>8</sup> The probabilities were calculated from the normal distribution table.

<sup>9</sup> The probabilities were calculated from the t-distribution.

<sup>10</sup> The courses examined were for students between grades 4 and 12 (Marzano 2000: 6)

<sup>11</sup> Marzano (2000) does not report the sample sizes for the seven studies; these averages are the aggregated average across the studies, not the means across the individual students.

<sup>12</sup> Percentages do not sum to 100 due to rounding.

<sup>13</sup> This is clearly a non-random sample, though I have no particular reason to think the syllabi are either better or worse than the typical ones. A majority of courses had no syllabi posted – this itself is a problem, as it makes it more difficult for the students to select appropriate classes. When multiple syllabi were posted by a single professor for a single course, I examined the most recent one.

<sup>14</sup> This does not include the vague ‘participation’ category, which is usually included.

<sup>15</sup> 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.

<sup>16</sup> Using online multiple choice exams raises a couple issues. Most importantly, one might question my reliance on so much ‘objective’ (and potentially trivial) testing. In addition, there are potential problems concerning cheating. I have taken several measures to reduce this potential; further details on the risks and opportunities are available from the author on request.

<sup>17</sup> In addition to providing additional scores, these preliminary projects allow me to give additional feedback to help the students refine and improve their research.

<sup>18</sup> For one review, see Topping (1998). In the tradition of ‘muddling through’, I used peer assessments for several years before actually trying to understand the conditions necessary for high quality peer assessments.

<sup>19</sup> Whether the greater variation leads to more randomness in the scores depends on how large the fluctuations are and how many scorers are used.

<sup>20</sup> 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.

<sup>21</sup> The rubric I have developed for the debates in my ethics class is attached as Appendix 2.

<sup>22</sup> Citation on French judge here?

<sup>23</sup> 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).

<sup>24</sup> 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.

<sup>25</sup> 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.)

<sup>26</sup> 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.)

<sup>27</sup> It is possible, though it seems less likely, that some professors will do just the opposite, perhaps to avoid grade appeals.

<sup>28</sup> I have used student assessments in other courses, but I have the most data from the Ethics course.

<sup>29</sup> My scores are for two policy memos, the debates, a final oral presentation, and a final research paper. The peer assessments are only for the debates and final oral presentations.

<sup>30</sup> I have not yet conducted regression analysis on all student assessments from my ethics classes, although I plan to finish the analysis in the future.

<sup>31</sup> I was a bit surprised to see that there was no difference in the scores awarded based on language ability, as in my view the native English speakers typically gave stronger presentations. This suggests that the student raters may have adjusted their expectations regarding the non-native presentations. It is possible that language affects the style (but not overall) scores, but I have not yet scrutinized the data for this.