

The Impact of Grading on the Curve: A Simulation Analysis

George Kulick

Le Moyne College
Syracuse, New York, USA
kulick@lemoyne.edu

Ronald Wright

Le Moyne College
Syracuse, New York, USA
wright@lemoyne.edu

Abstract

Grading on the curve is a common practice in higher education. While there are many critics of the practice it still finds wide spread acceptance particularly in science classes. Advocates believe that in large classes student ability is likely to be normally distributed. If test scores are also normally distributed instructors and students tend to believe that the test reasonably measures learning and that the grades are assigned fairly. Beyond an intuitive reaction, is there evidence that normally distributed test scores appropriately distinguish among student performance? Can we be sure that there is a significant correlation between test scores and student knowledge? Testing these assumptions would be difficult using actual subjects. In this paper we use mathematical models and Monte Carlo simulation to test the assumption that normally distributed grades assign the highest grades to the students who were best prepared for an exam.

Key Words: Grading on the Curve, Computer Simulation, Student Performance

Introduction

Grading on a curve, in one form or another, is a common practice in higher education. The practice is often criticized for ignoring the possibility that an instructor and the class may have together worked to the point where more than half of the students have earned a top grade (Roth, 2000). Critics also suggest that grading on a curve does not provide the ideal incentives for student motivation (Michaels, 1976). But the practice also has wide spread support. It is advocated as an antidote to grade inflation. It is also used in contexts in which institutions feel the obligation to distinguish among performances for the purpose of evaluating students for professional and graduate schools. While critics tend to argue that goals ought to focus more on teaching and less on evaluation and ranking, both sides seem to agree that grading on a curve is well suited for this ranking function. It is, in fact, this last assumption that this paper challenges.

Can we be confident that grading on a curve results in assigning the best grades to the best students? Addressing this question is complicated by a number of factors. If we try to measure the extent to which the higher test scores are associated with the higher levels of learning or achievement, we are confronted with the reality that the test itself is intended to be our measure of achievement. Furthermore we are likely to have difficulty agreeing on what should be meant by the best students. Is it the best prepared, those with the best ability, those who have learned the most? And how would we distinguish among and measure performance for these alternative definitions? Even when we address all these factors we will always find ourselves constrained by the relatively small samples from the student population which we can involve in our experiments. To address all these concerns we propose using Monte Carlo simulations to create large populations of students exactly matching a range of assumptions and then administering hypothetical exams to these

hypothetical students. These mathematical simulations will allow us to vary our assumptions about the students and the exams and consequently investigate the extent to which the best students (by whatever definition we elect to use) obtain the best scores.

The first section of the paper briefly examines the concept of grading on the curve. It is followed by a section that describes the simulation model and the assumptions on which the various models are based. Subsequent sections contain the results of the simulations as we progress from typical students at competitive institutions to an extreme case of highly motivated pre-med students at the most selective institutions. In each case we look at the correlation between performance on the exams and student preparedness for the exam. The final sections focus on initial observations, potential impacts on students, and directions for additional work.

The Concept of Grading on the Curve

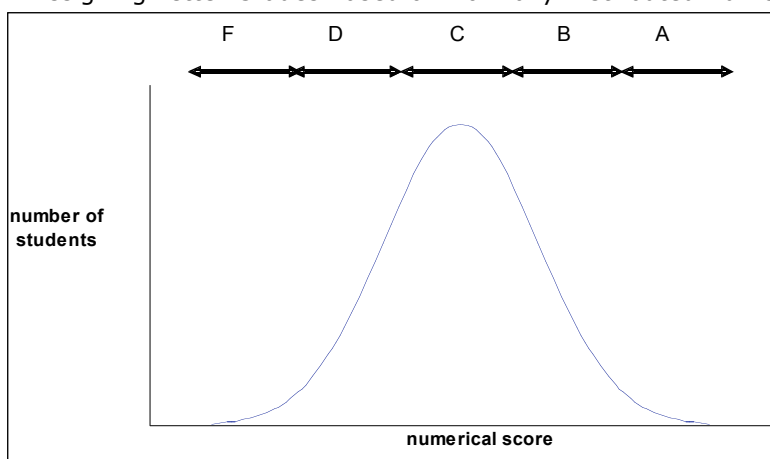
Grading on the curve has long been an accepted form of student assessment, particularly in large classes. In reality, grading on the curve has come to have a variety of meanings to students and instructors (Wall, 1987). The most simplistic, and perhaps the most commonly embraced by students, is the practice of adding points to all grades to bring the highest test scores up to the 100 point range. This practice is actually more commonly referred to as curving grades and an instructor might often say he or she decided to curve the grades on an exam that turned out to be more difficult than anticipated. In other instances curving grades is associated with predetermining, independent of performance on an exam, a fixed percent of students that receive each grade. This can be a means of assuring that in each class roughly the same percent of students receive the same final grades independent of the difficulty of specific exams or instructors. For example, the Psychology Department at a large US university has recommended distributions that result in 15% A's, 25% B's, 45% C's, 10% D's and 5% F's (Wedell, Parducci, & Roman, 1989). In this form, grading on a curve is seen as an antidote to grade inflation and can even be based on informal or formal institutional policy. For example, according to a very recent article in Boston University's *The Daily Free Press* (Maxwell, 2007), university policy suggests that the mean grade in large classes should be around B-minus or C-plus. In reality issues of grade inflation have existed for at least three decades (Abbott, 2008).

However, the original use of the phrase, grading on the curve, was based on the assumption that student abilities, particularly in large classes of hundreds of students, would most likely be normally distributed. It often argued that "exam scores tend to be normally distributed for well-constructed, norm-referenced, multiple choice tests (Wedell et al, 1989, p.239). In an experiment conducted by Wedell et al (1989), students were asked to assign grades to test scores that matched four different distributions, a normal distribution, a U-shaped distribution, and positively and negatively skewed distributions. Without any knowledge of the subject material, students were also asked how well the test measured knowledge for each of the distributions. The students rated the normally distributed test scores significantly higher than the other distributions. In general it is common to associate effective exams with normally distributed scores. Consequently, it is not a surprise that instructors feel they have created a fair and effective exam whenever they graph the test score distribution and see that it resembles the bell-shaped curve. As any student can attest, individual instructors use a variety of approaches to distributing the grades normally.

Johari and Sclove (1976) described one such approach, but most approaches end up assigning the letter grades in a manner represented in Figure 1.

In the case represented in Figure 1, grades are curved to an average grade of C. In other instances instructors curve to a low B, having A's and B's for students above the mean and C's and D's for those below the mean. In this case F's are assigned to outliers. Throughout the remainder of this paper we will use the phrases "grading on the curve" and "curving grades" to represent the practice of fitting letter grades to a normal distribution.

Figure 1. Assigning Letter Grades Based on Normally Distributed Numerical Scores



This commonly used form of grading has both its critics and proponents and the topic is one often discussed in education public forums ("Grading on a Curve", 2007). Many argue that it is unfair to automatically assign low grades to the lower end of the class when it is possible, in some classes, that all of the students will have achieved a certain mastery of the material and deserve grades that recognize their level of performance. Proponents of grading on the curve argue that as long as the performances on the exam vary according to a normal distribution, it is reasonable to have the grades also be normally distributed. The practice is perhaps the most commonly applied in introductory science classes in large universities (Maxwell, 2007).

In a large introductory chemistry class at a competitive university it is reasonable to expect that grades on a typical exam will be normally distributed and grading on a curve seems reasonable. But that same exam given to a chemistry class for pre-med students at a highly selective university might well result with half the students obtaining a near perfect score, hardly a normal distribution. In this case, grading on the curve would not be practical. Yet many highly selective universities seem to insist that all pre-med courses be graded on the curve. Consequently the chemistry exam at such a university must be more difficult in order to assure that the resulting exam scores are normally distributed. Although it may not be official policy, it is common practice at these highly selective institutions to ensure that the level of difficulty of exams result in normally distributed scores. Specifically, it is the practice in all pre-med courses at the two such universities which will form the context for our analysis. These two institutions were chosen merely to have a context for the discussion and were two which the authors were able to identify as having at least implicit grading policies. However there is no reason to imply that these outstanding institutions stand out in any way from their peers in this practice. Hence throughout the course of the paper we will refer to them only as institutions A and B.

Reasons for grading on the curve at these institutions are not publicly discussed. However we can assume some possible rationales. Supposedly these institutions and their professors feel compelled to distinguish performance among these outstanding students, to identify which students performed the best in the class and which were the poorer performers. Perhaps such rankings are designed to heighten the reputation of the institution by sending forward only the very best. Perhaps they are attempting to avoid grade inflation so that their best students will be clearly identified for the best medical schools. In many cases it is the very best institutions that are the most concerned about grade inflation (Gordon, 2006). However, any of these (or other reasons) must be based on the assumption that the best grades are going to the best students. It is this basic assumption that our work is designed to investigate. Does grading on the curve always, or even frequently, result in the best students getting the best grades?

Simulation Model

Investigating the assumption that the best students get the best grades would be difficult with samples of actual students. First defining what we mean by the best student is difficult. Do we mean the student with the best ability? Do we mean the student who is best prepared for the exam? Do we mean the students who know the most? And how would we assess the best by either of these measures other than by giving them an exam? How do we take into account how hard the student studied, or how well they take exams, or how they were feeling that particular day, or whether the exam was a fair exam?

A computer simulation model allows us to precisely define all of the assumptions in our analysis and to generate the observations accordingly. It gives us the ability to isolate the particular question at hand. We define our model based on 400 students taking three exams each made up of twenty questions. Initially we assume that all questions on the exam are equally difficult. In addition, we assume that the abilities of these 400 students are normally distributed, regardless on whether they belong to a "typical" or "highly selective" group. The following analysis distinguishes between the two groups by assuming the typical group has a smaller average level of ability with a larger standard deviation.

Our definition of the ability of a student is the probability that the given student will correctly answer a given question. Hence it actually refers to the how well the student is prepared to take the particular exam. Consequently we are not trying to determine whether that preparedness is a consequence of innate ability, hard work, positive attitude, or anything else. Throughout this paper, whenever we refer to student "ability" or student "preparedness", we use the terms in the precisely defined sense of the probability of getting a particular question correct on a given exam. To represent this probability, each of our 400 students is randomly assigned a value between 0 and 1 from a normal distribution with a given mean and standard deviation. The selected mean will indicate the overall ability of the group of students and the selected standard deviation will indicate the range of ability. A student assigned a value of 0.762, for example, would on average get 76.2% of the exam questions correct, assuming that all the questions on the exam are of equal difficulty. Subsequently, when we assume varying levels of difficulty for exam questions we will raise or lower these probabilities for particular questions. Since we selected the probabilities from a normal distribution we have met our assumption that the "abilities" of our 400 students are normally distributed.

While the student in our example is expected to get 76.2% of the exam questions correct on average, for any particular exam the student could obtain a much lower or higher score.

We are assuming that, given limited time for a test, no science exam contains questions from all the covered material. Consequently, at one extreme, our student, who in essence

knows 76.2% of the material well, might be given an exam in which all of the questions fall within that 76.2% and the student could get a perfect score on the exam. At the other end, if much of the material on the exam falls outside the 76.2% the student knows well, the resulting score would be much lower. Hence there is a random factor to the exam and it is this randomness that justifies the use of a computer simulation.

Randomness in a simulation model is produced by a random number generator. The most basic random numbers that initialize simulations are generated using a uniform distribution of numbers between 0 and 1. To simulate performance on an exam question, we randomly assign such a number to each question. A student is assumed to get the question correct if the value assigned to the student is greater than or equal to the value assigned to the question. This assumes each question is scored as correct or incorrect without partial credit. Again, consider the student with the assigned value of .0762. Since 76.2% of the numbers between 0 and 1 fall below 0.762, our student will on average get 76.2% of the questions correct. Of course students with higher ability (higher assigned values) are more likely to get any given question correct. As an illustration, Table 1 contains the results for one exam (20 questions) for a student with an assigned value (or measure of ability) of 0.762.

Table 1. Simulated Exam Grade for One Student

Assigned Value for Student			
			0.762
question	random number	comparison	question correct
1	0.684	$0.762 \geq 0.684$	yes
2	0.883	$0.762 < 0.883$	no
3	0.099	$0.762 \geq 0.099$	yes
4	0.496	$0.762 \geq 0.496$	yes
5	0.778	$0.762 < 0.778$	no
6	0.502	$0.762 \geq 0.502$	yes
7	0.715	$0.762 \geq 0.715$	yes
8	0.006	$0.762 \geq 0.006$	yes
9	0.289	$0.762 \geq 0.289$	yes
10	0.986	$0.762 < 0.986$	no
11	0.948	$0.762 < 0.948$	no
12	0.303	$0.762 \geq 0.303$	yes
13	0.048	$0.762 \geq 0.048$	yes
14	0.955	$0.762 < 0.955$	no
15	0.871	$0.762 < 0.871$	no
16	0.205	$0.762 \geq 0.205$	yes
17	0.741	$0.762 \geq 0.741$	yes
18	0.152	$0.762 \geq 0.152$	yes
19	0.728	$0.762 \geq 0.728$	yes
20	0.503	$0.762 \geq 0.503$	yes
			grade 70

The assigned score is the percent of questions the student got correct on this particular exam. In this instance the student with an assigned ability level of 0.762 obtained a score of 70. On a different exam of the same level of difficulty the student might well receive an 80.

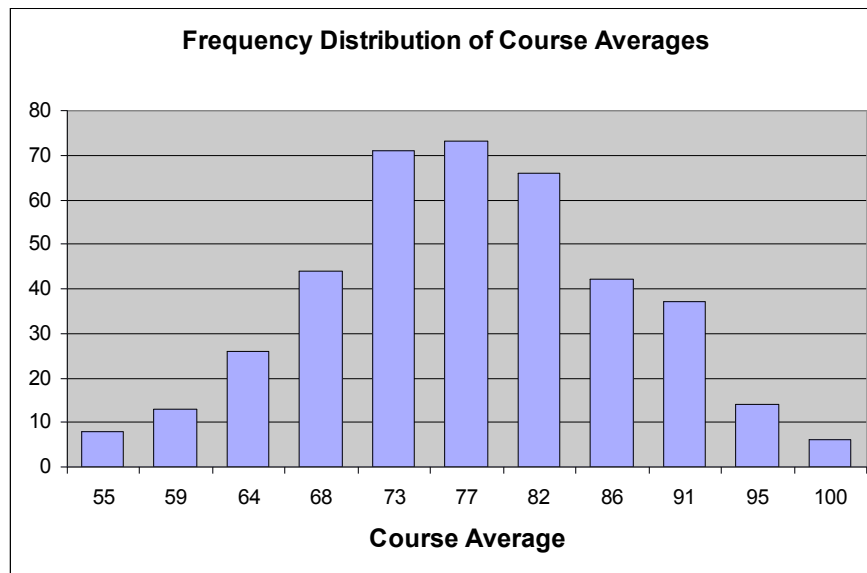
For an infinite number of exams we are of course assuming the student would have an average score of 76.2. Finally we calculate a student's course average as the mean score of three exams. A complete simulation would produce course averages for each of 400 students based on the three exam scores.

Simulating a Typical Group of Students

In our initial simulation we model typical students at a typical university. The first step is to define the normal distribution that will be used to assign random abilities (probability of answering a question correct) for the 400 students. We use a mean of 0.75 indicating that the average student will get (on average) 75% of the questions correct. In this case we will assume that our student ability ranges from 0.50 to 1.00. Since six standard deviations cover 99.74 % of all values, we will use as our assigned standard deviation $1/6^{\text{th}}$ of the range. An average of 0.75 and a standard deviation of 0.083 ($0.5/6$) of course allows for the possibility of a few values in excess of 1.00. All values in excess of 1.00 are modeled to represent students getting all questions correct and the distribution is effectively truncated at 1.00.

The results of our first simulation are summarized in Figure 2, a frequency histogram for the course average (the mean of the three exam scores) for our 400 students. As expected, the averages appear to be normally distributed with a reasonable range from 50's to 100. Grading on a curve to assign final grades seems perfectly reasonable.

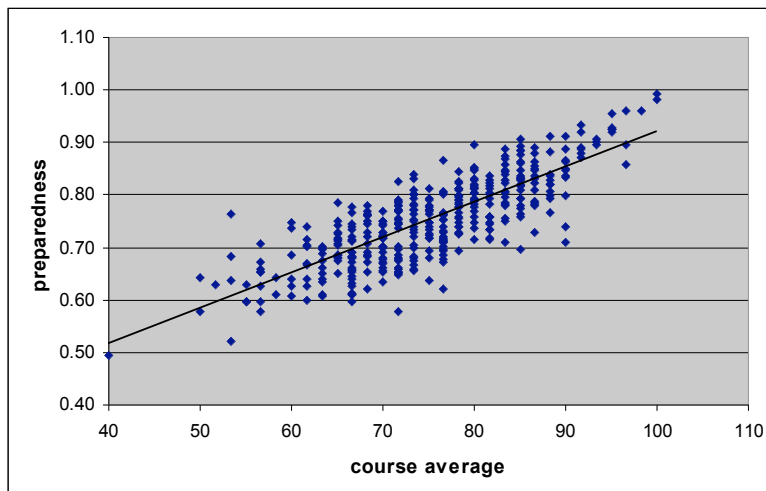
Figure 2. Frequency Histogram of Course Averages for Typical Class



As a result of the simulation, we have, for each of our 400 students, the assigned measure of ability (or preparedness) as well as the course average. We can, therefore, measure the correlation between the two. Figure 3 contains a graph of the preparedness versus course averages for the 400 students. The resulting correlation is 0.81 with a 95 % confidence interval of (0.76, 0.86). Hence we have strong evidence that a correlation exists between preparedness and course averages. However, from the graph in Figure 3 we can also observe that students with average preparedness (0.75) end up with course averages ranging from 60 to mid 80's. If these students, who are equally well prepared, are graded on a curve the resulting letter grades will range from a D to a B. Our assignment of 0.75 as

a level of preparedness suggests that each student knows 75% of all possible test material. Since few tests cover all possible material, the variation in the grades is a result of the extent to which the student was lucky enough to have the material he or she knew show up on the exam.

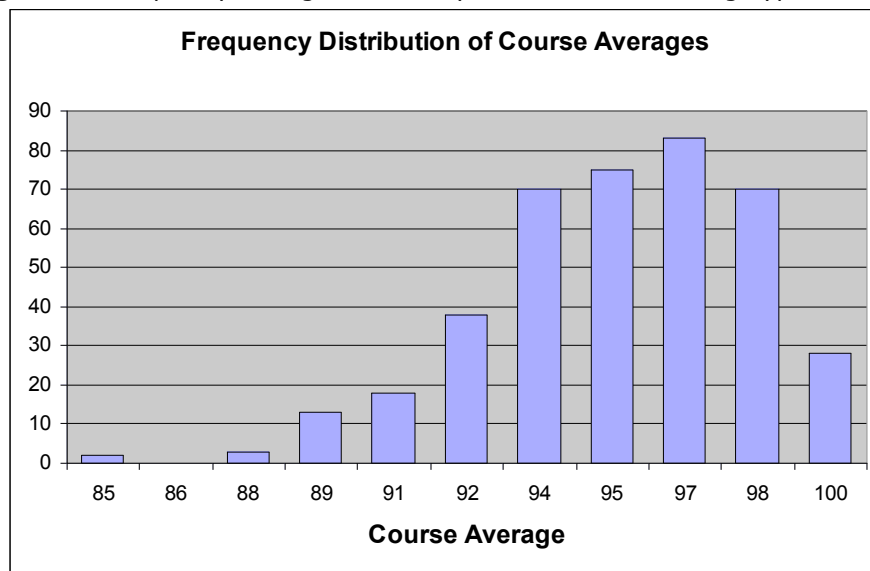
Figure 3. Relationship between Preparedness and Course Average for Typical Students



Modeling Students at Highly Selective Institutions

Some of the criticism of grading on the curve is directed to situations in which all of the students are highly qualified and well prepared. Is it fair to curve their grades so that some automatically receive C's and D's? Independently of whether it is fair, our analysis focuses on whether or not we can be sure that the best prepared students get the higher grades. Our next simulation examines the performance of these students. We have derived the context of this simulation from two actual institutions that do (at least by practice) grade pre-med classes on a curve. At these two institutions the average SAT scores for entering freshmen are 1350 and 1325. Hence the average student at each institution is in the top 7% of all senior college-bound SAT test takers ("SAT Percentile Ranks", 2008). We can further heighten the level of selectivity among these students by simulating grade distributions in pre-med courses. Both of these schools designate separate courses almost exclusively for pre-med students.

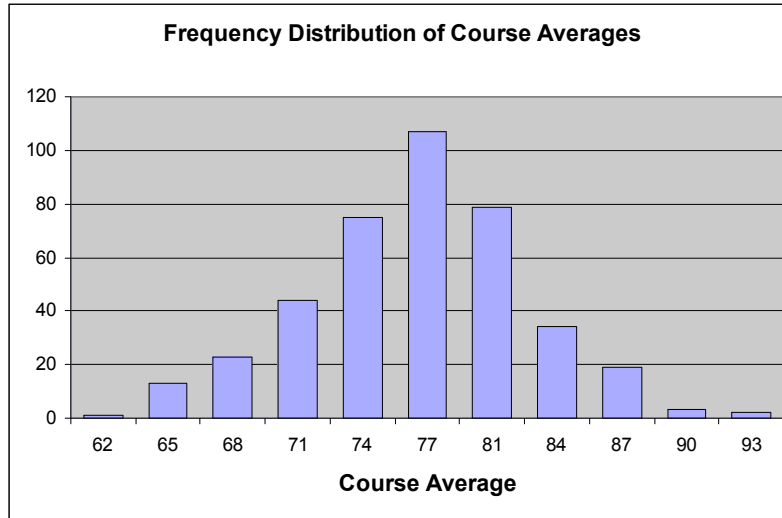
Therefore, the premise of our following argument is that these students are largely the best students from a very select group. To demonstrate how these students compare to the group described above, we simulate the outcome of them taking the same (typical) exam. To represent their higher ability/preparedness, we simulate the probability of each of them getting any given question correct from a normal distribution with a higher mean, say 0.95, and a smaller range of probabilities, say 0.9 to 1. This suggests that these highly talented and highly motivated students would virtually all get 90% of the questions correct in a typical introductory science class taught at a typical competitive university. The result from running our simulation with this group of very good students is summarized in the frequency histogram in Figure 4.

Figure 4. Frequency Histogram for Very Good Students taking Typical Exams

The course averages for the three exams are no longer normally distributed and grading on a curve is less practical. In addition, the highly capable and motivated students would not accept D's for scores in the high 80's or B's for scores in the mid 90's. But one could argue that the exams used at our typical institutions are not the exams used at these highly selective universities. In order to appropriately consider grading these students on a curve, instructors must make the exams more difficult, more specialized, or even more obscure to insure that fewer students get high scores. The consequence of a more difficult exam is to reduce the ability of these students to get the questions correct. We can model this increase in exam difficulty by lowering the average probability that students will get any given questions correct. So, in effect, when we reduce the mean of the normal distribution simulating the probability of getting a question correct we model either lowering student preparedness or a more difficult exam. Smaller standard deviations result in a smaller range and model a more homogenous group of students.

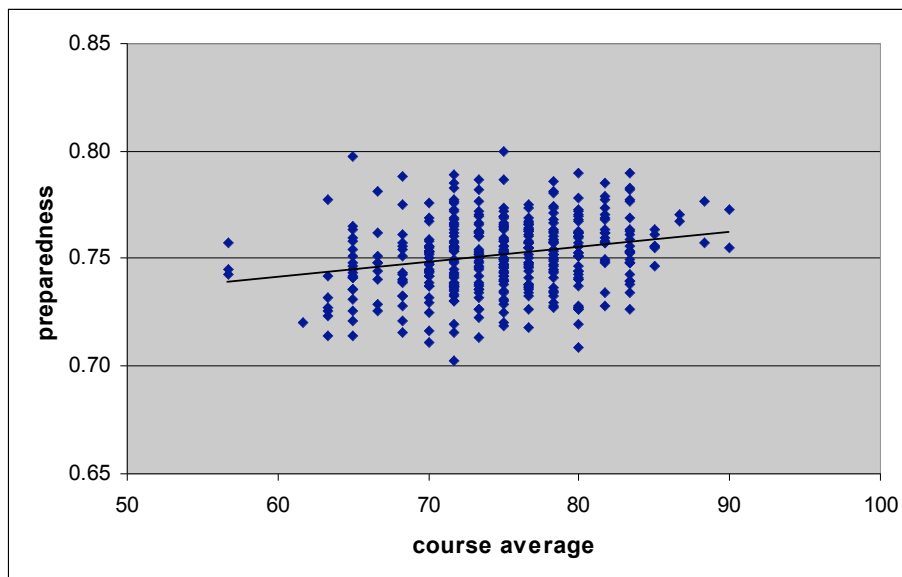
To simulate a harder exam for the top students, we will assign the random values for preparedness from a normal distribution with a mean of 0.75 rather than the previous 0.95. Thus the average grade on the exams should drop from around 95 to around 75. However the range in ability of the students did not change. We will make the modeling assumption that range of preparedness for the harder exam does not increase and hence we will use the same standard deviation as before and consequently the range of values will largely fall between 0.70 and 0.80 (rather than 0.90 and 1.00). Figure 5 contains the resulting frequency histogram for this simulation.

Figure 5. Frequency Histogram for Very Good Students Taking a Hard Exam



Now it appears that all is right with the world. The grades are normally distributed. The range is from the 60's to the 90's. Grading on a curve seems feasible at this point. Admittedly some outstanding students are going to receive low grades. But we have at least been able to distinguish among the performances of the students. But can we be confident that the better students got the better grades? Again we can calculate the correlation between student preparedness and the resulting grades. Figure 6 contains the results for the 400 students. In this instance the correlation has dropped to 0.23 with a 95% confidence interval of (0.15 to 0.31). We still have evidence that the correlation between preparedness and average scores is positive but it is certainly not strong.

Figure 6. Relationship Between Preparedness and Course Averages for Very Good Students



It is also of interest to begin to imagine the impact on specific (but still hypothetical) students. Our averagely prepared students (from among this group of top students) obtained averages from the 60's to the 90's and hence letter grades from D's to A's. Most notably, the two students with the highest level of preparedness (0.80) end up with averages of 65 and 75 and would get a C and a low B when graded on a curve. In the highly competitive environment getting a C or low B when you believed (correctly) that you were extremely well prepared for the exam can be very discouraging. Low B's are not going to be adequate for getting into good medical schools and many students will be sufficiently discouraged to drop out of the pre-med program. This "weeding out" is evidently a goal in some cases. However, it would be unfortunate if it is in fact the very best who are discouraged in this environment.

Modeling the Extreme Case

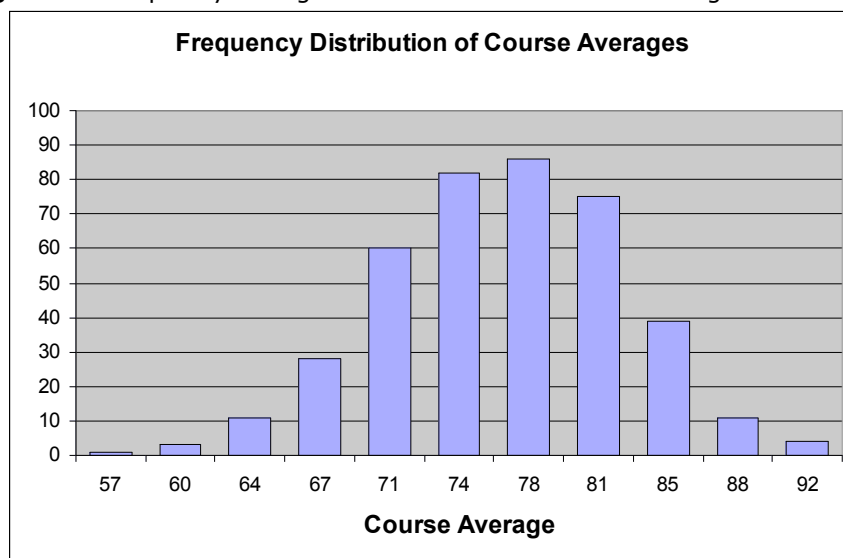
There is ample evidence that students drop out or are weeded out of the pre-med programs across the country. It is certainly the case in the two institutions we have used to provide the context of our simulation. When you compare the number of students in the pre-med sections of introductory chemistry to the number of students in the subsequent organic chemistry course, the drop is significant every year. Table 2 contains the counts for one year at the two universities.

Table 2. Course Enrollments in Pre-Med Chemistry Classes

	General Chemistry	Organic Chemistry
University A	1359	614
University B	701	361

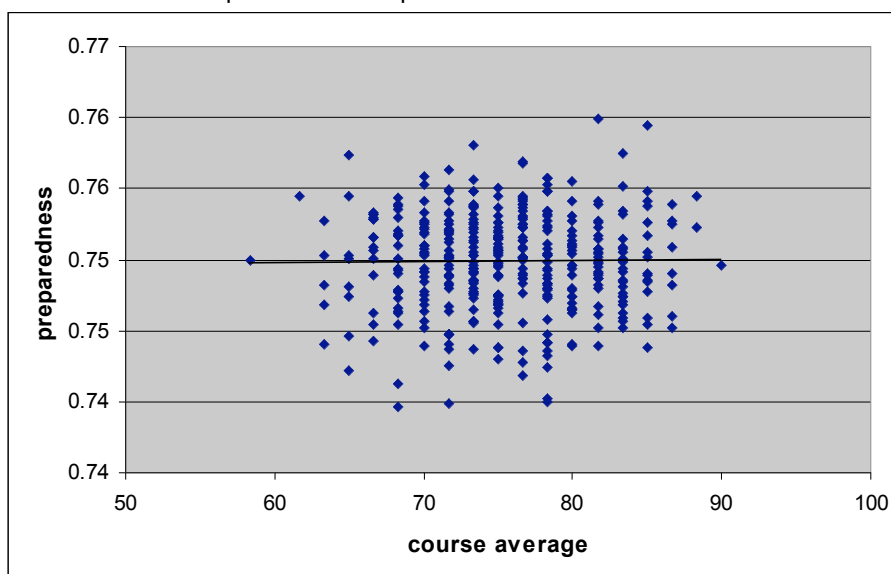
In each instance approximately one half of the pre-med students who take General Chemistry do not continue on to the Organic Chemistry class. Hence by the time we are grading these remaining pre-med students at these highly selective institutions we are certainly trying to distinguish performance between a group of students with very similar levels of ability and motivation. We model this most extreme case by selecting our random values for preparedness from a normal distribution with a smaller range, namely 0.74 to 0.76. We have assumed the instructors will continue to make the exams difficult enough to justify using a mean level of preparedness of 0.75. So essentially we are claiming that we now have a group of students in which the differences in ability or preparedness to take an exam are virtually indistinguishable. However, as illustrated in Figure 7, the exam results still show a nice normal distribution for our course averages, ranging from below 60 to the 90's.

Figure 7. Frequency Histogram for Excellent Students Taking a Hard Exam



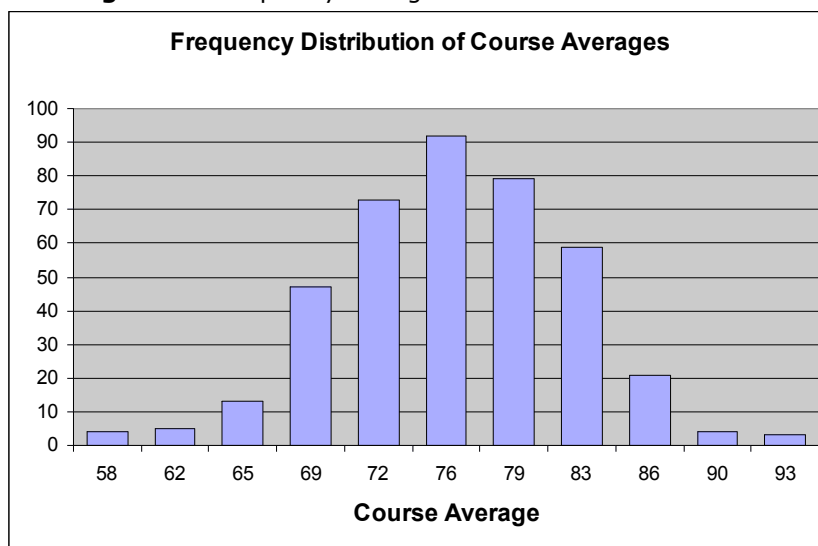
Greatly reducing the range of student ability did not reduce the range or grades or the likelihood that the grades would be normally distributed. Figure 8 contains the specific results for 400 students. Here, the calculated correlation is 0.01 with a confidence interval of (-0.07, 0.09). Consequently, we no longer have sufficient evidence that there is any correlation between preparedness of the very top students and their course average. In fact, in this particular simulation, the student with the highest level of preparedness ends up with a course average of 72, below the class average, and would receive a grade in the C+/B- range. Some students in the top 10% of the level of preparedness had course averages in the low 60's and could end up with C-'s or even D's, potentially fatal grades on organic chemistry for a pre-med student. It has become a matter of pure luck!

Figure 8. Relationship Between Preparedness and Grades for Excellent Students



What happens when we carry our model to the extreme and assume that all the students are precisely identical? We can certainly run one more simulation for this scenario using a mean student ability of 0.75 and a standard deviation of zero, and Figure 9 contains the resulting frequency histogram. Thus 400 identically prepared students taking three exams each made up of 20 questions will end up with grades ranging between 60 and 90. In this extreme case of 400 identical students we still observe an approximately normal distribution of test scores even though there is no variation in student ability (or preparedness). There is clearly no correlation between preparedness and test scores.

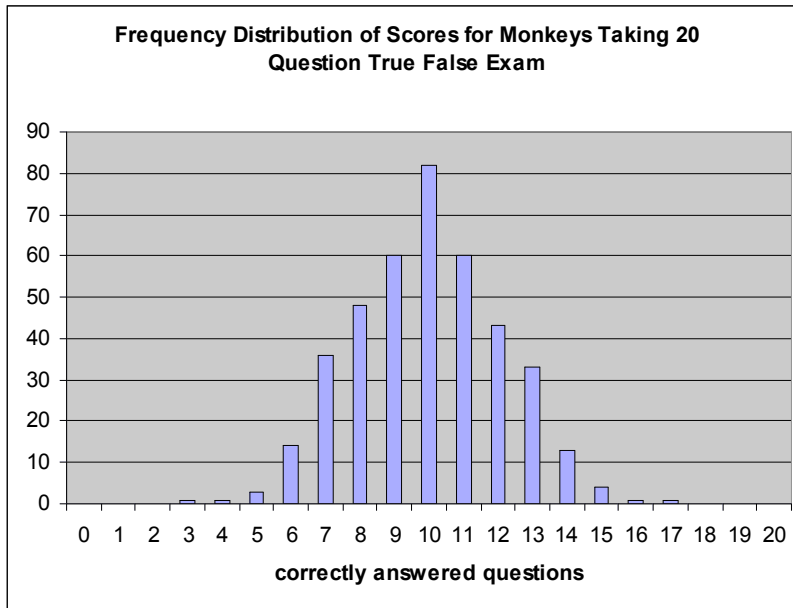
Figure 9. Frequency Histogram for Identical students



In scientific analysis, as well as in class room instruction, extreme cases often offer informative results. Perhaps an informative simulation would be one in which we administer a 20 question true false organic chemistry exam to 400 monkeys. It is a reasonable assumption that our monkeys have equal knowledge of organic chemistry. But, since it is a true false exam, each monkey has a 50% chance of getting any question correct. Some monkeys will be more lucky than others. Figure 10 contains the frequency histogram for the number of questions each monkey gets correct.

Of course our students are not monkeys. But our exams can approach levels that do no better job distinguishing among students, particularly as the task becomes more difficult when class populations are more homogenous. Merely making exams more difficult does not automatically improve the ability of the exam to rank student performance. Anecdotal evidence abounds of science exams with mean scores around 50% and with no students getting all the questions correct. Are these exams effective measures of learning or merely exams that will guarantee normally distributed test scores. As our monkey simulation illustrates, the more randomness and luck play a factor in these exams, the more likely the scores will automatically be normally distributed. Normal distributions alone are not evidence of an effective exam.

Figure 10. Frequency Histogram for Monkeys



Initial Observations

This analysis of extreme scenarios helps us to understand our problem in general. First our analysis focused on the extreme case of classes made up of the very best students. The scenario of the monkeys represents the complete extreme. The first and most obvious conclusion is that normally distributed test scores offer no independent evidence that the test has appropriately distinguished between the abilities of the test takers. Hence instructors cannot make the claim that just because they have test scores that are normally distributed they must have designed an exam that fairly distinguishes among student abilities. But although the results are most obvious in the extreme case, they also indicate that the problems apparent at the outermost levels of homogeneity, are also evident, all be it to a lesser degree, in all cases.

The extreme case brings the not so obvious to our attention. But our simplified models also provide potential insights into real life situations. This simulation model clearly provides the insight that the correlation between ability (as defined in the model) and final test averages is dependent on the both the mean and the standard deviation of the level of preparedness assigned to the 400 students. For a fixed mean (0.75) the correlation between ability and test scores declines as the standard deviation decreases as summarized in Table 3.

Table 3. Relationship between Standard Deviation and Correlation

Mean Preparedness	Standard Deviation of Preparedness	Correlation between Preparedness and Average Test Scores
0.75	0.083	0.81
0.75	0.017	0.23
0.75	0.004	0.01

Consequently the model suggests that as the students in a class become more similar in level of preparedness to take a test, the role of luck in determining their test score increases. Clearly the means and standard deviations in our model were selected somewhat arbitrarily. But the relationship between the standard deviation and correlation is unmistakable. Specific instructors can argue that as they make their exams more difficult they are also somehow making them such that the standard deviation of student preparedness also increases. However they cannot continue that argument forever. Clearly there is no ability to design an exam that will distinguish ability in the extreme case of a standard deviation of zero. And again, just as clearly, a normal distribution of test scores, by itself, provides no evidence of the exam's capacity to correlate grades with ability. Just as importantly, the model suggests that even when there is more variation in student ability, luck still plays a role that can affect some students significantly.

The primary conclusions of these simulations are best illustrated in the context of the extreme case of the outstanding students in large science classes. However the results can contribute to discussions about student assessment in all disciplines. In many fields text books come with test banks of multiple choice questions and instructors can randomly select the questions for their exams based on the chapters they have covered. There is certainly a high level of randomness associated with this approach. At the same time the results might well suggest that science faculty whose primary training for their profession took place in their research Ph.D. programs should reach out to assessment and evaluation experts to seek better tools for determining the validity of their testing schemes. None of us should any longer find comfort in the fact that our student test scores are normally distributed.

Assessing the Validity of the Simulation Model

As with any model, there can be questions about the validity of the assumptions that drive the model. In many cases we use our models to evaluate the extent to which we should consider more expensive testing in actual settings. Industrial chemists use simulation models to conduct lab experiments. Simulation results that show promise are then actually performed in the lab. As educators react that this simulation model raises concerns about grading on a curve many of the assumptions and conclusions of the model could be verified from real data. Given willing participation from enough institutions, the assumption that exams are designed to produce normal distributions of test scores can be investigated by looking at actual test scores in large sections for a variety of courses and universities. The model's actual conclusion that eventually there is little correlation between test scores and student preparedness could in part be investigated by measuring the correlation between test scores for individual students. If luck plays an increasing role in test scores there should be evidence that students' scores will vary significantly from one test to another. Conversely, the same students getting the higher scores all the time would argue against the conclusion of the model.

Our use of a value to represent the probability that a student will get a question correct appears to match reality for exams in science classes. In the typical environment it is easy to accept that most exams cannot include questions about everything that was covered in the class or in the text. Hence if a student has mastered only 75% of the material, there is a chance that the exam will include questions only from that 75% and the student could get a 100. Of course many students believe the luck only goes the other way. In the environment where the classes are filled with excellent students, the exams must be made so difficult that luck is based on whether the student happened to remember tiny obscure details. In fact it is not uncommon for these very difficult exams to have mean scores in the 50's or 60's with no students getting all questions correct. As our simulations illustrate, luck

becomes even more a factor in this case. Fully investigating these assumptions would probably require a more qualitative analysis based on interviews with students and faculty.

All of these investigations would be non-trivial. The power of the simulation model is that it suggests that further discussion and investigation is warranted, particularly given the impact the practice of grading on a curve can have on individual students.

Potential Impact on Students

The simulation results emphasize the role luck could play in student performance on exams. The resulting impacts on students can be substantial. In the case of the pre-med students at the highly selective institutions, those who are discouraged from pursuing a medical career are perhaps chosen through a random process. While grades are by no means a sole determinant of admission to med school, students who do poorly in initial courses will often eliminate themselves in the first two years of undergraduate work. And there is evidence (Crocker, Quinn, Karpinski, & Chase, 2003), that female students who have worked hard and fail to achieve the results that they expect are more likely to be effected than their male counterparts. There is also evidence (Schoon, Ross, Martin, 2007), to support that female students are more likely to drop out of highly competitive majors. Serious students who expect their high level of hard work and preparation to pay off are also more likely to be discouraged than those who have learned to lighten up and expect luck to play a role in assessment. Is it possible that this process in fact "weeds out" personality characteristics rather than academic ability? And if so, are we confident that these are the characteristics we wish to discourage? For these excellent students, don't we owe them a form of assessment that reduces luck and encourages all highly capable students to pursue the best possible use of their talents? Furthermore, we should also acknowledge that grading on the curve can introduce a luck factor to some degree in assessing all students (not just the top students).

Final Observations

The simulations described in this paper point out potential flaws in commonly used grading practices. Unfortunately it does not necessarily point to suggestions for addressing those flaws. Further simulations could be used to evaluate the design of alternative testing options. Some simple questions can be asked. For example, our analysis assumed that all the questions were equally difficult. Further simulations were run to see if varying the level of difficulty among questions had any significant impact. As an example, we tried exams made up of 5 easy question, 10 moderate questions, and 5 hard questions. In most cases the correlation between ability and grades did not change. It did have the impact of making the results for the easy exam for very good students more normally distributed but at the same time it lowered the correlation between ability and grades. One could test the impact of more questions or more exams, or a wider variety of difficulty. In the end, however, the ideal assessment tool will likely require a reduction in the random factor.

The simulations in this paper assumed that all test questions were graded as correct or incorrect. While the exam questions are seldom true/false, they are often multiple choice or problem solving exercises that are graded without partial credit. This is a common format for large introductory science classes and one used almost exclusively in the chemistry and organic chemistry classes at one of our case institutions. In part, grading on a correct/incorrect basis is assumed to be necessary due to the large class sizes. One possible improvement to the standard multiple choice exams could be to allow for the possibility of

second best answers (worth fewer points than the best answer). Simulation models could be used to investigate the extent to which the luck factor is reduced by this type of exam.

Roth (2000) made a connection between assigning grades and salaries. Salaries and in particular merit pay are often discussed in the context of the extent to which merit pay increase motivation for higher performance. Among those who argue in favor of merit pay, there is an understanding of the importance of the evaluation process in determining levels of merit. Quality control guru W. Edwards Deming (1986) introduced his famous Red Bead Game precisely to illustrate the extent to which randomness in evaluation systems can lead to highly inappropriate conclusions. Hence, perhaps our primary task is to look for forms of evaluation that minimize random factors and produce results that in the end would make us more comfortable evaluating students on an absolute level. Undoubtedly such forms of evaluation would require far more work than merely producing increasing difficult examinations. The results from this paper at least point out that normal distributions of test scores do not provide independent verification of effective education or effective evaluation of students. We certainly are obligated to all of our students to do better not just in our teaching but in our assessment of student performance as well.

References

- Abbott, W. (2008). The politics of grade inflation: a case study, *Change*, 40, 32-37.
- Crocker J., Quinn D., Karpinski, A., Chase, S. (2003). When grades determine self-worth: Consequences of contingent self-worth for male and female engineering and psychology majors. *Journal of Personality and Social Psychology*, 85, p 507.
- Deming, W.E., (1986), *Out of Crisis*, MIT Center for Advanced Engineering Studies, Cambridge, MA.
- Gordon, M. (2006). When B's are better. *The Chronicle of Higher Education*, August 11.
- Grading on a Curve*. (2007). Retrieved July 24, 2007 from The Volokh Conspiracy site: <http://www.volokh.com/posts/1149788426.shtml>
- Grading on a Curve*. (2007). Retrieved July 24, 2007 from the Young Mathematician Network site: <http://concerns.youngmath.net/story/2002/12/5/155959/298>
- Johari, S. and Sclove, S. (1976). Partitioning a distribution. *Communications in Statistics*, 5, 133-147.
- Maxwell, Y. (2007). Profs. split on curving class grades. *The Daily Free Press*, Boston University. Retrieved July 24, 2007 from <http://dailyfreepress.com>
- Michaels, J.W. (1976). A simple view of the grading issue, *Teaching Sociology*, 3, 108-203.
- Roth, W.F. (2000). Our grading system throws kids a curve, *The Education Digest*, 65, 27-31.
- SAT percentile ranks, 2007 college-bound seniors-critical reading+mathematics*. (2008). Retrieved June 15, 2008 from the College Board site: <http://professionals.collegeboard.com/data-reports-research/sat/data-tables>

Schoon, I., Ross, A, Martin, P. (2007). Science related careers: aspirations and outcomes in two British cohort studies. *Equal Opportunities International*, 26, p 129.

Wall, C. (1987). Grading on the curve. *InCider*, 5, 83-85.

Wedell, D., Parducci, A., Roman, D. (1989). Student perceptions of fair grading: A range-frequency analysis. *The American Journal of Psychology*, 102, 233-248.