A Case for Stricter Grading

Aaron F. Archer
Andrew D. Hutchings
Brian Johnson
Harvey Mudd College
Claremont, California 91711
{aarcher,ahutchin,bjohnson}@hmc.edu

Advisor: Michael Moody

Abstract

We develop a ranking method that corrects the student's grades to take into account the harshness or leniency of the instructor's grading tendencies.

We simulate grade assignment to a student based on the student's inherent ability to perform well, the student's specific aptitude for the course, the difficulty of the course, and the harshness or leniency of the instructor's grading.

We assume that we have access to a instructor's previous grading history, so that we can judge how harsh or lenient a grader each instructor is. After making this determination, we adjust each grade given by that instructor to systematically correct for that instructor's bias.

After correction, the student body has an aggregate GPA of approximately 2.7, corresponding to an uninflated grade of B−. The corrected GPA values do a considerably better job of accurately ranking the students by ability, especially for students in the bottom eight deciles.

Assumptions

1. We wish to evaluate students purely based on their ability to perform well in courses.

2. Each student has a quality attribute that is not directly measurable but influences the ability to do well in courses. The ideal ranking of students is by highest quality attribute.

The UMAP Journal 19 (3) (1998) 299–313.

3. The instructor has an accurate perception of each student's performance in a course.

4. The cause of high grades is lenient grading practices by the average instructor at ABC College.

5. A more lenient instructor tends to grade all students higher, not just students of a certain ability level.

6. Scholarship selection is completed in the first half of a student's undergraduate career, to allow her to enjoy the scholarships while she is still in school.

7. Because the students are early in their careers at ABC College, they are still taking primarily general education courses, rather than courses in their major. Therefore, we assume that they select courses randomly, and thereby we model the breadth of course selection across disciplines.

8. Since the students know that their grades are going to be adjusted to filter out the harshness of their instructors' grading, they do not gravitate toward courses taught by lenient instructors.

9. Each student has a varying aptitude for each course. Presumably, a student has more aptitude for courses in her major. But since general course requirements tend to be broad and these are the courses we are examining, we assume that a student's aptitude for a course is random.

10. Each course has an inherent difficulty. In an easy course it is difficult to differentiate the high ability students from the rest, whereas tougher material produces a greater spread of performances.

11. Instructors know when a course is difficult. Presumably all students (even the top ones) will attain a lesser mastery of more difficult material, but the instructor will take this into account when assigning grades.

12. The college is on a semester system and each student takes four courses per semester.

13. A student's performance in a course is not influenced by which other students are taking the course. Neither is the student's grade, since we assume that instructors do not grade the students in a given course on a curve but rather on some absolute standard of performance.

14. An instructor's harshness in grading does not depend on the course and remains constant over a period of several years. Data on instructors' grading histories are available.

15. All instructors rate a student's performance the same, but they have different standards for what grade that performance should earn.

Practical Considerations

The concept of a single quality attribute that describes each student is not one that plays well politically and in the media. Not many people would advocate that a student's overall ability to do well in courses can be accurately characterized by a single real number. Therefore, our adjusted measure of student ability should be some sort of adjusted GPA, which will be easier for a general audience to accept and understand. This does not present a problem from the modeling point of view, as long as we know how quality rankings correspond to GPA values, and vice versa.

Ultimately, as we construct our model, we will run into a fundamental grading problem. The average grade at ABC College is an A−, which corresponds to a 3.67 GPA. Grade point averages that are this high result in very uninteresting grade distributions. The majority of the grades must be A+, A, or A−. In other words, if we look at the transcript of any above-average student at ABC, we will probably see a page full of A+, A, and A− grades. In this kind of environment, it will be extremely difficult to pick out the top few students, because the top half of the school is separated by only about 0.6 grade points. In contrast, the bottom half of the school is spread over the remaining 3.67 grade points, so it will be much easier to rank them by ability.

One radical solution to this dilemma is to require additional feedback on student performance from the instructors. We outline one possible system here, before we move on to a less radical approach. In addition to giving grades on the usual A to F scale, we could require an instructor to give each student a ranking between 1 and 10. At least one student in each course must receive a 1, and at least one student must receive a 10. This forces a spread in the instructor's rankings, so that even an easy-grading instructor (all A+ grades) must rank the better-performing students above the less able students. Next, the instructor is allowed to give a context to the scale. If the instructor has taught that course before, she would be asked to rate the current course in terms of previous ones. We ask the instructor to identify, on some absolute scale of ability, which interval corresponds to the 1 to 10 relative scale for the course. For example, if the instructor felt that her best student was about as competent as a student at the 90th percentile, then she would identify the right endpoint of the scale with the 90th percentile of absolute student ability. If she felt that her worst student was the poorest student to attend the college over an entire five-year span, then she would identify the left end of the relative scale with that point on the absolute scale. This two-stage evaluation system forces the students to be differentiated by performance puts the measures of performance into an absolute (rather than instructor-dependent) context.

What Characterizes a Good Evaluation Method?

As we attempt to rank the students at ABC College, we assume that the students have underlying quality scores that are reflected in their grades. We try to approximate the ranking induced by the hidden quality values. It may be inappropriate (for political reasons) to refer to our rankings as "estimated student qualities," so we instead calculate an adjusted GPA.

As we calculate adjusted GPA values, we keep in mind several goals:

- We wish to allocate correctly the available scholarships to the top 10% of the student body. To test whether or not we succeed, we must compare the ranking induced by our adjusted GPA values with the actual ranking of the students by intrinsic quality. Our first measure of the accuracy of our adjusted GPAs will just be the number of scholarships that we correctly allocated to deserving students.
- If the top-ranked student somehow fails to receive a scholarship, this is considerably more unjust than if a student who just barely deserves a scholarship misses out. Thus, we compute a second measure of accuracy by summing the severity of the mistakes made in awarding scholarships.
- It is important for all of the student rankings to be accurate, not just the top 10%, because they are used for much more than just scholarship determination. For instance, class rank is often cited in graduate school and job applications. Therefore, we consider a third measure of accuracy that gives a total error measure for our entire set of adjusted GPA rankings, rather than for just the top decile.

Modeling College Composition and Grade Assignment

According to Assumption 13, we do not need to consider the other students in a course when we determine a student's performance in the course and the grade the student receives; in other words, the composition of students in the course does not significantly affect the students' ability to learn, and none of the instructors grades on a curve. Thus, we model a student's grade as a function of

- her inherent quality,
- her aptitude for the specific course,
- the difficulty of the course, and

- the harshness of the instructor grading the course.

We treat each of these quantities as real-valued random variables and generate their values by computer.

We let q_i denote the inherent quality of student i. We will consider q_i to be distributed normally with mean 0 and standard deviation σ_q. This is reasonable, since we know that the normal distribution gives a good approximation for many characteristics of a large population.

We let $c_{i,j}$ represent the random course aptitude adjustment for student i when she takes course j. Again, it makes sense to let $c_{i,j}$ be normally distributed about 0, and we denote the standard deviation of this aptitude adjustment by σ_c. We let the net aptitude of student i in course j be $q_i + c_{i,j}$, which is normally distributed with mean 0 and standard deviation $\sqrt{\sigma_q^2 + \sigma_c^2}$. We choose our unit of measure so that $\sigma_q^2 + \sigma_c^2 = 1$. Furthermore, we estimate that a student's intrinsic quality influences her success at least five times as much as her aptitude adjustment for the particular course she is taking. Hence, we choose $\sigma_c < 0.2$.

Next, we consider how the difficulty of a particular course affects the grades that the instructor gives. We assume (see Assumption 10) that a difficult course spreads out the distribution of grades given; this means that poor students tend to do worse in difficult courses, but also that excellent students will do better, since they are being given an opportunity to excel. Conversely, in an easy course, the grades tend to bunch closer together, since the poor students are being given an opportunity to excel and the best students' performances are limited by the ease of the subject matter. Let d_j denote the difficulty of course j; then this interpretation leads us to consider a performance rating $N_{i,j}$ of student i in course j given by

$$N_{i,j} = (q_i + c_{i,j})d_j,$$

where d_j is a positive number, equal to 1 for a course of average difficulty, greater than 1 for a difficult course, and less than 1 for an easy course. Note that we are assuming the performance of a student in a course is random only in that the student's inherent ability is modified by a random aptitude adjustment factor. Once this factor is applied, the student's performance is determined, given the difficulty of the course.

Finally, we must take into account the grading philosophy of the instructor. Notice that a difficult course does not shift the performance distribution to the left, because the performance is measured relative to the instructor's expectation. We assume that the instructor is aware of the difficulty of the course and compensates accordingly in grading. This brings up a delicate distinction. An instructor's harshness does not reflect her expectation level but only her tendencies in grading. That is, we assume that the instructor's harshness does not pertain to her assessment of a student's performance but rather to what grade she thinks that performance deserves.

Let h_k denote the harshness of instructor k; then we should let the student's grade depend on $N_{i,j} - h_k$, since the harshness causes a systematic bias in all

of the grades that the instructor gives. We let a harshness of 0 correspond to an average instructor at an institution without grade inflation. At Duke University and presumably at other institutions, 2.7 was the average GPA prior to the grade inflation that began to appear in the 1970s [Gose 1997]. Therefore, letting G denote the grading function that maps real numbers to discrete letter grades, we should center $G(0)$ on a grade of B$-$. Furthermore, among instructors who grade on a curve, an interval of one letter grade is often equated with one sample standard deviation in the course scores. So, we let one standard deviation for a course of average difficulty correspond to a whole letter grade in our model. Thus, the instructors in our model are grading on a virtual curve; that is, they grade on an absolute standard that simulates grading on a curve in a hypothetical course in which the full distribution of students is enrolled.

This analysis leads to a grading function

$$G: \quad \rightarrow \{0, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13\}$$

defined by

$$G(x) = \begin{cases} 0, & \text{if } x \leq -\frac{11}{6}; \\ 3, & \text{if } -\frac{11}{6} < x \leq -\frac{9}{6}; \\ 4, & \text{if } -\frac{9}{6} < x \leq -\frac{7}{6}; \\ \vdots & \\ 12, & \text{if } \frac{7}{6} < x \leq \frac{9}{6}; \\ 13, & \text{if } x > \frac{9}{6}, \end{cases}$$

where the values $0, \ldots, 13$ represent the letter grades F, D, D+, C$-$, C, C+, B$-$, B, B+, A$-$, A, A+. To convert the numeric value to grade points, we divide by 3; thus, an A$-$ average means a GPA of 3.67. If student i takes course j taught by instructor k, she will receive a grade of

$$G(N_{i,j} - h_k) = G\big((q_i + c_{i,j})d_j - h_k\big).$$

For convenience, we define $l(g)$ and $r(g)$ to be the left-hand and right-hand endpoints of the interval on which $G = g$. For instance, $r(13) = \infty$ and $l(12) = \frac{7}{6}$.

A simple calculation reveals that in a course where $d = 1$ and $h = 0$, the expected grade is 2.63 (see **Table 2** and **Figure 1**). We intend harshness 0 to represent a reasonable level of strictness in grading. It centers the grades at B$-$, which is the exact middle of all passing grades, and yields a GPA in line with the "reasonable" historical number of 2.7 at Duke University.

We can visualize the grading method by graphing a $\text{normal}(0, 1)$ density function, which represents N (in the case where difficulty is 1), with the x-axis partitioned into intervals representing grades according to the grading function G (see **Figure 2**). A difficult course spreads out the distribution, resulting in more Fs (because the poor students cannot keep up) and more As (because the top students have an opportunity to shine). A positive (negative) harshness effectively shifts the grade intervals to the right (left).

Table 1.
Symbol table.

$c_{i,j}$	course aptitude adjustment for student i taking course j
d	estimate of course difficulty
d_j	difficulty of course j
G	grading function, from performance rating to letter grade
$\overline{g}$	average grade given by instructor, from historical data
g_{adj}	adjusted grade that an instructor of harshness zero would give
h_k	harshness of instructor k
I	interval in which student's performance value is estimated to lie
$l(g), r(g)$	endpoints of performance rating interval corresponding to letter grade g
$N_{i,j}$	performance rating of student i in course j
N_{est}	estimate of student performance value
$N_{i,\mathrm{est}}$	estimate of student i's performance value
Φ	standard normal cumulative distribution function
q_i	inherent quality of student i
$q_{i,\mathrm{est}}$	estimate of inherent quality of student i
σ_q	SD of inherent quality of student i
σ_i	SD of course aptitude adjustment of student i taking course j
$\sigma(d,h)$	SD of grades given by a professor of harshness h in a course of difficulty d
t_0	left endpoint for performance rating interval corresponding to grade of A+
t_1	right endpoint for performance rating interval corresponding to grade of F

Table 2.
Expected grade on a 4-point scale as a function of course difficulty d and instructor harshness h.

	d										
h	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5
−3.0	4.33	4.33	4.32	4.31	4.30	4.29	4.27	4.25	4.23	4.20	4.18
−2.8	4.33	4.32	4.31	4.30	4.28	4.27	4.24	4.22	4.19	4.16	4.13
−2.6	4.32	4.31	4.30	4.28	4.26	4.24	4.21	4.18	4.15	4.12	4.08
−2.4	4.31	4.30	4.28	4.25	4.22	4.19	4.16	4.13	4.10	4.06	4.03
−2.2	4.29	4.27	4.24	4.21	4.18	4.14	4.11	4.07	4.03	4.00	3.96
−2.0	4.26	4.22	4.19	4.15	4.11	4.08	4.04	4.00	3.96	3.92	3.88
−1.8	4.19	4.15	4.11	4.07	4.03	3.99	3.95	3.91	3.87	3.83	3.79
−1.6	4.10	4.06	4.02	3.98	3.94	3.90	3.86	3.82	3.78	3.73	3.69
−1.4	3.97	3.94	3.90	3.86	3.82	3.78	3.74	3.70	3.66	3.62	3.58
−1.2	3.82	3.79	3.76	3.72	3.69	3.65	3.62	3.58	3.54	3.50	3.46
−1.0	3.64	3.62	3.60	3.57	3.54	3.51	3.48	3.44	3.41	3.37	3.33
−0.8	3.45	3.44	3.43	3.41	3.38	3.35	3.32	3.29	3.26	3.23	3.19
−0.6	3.26	3.25	3.24	3.23	3.21	3.19	3.16	3.13	3.10	3.07	3.04
−0.4	3.06	3.06	3.05	3.04	3.03	3.01	2.99	2.96	2.94	2.91	2.88
−0.2	2.86	2.86	2.86	2.85	2.84	2.82	2.80	2.78	2.76	2.74	2.72
0.0	2.66	2.66	2.66	2.65	2.64	2.63	2.61	2.60	2.58	2.57	2.55
0.2	2.46	2.46	2.46	2.45	2.44	2.43	2.42	2.41	2.40	2.39	2.38
0.4	2.26	2.26	2.25	2.24	2.23	2.23	2.22	2.21	2.21	2.20	2.20
0.6	2.06	2.05	2.04	2.03	2.03	2.02	2.02	2.02	2.02	2.02	2.02
0.8	1.85	1.84	1.83	1.82	1.81	1.81	1.82	1.82	1.83	1.84	1.85
1.0	1.63	1.62	1.61	1.60	1.60	1.61	1.62	1.63	1.64	1.66	1.67

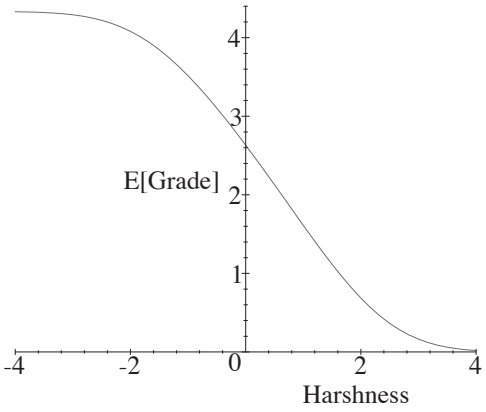

Figure 1. Expected grade on a 4-point scale as a function of instructor harshness h, given that course difficulty $d = 1$.

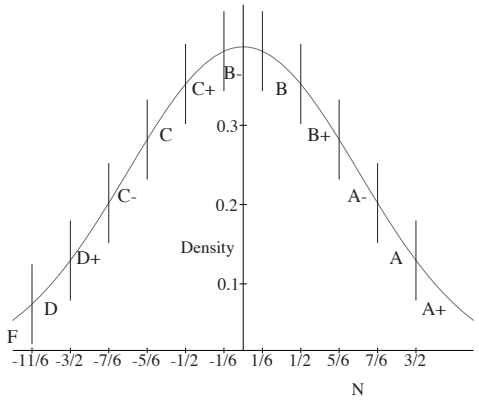

Figure 2. The probability density of the performance variable N. The vertical bars represent the grade ranges for an instructor of zero harshness. For $h > 0$, the ranges shift to the right by h, making it harder to earn a high grade.

Since we have no data on the students, courses, instructors, or grades at ABC College, we generated a random set of students, courses, and instructors. Since we don't know the exact composition of the college, we generated various scenarios.

- We assigned each instructor to teach five courses per year, which is a typical teaching load at many colleges.
- We computed a quality variable q for each student by sampling randomly from a $\mathrm{normal}(0, \sigma_q)$ distribution.
- We assigned each student to eight courses per year (also a standard load for many colleges), for either one or two years, using a uniform probability of selecting each course.
- We generated course aptitude adjustments c for each course that a student enrolled in by sampling from a $\mathrm{normal}(0, \sigma_c)$ distribution.
- We assigned difficulties to each course by sampling from a symmetric beta distribution centered at 1. A typical choice would be $\mathrm{beta}(3, 3)$ on $[0.7, 1.3]$ (see **Figure 3**).
- We assigned harshnesses to instructors by sampling from an asymmetric beta distribution skewed and translated towards leniency (to represent the tendency to inflate grades at the college). A typical choice would be $\mathrm{beta}(2, 3)$ on $[-2, 0]$ (see **Figure 4**). One can use **Table 2** to guide the choice of distribution according to the average GPA that we desire.

We did not restrict ourselves to considering the case where the average GPA at the college is 3.67. In early 1997, Duke University considered revising

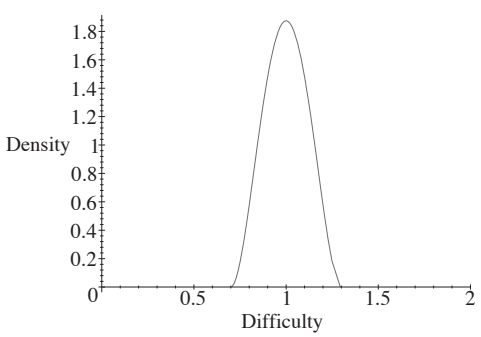

Figure 3. Typical course difficulty distribution: $\mathrm{beta}(3, 3)$ on $[0.7, 1.3]$.

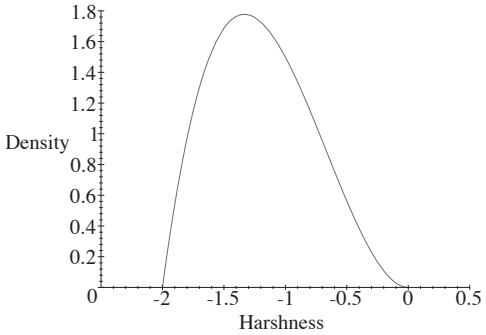

Figure 4. Typical instructor harshness distribution: $\mathrm{beta}(2, 3)$ on $[-2, 0]$.

its calculation of GPAs to take into account the difficulty of the course in which a grade was received, the quality of the other students in the course, and the historical grading tendencies of the instructor. The reason for this move was alarm that the average GPA at the University had risen from 2.7 in 1969 to 3.3 by fall 1996 [Gose 1997]. If an average GPA of 3.3 is considered evidence of rampant grade inflation, then 3.3 is a more likely estimate of the average GPA at ABC College than 3.67 is.

Some word on our rationale for choosing distributions is in order here. The normal distribution is a standard choice for representing abilities in a population. Our model of how course difficulty affects student performance and grades loses validity for d outside the range of approximately $[0.5, 2]$, and of course a negative difficulty makes no sense at all. Thus, the normal distribution is not an appropriate choice. We chose a beta distribution because it takes values on a finite interval. Similarly, any harshness value outside the range $[-3, 2]$ is patently ridiculous (see **Figure 1**). In fact, a harshness value of -2 is fairly ridiculous; but to obtain a school-wide average GPA of 3.67, we have to allow that some instructors are that lenient. In any event, it behooves us to choose a distribution over a finite interval. We also desire an asymmetric distribution, to represent the tendency at the college toward lenient grading. Thus, a $\mathrm{beta}(a, b)$ distribution with $a < b$ is appealing for our purposes.

Our model indicates that the ABC College administrators' concerns about being unable to distinguish among the top students are justified. Indeed, when we generate an incoming freshman class of 500 students and make the instructors lenient enough to yield a 3.64 average GPA, 60 of these students *still have a straight A+ average after two years!* In this environment it is clearly necessary to search for a better evaluation method than simple GPA.

The Modified GPA algorithm

Our algorithm for establishing a class rank involves a number of distinct stages. We first attempt to gain additional information from the instructor's historical grade awards and use this information to refine our knowledge of the courses and instructors that the students are taking currently. Using an estimate of the instructor's harshness based on the grades that he has given historically, we correct the grades awarded in a particular course by estimating the mean value by which any leniency or harshness changed a student's letter grade. Incorporating this correction factor allows us to provide an adjusted GPA measure that represents more fully the actual performances and, hence, the quality of the students.

We assume that the instructors have been teaching at the college for at least two years and that we have access to the grades that they assigned during those years. We simulate these data just as we generated the data for the students, as described in section **Modeling College Composition and Grade Assignment**. New students are generated randomly as in the original student data. Given the actual harshness of the instructors and the difficulties of the courses taught, numerical performances and grades for all of each instructor's courses are generated as above.

From these historical data, we compute the average grade $\overline{g}$ granted by each instructor. One can calculate the expected grade granted in a course as a function of difficulty d and harshness h (see **Table 2** and **Figure 1**). For a given value of d, this function decreases monotonically with h, so we can calculate the inverse function. Assuming that $d = 1$ we estimate the instructor's harshness h_{est} by evaluating this inverse function at $\overline{g}$.

Notice that we never even try to estimate the difficulty or take the actual grade distribution into account. Despite this crude method of estimating harshness, we achieve surprisingly good results. Using courses of 40 students each, the harshness that we estimate is usually within about 0.05 of the actual harshness, though it is not too uncommon to err by as much as 0.12. The error tends to decrease the closer the actual harshness is to zero.

We can now adjust the grades of the students in each of an instructor's courses based on our harshness estimate h_{est} for that instructor. For simplicity we assume $d = 1$ for the course. The fact that a student receives a grade g in a course with a instructor of harshness h means that the student's performance value N lies in the interval

$$(l(g) + h,\ r(g) + h]\ .$$

Thus we estimate that N lies in the interval

$$I = (l(g) + h_{\text{est}}, r(g) + h_{\text{est}}]\,.$$

We estimate N to be the expected value of the distribution of N given that N lies in I. For grades other than A+ and F, this interval has width $\frac{1}{3}$. Assuming

$d = 1$, the a priori distribution of N is standard normal. Given that it lies in I, the probability density function is just the indicator function for I times the standard normal density times a constant factor. Since the density function for the standard normal distribution is almost linear over any interval of width $\frac{1}{3}$, we estimate N as

$$N_{est} = h_{\text{est}} + \frac{l(g) + r(g)}{2}.$$

If the grade is A+ or F, we can calculate the expected value analytically, with the following results:

$$E[N|N > t_0] = \frac{e^{-t_0^2/2}}{\sqrt{2\pi}(1 - \Phi(t_0))} \quad \textbf{(1)}$$

$$E[N|N \leq t_1] = -\frac{e^{-t_1^2/2}}{\sqrt{2\pi}\Phi(t_1)} \quad \textbf{(2)}$$

where Φ is the standard normal cumulative distribution function and the relevant t values are $t_0 = l(A+)$ and $t_1 = r(F)$. So when the student's grade is A+ or F, we set N_{est} to **(1)** or to **(2)**, respectively.

Now that we have a value for N_{est}, we assign the student an adjusted grade for the course. The adjusted grade g_{adj} is the grade that an instructor of zero harshness would have given, except that we assign a real number grade instead of an integer. Specifically,

$$g_{\text{adj}} = 3N_{\text{est}} + 8.$$

To avoid a discontinuity at $N_{\text{est}} = r(F)$, we treated an F as a grade of 2 for this purpose.

Since the grades given by a lenient grader exhibit a smaller spread and hence do not differentiate the students as well as those given by a strict grader, we grant them less import when calculating a student's adjusted GPA. Specifically, the student's GPA is a sum of the grades received weighted by $\sigma(1, h_{\text{est}})$, where h_{est} is the estimated harshness of the instructor who assigned the grade and $\sigma(d, h)$ is the standard deviation of grades given by a instructor of harshness h in a course of difficulty d (see **Figure 5**).

If we wished to refine this method, we might use the spread of grades in a course to estimate its difficulty both when examining the instructors' grading histories and when adjusting grades at the end. However, even without this enhancement, the basic method that we have just outlined performs well, as we demonstrate in **Results of the Model**. One has to be very clever to estimate the difficulty of a course in a way that is numerically stable. The most obvious method is to note that for student i, we have $N_i = q_i d$, then use $N_{i,\text{est}}$ and some estimate $q_{i,\text{est}}$ of q_i that we pull from some other source to estimate

$$d \approx \frac{N_{i,est}}{q_{i,est}}.$$

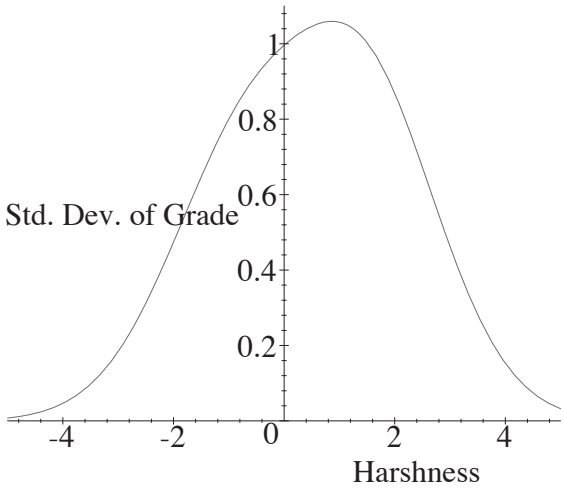

Figure 5. Standard deviation of grade distribution as a function of h, assuming difficulty $d = 1$.

The difficulty is that q_i is likely to be near zero, so the error $|q_i - q_{i,\text{est}}|$ is magnified when we divide.

Results of the Model

We generated a number of scenarios to elucidate the features both of our simulated data and of the modified GPA algorithm. We ran these simulations on a test student population of 500 students, each of whom took 16 courses, with average course size 40.

We use a number of these scenarios to demonstrate the results of our simulations. Plots of actual quality ranks vs. rank by GPAs or adjusted GPAs demonstrate the effectiveness of each ranking method over all tiers of students. For a perfect ranking of students, this plot would lie along the line $y = x$.

We define three error metrics to aid in the comparison between the ranking generated by our revised GPA method and the raw GPA ranking.

- We define a *misassigned scholarship candidate* to be a student who either received a scholarship but was not in the top 10% in quality, or who was in the top 10% of quality but did not receive a scholarship. A simple count of the number of misassigned scholarship candidates measures the method's effectiveness at identifying the highest caliber of student. We refer to this quantity as the MS (Missed Scholarship) metric for a given estimated ranking.

- For each student who is ranked incorrectly, the rank errs by some number of places. Summing these rank errors over all students gives us a measure of how our ranking compares to the actual quality ranking across the entire spectrum of students. We refer to this as the SE (Scaled Error) metric.

- Finally, to determine the injustice with which scholarships are assigned, we sum over all misassigned scholarship candidates the distance between their quality ranking and the scholarship cutoff rank. We refer to this measure of error as the SI (Scholarship Injustice) metric.

The first scenario has difficulty scaled to be between 0.7 and 1.3, while the harshness distribution is relatively lenient, with values ranging between -2.1 and -0.1. The variation due to course material is set to be 0. This yields, as one might expect, a student population with rampant grade inflation. Overall GPA is 3.64, with 60 students receiving perfect A+ averages. **Figure 6** plots GPA rank against actual quality rank, where we observe significant discrepancies between the estimated and actual rankings. At this level of grade inflation, the top tiers of students are almost entirely indistinguishable by GPA. Attempting to correct for harshness by using the corrected GPA does not significantly improve the results at the high ranks. It does, however, improve the SE number from 9,124 to 5,352, representing a superior evaluation of the middle and lower tiers of students (see **Figure 7**).

We now alter the parameters in our model to fit what we feel is a far more realistic situation. Harshness is set to vary between -1.569 and $.431$, yielding a scenario that has an average GPA of 3.34. Now only 38 students have perfect $A+$ averages. The effect of ranking based on the adjusted GPA is definitely apparent, in **Figures 8–9**. The discrepancies are clearly less for middle-ranked and low-ranked students. The SE measure improves from 6,916 using simple GPAs to 4,842 using adjusted GPAs.

Note the loss of ranking accuracy at high quality levels. The two methods of ranking perform nearly identically, with raw GPA giving an MS of 6 while the adjusted GPA rank gives an MS of 7.

Table 3 summarizes results for a sampling of the scenarios, all of which use the usual range of difficulty from 0.7 to 1.3.

As our simulations show, the modified GPA ranking fares well against the raw GPA method, with SE numbers significantly lower in each trial. This means that the students in the lower deciles are ranked much more accurately in each case. This suggests a certain robustness and indicates that judgments based on the modified GPA rank will be more fair.

As in all other trials performed, both the raw and adjusted GPA ranking methods performed poorly at the high end of the ability curve, according to all three measures.

Strengths and Weaknesses

One weakness of our model is that it does not allow for a completely analytic solution to the scholarship selection problem. Computer simulation is the only means we have to test and evaluate our methods against the simple-minded raw GPA ranking method.

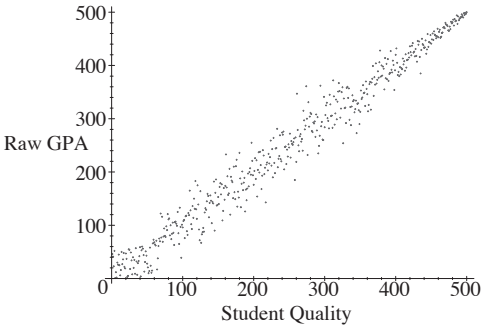

Figure 6. Wild grade inflation resulting in an average GPA of 3.64. The raw GPA estimate makes significant mistakes in the entire range but is especially inaccurate in the top two deciles.

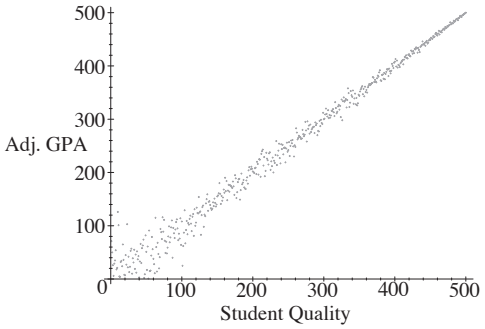

Figure 7. The same scenario as **Figure 6** with rank determined by GPAs modified for instructor harshness.

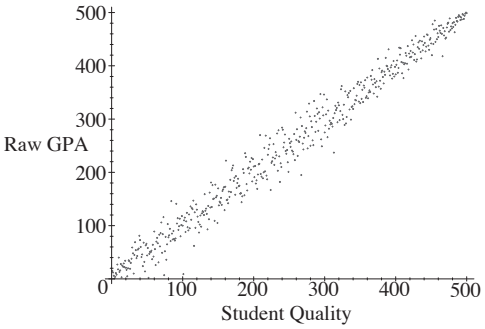

Figure 8. A more reasonable scenario. The raw GPA rank maintains some level of inaccuracy throughout the spectrum of student ability.

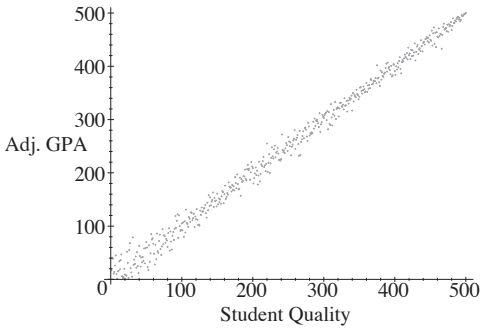

Figure 9. Same scenario as **Figure 8** but with ranks determined by the modified GPAs.

Table 3.

The relevant information for several simulations. Note how the modified GPA ranking produces smaller SE numbers in all cases, representing greater overall accuracy.

Trial	Harsh low	Harsh high	Raw GPA	Raw MS	Adj. MS	Raw SE	Adj. SE	Raw SI	Adj. SI
1	−2.1	−0.1	3.64	9	16	9124	5352	139	118
2	−1.55	0.35	3.22	8	10	7178	3122	247	596
3	−1.569	0.331	3.34	6	7	6916	4842	215	197
4	−1.9	0.1	3.48	8	10	8896	5424	101	226
5	−1.49	0.51	3.19	8	5	7454	4410	153	81

Potentially the greatest weakness in our model and techniques is the lack of a good ranking of the top two deciles. Whether or not a robust method exists is, we believe, debatable. Nothing we have seen indicates that the information required to form a confident ability ranking is even contained in the GPA information we have. It is likely that complete rank-ordering cannot be achieved given the information our model provides. We do not know this to be true, but it is certainly consistent with the results we have witnessed.

Another weakness is that our model cannot take into account the effect of curved grading systems and the possibility of student grades being altered by the performances of the fellow students in the course. Similarly, other interactions between the entities in our model, such as the formation of study groups, can affect the performance (as distinguished from grade) of a student in a course in a way that is dependent upon the other members of the course. Our model also includes parameters, namely, the course difficulties, that are difficult to estimate accurately, and thus remain a complete unknown throughout our attempts to rank based on ability.

In spite of these features, our model has a number of compelling features. By changing just a few parameters, one can generate an entirely new scenario that has a plausible distribution of grades and GPAs. Furthermore, it takes into account the three functional parts of any educational experience: the students, the instructors, and the courses. Arguably, no model could be complete without accounting for variations of each of the three parts.

Despite all of our problems in classifying the scholarship winners, the adjusted GPA method we use is almost uncannily good at identifying the lower deciles, which in a real context is important to the students and the school.

From a practical standpoint, our model and methods are fairly simple to implement. The number and size of the calculations performed is linear in the size of the student body and could be executed with modest computer resources at even a large institution.

To sum up, it would behoove ABC College to use our ranking system, since it more accurately identifies the bottom eight deciles of student ability. However, if the administration seeks to accurately rank the top tier of students, it must realize that a bloated aggregate GPA from excessively lenient grading can quickly lead to a situation where no amount of calculations and statistics can recover the desired information about the intrinsic quality of the students.

References

Gose, Ben. 1997. Duke rejects a controversial plan to revise the calculation of grade-point averages. *Chronicle of Higher Education* 43 (21 March 1997): A53.