Grade Inflation

Team 226

February 9, 1998

1 Restatement of the Problem

The A Better Class (ABC) College needs to rank its students to determine the winners of a generous merit scholarship which is only awarded to students among the top 10%. Unfortunately, due to grade inflation, the average grade given at ABC College is an A-. Traditional GPA's are thus nearly meaningless, since so many students have practically the same GPA, with so many A's and A-'s given out. The traditional GPA also punishes students for taking difficult courses, especially when the grade average is so high. One lower grade from a difficult course can make a student fall in class rank behind students who take only easier courses. The task is to devise a method that will separate and rank the students, so that the scholarship may be fairly awarded.

The dean of the college thought that comparing each student to the other students in each course would be an effective way to build a ranking. Each grade would be compared against other grades from the course to determine if a student was above average, average, or below average in the class. Combining the information from all courses could allow students to be ranked in deciles.

The problem has four major questions to be answered:

- Assuming that the grades given out have pluses and minuses, can the dean's idea be made to work?
- Assuming that the grades given out are without pluses and minuses, only flat letter grades, can the dean's idea be made to work?
- Can any other schemes produce a desired ranking?
- A concern is that the grade in a single course could change many student's deciles. Is this possible?

To avoid confusion, we will use the following definitions for ambiguous words. A "class" is a group of students who all graduate at the same time, for example, the class of 1999. A "course" is a group of students being instructed by a professor, who assigns a grade to each student.

2 Further Considerations

The method should reward students a greater amount for scoring well in more difficult courses than in easier courses. It would also be positive to discourage grade inflation. At the same time, the college may wish to avoid discouraging students from taking courses outside their majors, for fear of scoring worse than students who major in those courses' departments.

3 Assumptions and Hypotheses

The following assumptions were made when tackling the problem:

- It is possible to assign a single number, or "ability score", to each student, which indicates their relative scholastic ability and, in particular, their worthiness of the scholarship. In other words, we can rank students.
- The rank should be transitive; that is, if X is ranked higher than Y, and Y is ranked higher than Z, then X should be ranked higher than Z. We can therefore sort the students by rank.
- The performance of an individual student in all courses is positively correlated, since:
  - There is a degree of general aptitude corresponding to the ability score which every student possesses.
  - All professors, while their grade averages may differ, rank students within their courses according to similar criteria.
- While there may be a difference between grades in courses which reflects the student's aptitude for the particular subjects, this has only a small effect, because:
  - Students select electives in a manner highly influenced by their skill at the subjects available, i.e., students tend to select the courses at which they are most talented from.
  - All students should major in an area of expertise, so that they are most talented at courses within or closely related to their majors.
  - Courses which are required by the college reflect an emphasis on the part of the college; i.e., even if the required courses are "unfair" because they are weighted towards one subject, e.g., writing, that is taken to be the choice of the college, that highly ranked students must do well on the required courses.
- Not all courses have the same difficulty. That is, it is easier to earn a high grade in some courses than in others.

- The correspondence of grades to grade points is as follows: $A = 4.0$, $B = 3.0$, $C = 2.0$, $D = 1.0$, $F = 0.0$. A plus following a grade raises the grade point by one-third, while a minus lowers it by the same amount (i.e., an $A- \approx 3.7$, while a $C+ \approx 2.3$).
- Students take a fixed courseload for each semester for eight semesters.
- The average grade given at ABC College is A-. Thus we assume that the average GPA of students is at least 3.5, the smallest number which rounds to an A-.
- In general, student X should be ranked ahead of student Y if:
  - X has better grades than Y.
  - X takes a more challenging courseload than Y.
  - (We recognize that this point is debateable) X has a more well-rounded courseload.

4 Analysis of Problem and Possible Models

4.1 The Problem with Traditional GPA Ranking

The traditional method of ranking students, commonly known as the Grade Point Average, or GPA, consists of taking the mean of the grade points that a student earns in each of his courses, and then comparing these values to determine class rank.

The immediate problem with the traditional GPA ranking is that it does not sufficiently distinguish between students. When the average grade is an A-, all above-average students within any class receive the same grade, A. Thus, with only four to six classes per semester, fully one sixth of the student body can be expected to earn a 4.0 or higher GPA.[1] This makes it all but impossible to distinguish between the first and second deciles with anything resembling reliability. Furthermore, any high-ranking student earning a below-average grade, for any reason, is brutally punished, dropping to the bottom of the second decile, if not farther. This is a result of the extremely high average grade; if the average grade were lower, there would be a margin for error for top students.

Unfortunately, the traditional GPA exacerbates its own problems by encouraging the grade inflation which makes it so useless. Since it does not correct for course difficulty, students will seek out easily graded courses. Faced with the prospect of declining enrollment and poor student evaluations, professors who grade strictly will feel pressure to relax their grading standards. Instructors who grade easily will be rewarded with high enrollment and excellent evaluations, potentially leading to promotion. This creates a strong push towards grade inflation, since the traditional GPA punishes both the student taking a difficult course and the professor teaching it.

[1] Repeated trials of the process described in Section E yield this result.

Any system intended to replace the traditional GPA should address this problem, so that grade inflation will be arrested and hopefully reversed.

Another potential concern is that the traditional GPA encourages specialization by students. Since students tend to perform better in courses related to their majors, the GPA will reward students who take as few courses outside their "comfort zone" as possible, and punish students who attempt to expand their horizons. We note, however, that individual colleges may or may not regard this as a problem; the relative values of specialization and well-roundedness are open to debate.

4.2 Three Possible Solutions

In a ranking of students based on their performances relative to each other, the following considerations also come into play.

- It is not possible to compare students just to others in their class. Students often take courses in which all other students belong to another class.
- We have to compute rankings separately each semester, because the pool of students changes due to graduation and matriculation.
- It is not possible to take into account independent studies, because there is nobody to compare to.
- It is not possible to take into account pass/fail courses because they do not assign relative grades.

We recognize three potential solutions to this problem. The following sections will describe them in more detail.

The first is the Standard Normalized GPA. It assumes that grades within a course should be normally distributed, and maps all grades within each course to a bell curve. Each student is given a revised GPA based on his average position on the bell curves for each of his courses.

The second is the Iterated Normalized GPA. It attempts to correct for the varying difficulties of courses. In theory, every grade given to a student should be approximately equal to his GPA, so the average grade given in a course should be about equal to the average GPA of students in that course. This scheme repeatedly adjusts all the grade points in each course until the average grade in every course equals the average GPA of the enrolled students.

The third is the Least-Squares. It assumes that, other things being equal, the difference between two students' grades will be equal to the difference in their ability scores. It attempts to find these ability scores by solving the system of equations generated by each course (for example, if student X gets an A but student Y gets a B, then $X - Y = 1$.) Since, in any non-trivial population, this system has no solution, methods of least-squares approximation are used to approximate these values. The students are then ranked according to ability score.

5 The Standard Normalized GPA

5.1 How It Works

The Standard Normalized GPA is perhaps the simplest method, and the method most in keeping with the Dean's suggestion. For the grades in each course, the mean and the standard deviation are computed. We then determine how many standard deviations above or below the mean each student's grade is. This standard deviation measurement then becomes the student's "grade" for the class. Each student's standard deviation "grades" are then averaged for a overall "GPA" measurement. The students are then ranked by this average GPA performance. This is a quantified version of the Dean's suggestion to rank each student as average, below average, or above average in each class, and then combine the information for a ranking.

5.2 Strengths and Weaknesses

This method has several advantages. It is fairly easy to calculate, not too much more difficult than the standard GPA measurement. Each course can be considered independently, which can be a big advantage when final grades are calculated. Instead of waiting for all results to come in, the registrar may calculate the students' ability score (which in this case is the number of standard deviations above the mean) for that course immediately. This can save time when sending grades out at the end of the semester. The standard deviations do correct for differing course averages, e.g., making a B+ when the course average is a C+ look better than an A- when the course average is an A. At the same time, it continues to rank students in the order in which they scored in each course. Student X is thus always ranked above Student Y if X and Y take similar courses, and X has better grades.

The Standard Normalized GPA suffers from many of the same problems as the traditional GPA. It does not reward students who have a more well-rounded course load. Instead, students are punished severly if they perform at less than the course average. This is likely to punish a student who takes a course outside his major, since he is likely to score worse than those students who are majoring in the course's subject. This is not nearly as bad as the treatment of difficult courses, however. The standard GPA makes no distinction between easy and difficult courses, and thus encourages easy courses, as previously discussed. The Standard Normalized GPA attempts to correct this, but ends up claiming that a low average grade is equivalent to a difficult course. This is not always true, and has some interesting quirks. Higher level courses may be populated by only students who excel both in the subject of the course and in general, and so only high grades are given. But if all grades are high, this method treats the course as easy. This method boosts one student's grade if the other students in one's course have lower scores. Additionally, ability scores may be signifcantly raised by adding poor students to the course. This method has no feedback mechanism, as it does not compensate for the skill of the students when deciding

the difficulty of a course. A good student who takes courses with other good students will look worse than a slightly less able student who takes courses among significantly less able students. It is clear that the difficulty of a course is decided not only by the grades of its students, but also by the aptitudes of the students in the course.

5.3 Consequences

The Standard Normalized GPA also assumes that all teachers assign grades based on a bell curve. Its ability scores are calculated with this in mind. While a bell curve is the most likely distribution, not all teachers grade on a curve, and some classes, as mentioned before, may require grades to fit some other distribution in order to be fair; e.g., if all the students are extraordinarily talented. However, the college may regard it as positive to force the professors to assign grades on a bell curve. In this case, this method would be useful, since its ability scores are more meaningful when the grades for a course are normally distributed. The closer the grades for a course are to a curve, the more closely the ability scores match the grades. Grading on a curve also fosters cutthroat competition among students, since any student's ability score may be significantly raised by lowering the ability scores of other students.

6 The Iterated Normalized GPA

6.1 How it Works

Rather than directly comparing students, this algorithm compares courses. Suppose we have a course which is unusually difficult. Then students should be given lower grades in that course relative to their others. The average grade from that course should therefore be lower than the average GPA of all students enrolled in it.

We should therefore be able to correct for courses that are unusually difficult by adding a small amount to the point value of every grade given in that course. Likewise, we can correct for easy courses by subtracting a small amount.

Of course, once we have corrected everyone's grades, their new GPA's will be different, and most likely some courses will need further correction. The Iterated Normalized GPA method of ranking makes ten corrections to all grades, then sorts students in order of corrected GPA. According to numerical experiments explained in Section A, ten iterations are sufficient to bring the difference between the average GPA and the average grade down to zero.

6.2 Strengths and Weaknesses

This algorithm is fairly quick to compute, taking only a couple of minutes for 1000 students, 200 courses, and a course load of 6. The computation is straightforward to explain, and easily understood by non-experts.

However, all the grades from all courses must be known to run the computation. The corrected grades cannot be computed independently by students. There is also no guarantee that the corrected GPA's will be comparable across semesters. To compute overall class rank at graduation, it will be necessary to average the ranks across semesters, rather than corrected GPA's.

6.3 Consequences

The Iterated Normalized GPA systematically corrects for teacher bias in giving grades, thus eliminating the tendency of students to select easy courses, and therefore makes progress toward reversing grade inflation. The total correction made for each course may be used as an indicator of the course's grade bias.

This algorithm tends to "punish" students in courses where the grades are unusually high. It should be noted that if students score high in a course relative to their other grades, it could be because either the course was easy, or the students put forth an extra effort. If the course was easy, then the punishment is due. If the difference was due to extra effort, then such effort is not typical of the students in question, and the punishment is arguably due.

Although the correction can be applied to very small classes and independent studies, strange things are likely to happen. If a student in an independent study gets a grade above his GPA, he gets punished by the correction, and if he gets a lower grade, he gets rewarded, which is clearly undesirable. The first round of corrections will replace the F with the student's GPA, but since his grades were worse to begin with, the average GPA of his other courses will be smaller than if he'd made a better grade in his independent study, so the difficulty correction for those is smaller. Using the sample data set presented in Section C as a basis, we experimented with independent studies and determined that they had minimal impact on the rank order.

Therefore, to avoid such strange results, independent studies should be ignored when running the computation.

7 The Least-Squares Algorithm

7.1 How it Works

The Least-Squares method assumes that the difference between two students' abilities will be reflected in the difference between their grades. Hence, if student X and student Y take the same course, and get an A and B, respectively, $X - Y = 1$.

We further assume that students majoring in hard science fields will perform better in hard science courses than in humanities courses, and vice versa. We also assume that this will be reflected by a drop in their grade, and that this drop is of approximately the same magnitude for all students; we call it H_H. Hence, if, in the example above, students X and Y are taking a mathematics

course, but student X is majoring in physics and student Y is majoring in literature, we have $X - (Y + H_H) = 1$.

A course with N students will generate $\frac{N^2+N}{2}$ such linear equations; the abilities of each student can be expected to be the solution to the set of all such equations from every course offered during the semester.

In practice, these equations never have a solution. A student body of 6000, for example, will generate roughly 250,000 equations;[2] the chance that they will be consistent is essentially zero. Hence, methods of least-squares approximation must be employed. The system is converted into the matrix equation $Ax = b$, where A is the matrix of the coefficients of the left hand side of each equation, x is the vector of the abilities of each student and the constant HH, and b is the right hand side of each equation. This matrix equation has no solution, but multiplication through by the transpose of A yields the equation $A^T Ax = A^T b$, where x is now the least-squares approximation to the original system. This matrix equation, instead of having no solution, has a one-dimensional solution set; with nullspace equal to scalar multiples of $(111111...11110)^T$, where the 1's correspond to the student's abilities and the 0 to the constant H_H. Thus, one student's ability score may be assigned arbitrarily, and the rest will then be well-determined. This arbitrary assignment will in no way affect the ordering of any two students' ability scores, or the magnitude of the difference between two students. After these scores are determined, the difference between a 2.0 and the median score is added to every student's score, so that the scores will be easily interpretable in terms of the traditional GPA. These scores can be averaged over all 8 semesters to produce a ranking at graduation.

7.2 Consequences

An immediate consequence of changing to this ranking will be that, so long as the average grade remains an A-, all ability scores will be tightly packed into a range between about 1.0 and about 3.0; no student will appear to carry an A average. This will likely result in professors widening their grading scales, in order to reward their best students, thus reducing grade inflation to something more reasonable.

7.3 Strengths

This method has two major strengths; the first is that it corrects for the difficulty of every student's course load, the second that it can reward students for carrying a well-rounded course load.

This second strength is extremely flexible, and deserves further enumeration. If a school wishes not to account for well-roundedness, the factor H_H may be omitted from the algorithm above, with no consequence except that the ability scores will no longer consider the balance or specialization in each

[2] Assuming a typical course has 20 students and there are 1200 courses, we have $21 \times 10 \times 1200 = 250,000$ pairs of grades.

student's course load. If a school wishes to emphasize several different areas of specialization rather than just two, say, Arts and Literature, Social Sciences, Natural Sciences, and Hard Sciences, it could do so by replacing H_H with several constants representing the difficulty of the transitions between each pair, in this case, A_S, A_N, A_H, S_N, S_H, and N_H. A school may also dictate that certain emphasized courses, e.g. a freshman writing course, not benefit students majoring in some departments over others, it may categorize that course as belonging to every area of specialization, or to none; similarly, if a school wishes to dictate that certain de-emphasized courses, e.g. Physical Education, not reward students with a Well-Roundedness correction, it may also dictate that they be categorized in every area of specialization or in none. Lastly, other corrections may be made for students with special circumstances, for example, if a student double-majors in two different areas of specialization, as each Well-Roundedness correction might be replaced by the average of the two corrections from each of his major areas.

7.4 Weaknesses

The most glaring weakness of this method is that it involves huge amounts of computations, and may severely tax the computing resources of larger universities. For example, a school with a student body of about 6000 will need approximately 200 megabytes of RAM and 14 hours at 200 megahertz to perform the necessary calculations. For any significantly larger school, the operation quickly becomes infeasible with current technology. See Section B for the details of these calculations.

8 A Small Test Population

We postulate a very small school, which has a student body of 18 (students A - R), and offers only the following classes: Math, Physics, Computer Science, Physical Education, Health, English, French, History, Philosophy, Psychology, Music History.

Math, Physics, and English are generally believed to be prohibitively difficult courses, while Physical Education, Health, and Music History are generally considered to be very easy.

Students' transcripts are listed in C. Looking just at these transcripts, without analyzing them numerically, we find that the following are self-evident:

- Student A should be ranked ahead of Student B, Student C ahead of Student D, Student E ahead of F, and so on, because they carry better grades in courseloads of similar difficulty.
- Students O and D should be ranked ahead of Student J because they have slightly better grades in more difficult courseloads.
- Student E should be ranked ahead of Student D because he has better grades in a more difficult courseload.

Any valid ranking system must satisfy these conditions.

We also recognize the following relationships as preferable but not absolutely necessary:

- Student O should be ranked ahead of Students Q and R, and student P should be ranked ahead of student R, because they have almost as good grades and much more difficult schedules.
- Student M should be ranked ahead of Student Q, and Student N ahead of student R, because they have similar grades in a more difficult schedule.
- Student K should be ranked ahead of Students M, N, Q, and R because he has similar grades in a much more difficult schedule.
- Students C, G, and K should be ranked near each other because they have similar grades in similar schedules.
- Student P should be ranked ahead of Student J, because he has similar grades against a significantly more difficult schedule, and has higher grades in the two classes that they share.

If we postulate that the well-roundedness, or lack thereof, of a student's schedule should affect his rank, we also find the following relationships:

- Student E should be ranked ahead of Students C and D, because he has almost as good grades in a more difficult, much more well-rounded schedule.
- Student I should be ranked ahead of Students K and M because he has similar grades against a more well-rounded schedule.

We now have some basis to analyze the effectiveness of our ranking systems.

The rankings of this sample population are given in Section D.

First, we consider the traditional GPA. It does satisfy the relationships we had required, and it does rank students C, G, and K near each other, and it ranks student K higher than students N and R. However, it does not rank student M above student Q , nor student N above student R, nor student K above students M or Q, nor students O and P above students Q and R. The traditional GPA passes only five of the thirteen tests we had set for it.

If we also stipulate that well-roundedness should be a factor, the traditional GPA fails 3 of the four additional tests we had set, only ranking student E higher than student D.

Next, we consider the Standard Normalized GPA. It also satisfies all of the required relationships, and satisfies student C close to student G, student N ranked ahead of student R, and student K ranked ahead of students M, N, Q, and R. It passes these six tests and fails the other seven.

If we also stipulate that well-roundedness should be a factor, the Standard Normalized GPA passes only two of the four additional tests we had set, ranking student E ahead of student D and student I ahead of Student M.

Third we consider the Iterated Normalized GPA. It satisfies all the relationships we had required, and also satisfies eight of the thirteen tests we had set, failing only to rank student P ahead of students J and R, student O ahead of student Q, and students C and G near student K.

If we also stipulate that well-roundedness should be a factor, the Iterated Normalized GPA passes two of the four additional tests we had set, ranking student E ahead of student D and student I ahead of student K.

Lastly, we consider the Least-Squares. It also satisfies all of the required relationships, and passes nine of the thirteen tests we had set, failing only to rank students N and P ahead of student R and to rank students C and G near student K.

If we further stipulate that well-roundedness should be a factor, Least-Squares satisfies all four of the additional tests.

9 Test Population Redux (no +/-)

To answer the second question, we take the initial test population, and drop all pluses and minuses from the rankings. This leaves a similar set of rankings, but with far less information. Yet we are still able to determine some basic intuitive relationship from these transcripts:

- A should be ranked above B, C above D, and G above H since they have better grades in similar courses.

Any valid ranking system must satisfy these conditions.

We also recognise the following relationships as preferable but not absolutely necessary:

- O should be ranked ahead of P because he has slightly better grades in the same courseload.
- E should be ranked ahead of F because he has the same grades in a more difficult courseload.
- O should be ranked ahead of Q and R because he has almost equivalent grades in a much more difficult courseload.
- C should be ranked ahead of I and G because he has the same grades in a more difficult courseload.
- I should be ranked ahead of K and L because he has the same grades in a more difficult courseload.
- K and L should be ranked ahead of M and N because they have the same grades in a more difficult courseload.
- M and N should be ranked ahead of Q and R because they have the same grades in a more difficult courseload.

If we postulate that the well-roundedness, or lack thereof, of a student's schedule should affect his rank, we also find that C, E, G, and I should be ranked near each other because:

- E has slightly worse grades in a more difficult, better-rounded courseload.
- C has the same grades as G and I in a slightly more difficult, slightly less well-rounded courseload.

The rankings of this sample population are given in Section D.

First we consider the traditional GPA. It does satisfy the relationships we had required, and it does rank O ahead of P. However, it fails every one of the other tests we had set for it, passing only one of twelve.

If we also stipulate that well-roundedness should be a factor, the traditional GPA passes three of the six tests we had set, failing to rank E close to I, C, or G.

Next we consider the Standard Normalized GPA. It also satisfies all of the required relationships, and ranks O above P, C above I, I higher than K, both K and L ahead of M, and both M and N ahead of Q and R. It passes these six tests, but fails in the six others, not ranking E above F, I ahead of L, both K and L higher than N, C above G, O ahead of Q, and O higher than R.

If we stipulate that well-roundedness should also be a factor, the Standard Normalized GPA passes only one of the six additional tests, ranking only E near C.

Third we consider the Iterated Normalized GPA. It satisfies all the relationships we had required, and also passes nine of the twelve other recommended tests, failing only to rank C ahead of G, C ahead of I, and I ahead of L.

If we also stipulate the well-roundedness should be a factor, the Iterated Normalized GPA satisfies three of the six additional test, failing to rank E near I, C, or G.

Lastly we consider the Least-Squares method. It also satisfies all of the required relationships, and additionally satisfies nine of the thirteen recommended relationships, failing to rank C above G or I, and to rank both M and N ahead of R.

If we further stipulate the well-roundness tests, then Least-Squares satisfies four of the six tests, failing to rank G near C and E (One can see that *near* is a fuzzy term, not transitive, since G is near I, and I is near E and C, but G is not near E or C).

10 Stability

10.1 How Well do the Models Agree?

We have four ways of ordering students: plain GPA, Standard Normalized GPA, Iterated Normalized GPA, and Least-Squares. Since all four are more or less reasonable, they should agree fairly well with each other.

One way to test agreement is to plot each student's rank under one method with his rank under the others. If the plot is scattered randomly, then the rankings do not agree about anything. If the plot is a straight line, then the rankings agree completely.

To get an idea for how each model works, a large population of students and courses was created using a numerical simulation, with 1000 students, 200 courses, and 6 courses per student. The details of the simulator are explained in Section E. All the algorithms were applied to them except Least-Squares, which was too difficult to implement in the available time.

10.1.1 With Plus and Minus Grades

See Figures 1, 2, and 3 for graphs of the agreement, using simulated students and courses, and allowing plus and minus grades. The comparisons to plain GPA rankings are rather scattered, especially toward the lower left corner where the highest rankings are. The plain GPA rankings do not appear to agree particularly well with either the Iterated Normalized or the Standard Normalized rankings. There are lots of scattered points. This is due mostly to the fact that there are lots of ties in plain GPA rankings, especially near the top of the class, and tied students are ordered more or less at random. Very few ties are present in any of the other methods. The Iterated Normalized and Standard Normalized rankings are in better agreement, with fewer outlying points.

The two non-traditional methods do agree mostly, but not completely, on the first decile. They agree on 89 of the 100 students in the top decile.

A single run of the simulation is analyzed here, but these results are typical of other runs.

10.1.2 Without Plus and Minus Grades

See Figures 4, 5, and 6 for graphs of the agreement, using simulated students and courses, and disallowing plus and minus grades.

A great deal of information is lost without the use of plus and minus grades. In particular there are many more ties in the plain GPA based ranking which show up as large squares of scattered points. The large square at the bottom left shows the massive tie among people with 4.0 averages. Again, the plain GPA is not in good agreement with the non-traditional methods due to these ties. Both new models agree with each other on 79 of the 100 students in the top decile. Apparently, the loss of information is responsible for the greater lack of agreement.

10.2 How Much Does Changing One Grade Affect the Outcome?

If one grade of one student is changed, his rank can be expected to change as well. For plain GPA rankings, changing one student's grades can only move that student from one place to another. In the non-traditional rankings, each

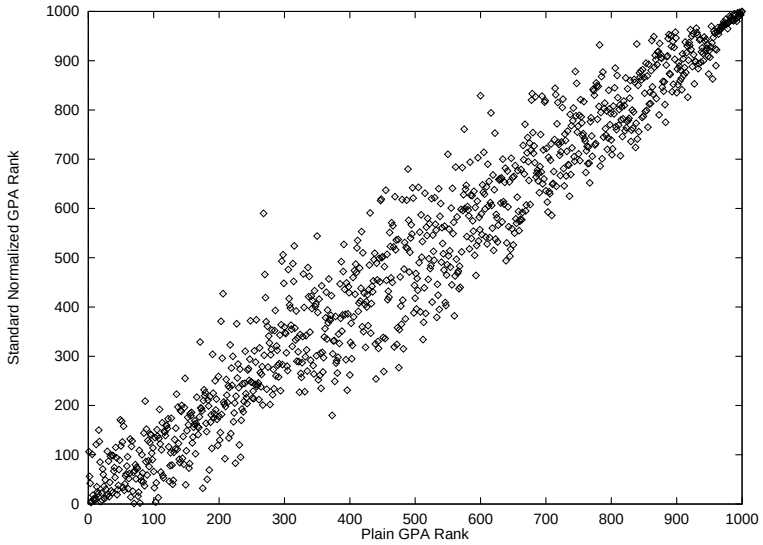

Figure 1: Plain GPA rankings plotted against Standard Normalized GPA rankings, using simulated students.

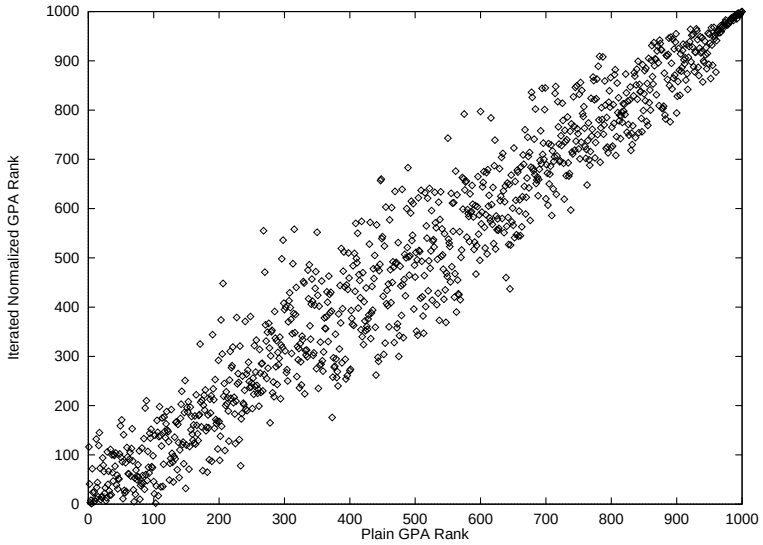

Figure 2: Plain GPA rankings plotted against Iterated Normalized GPA rankings, using simulated students.

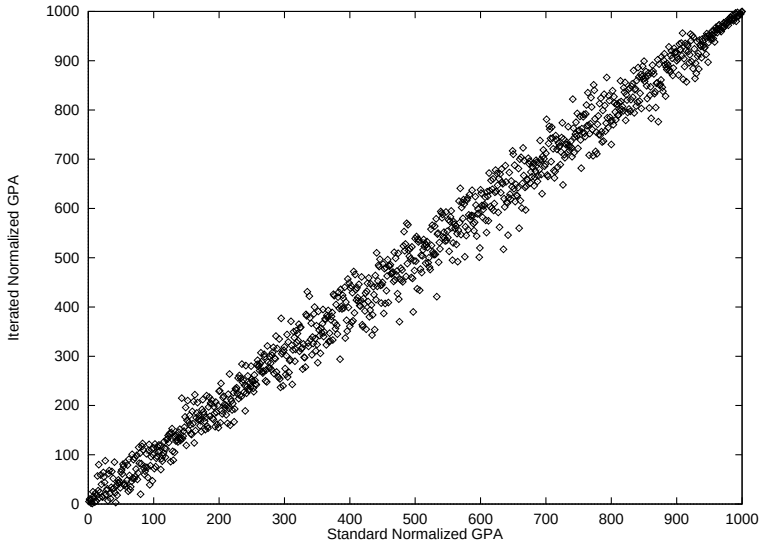

Figure 3: Standard Normalized GPA rankings plotted against Iterated Normalized GPA rankings, using simulated students.

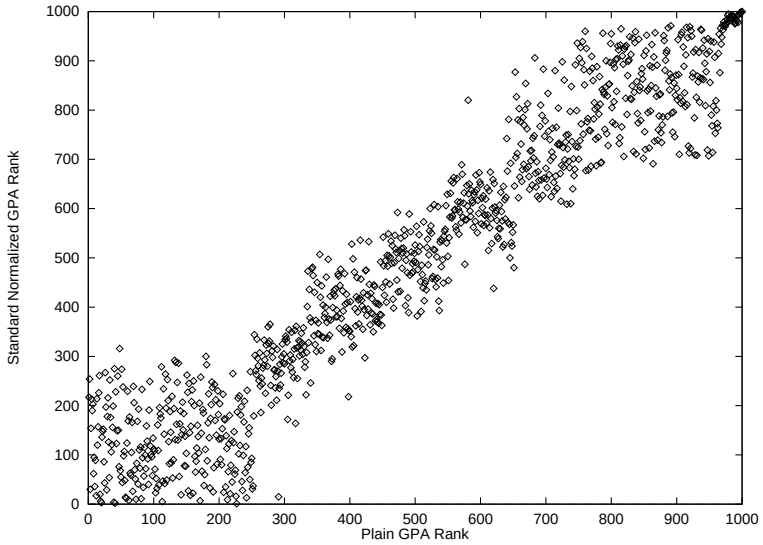

Figure 4: Plain GPA rankings plotted against Standard Normalized GPA rankings, using simulated students, and no plus or minus grades.

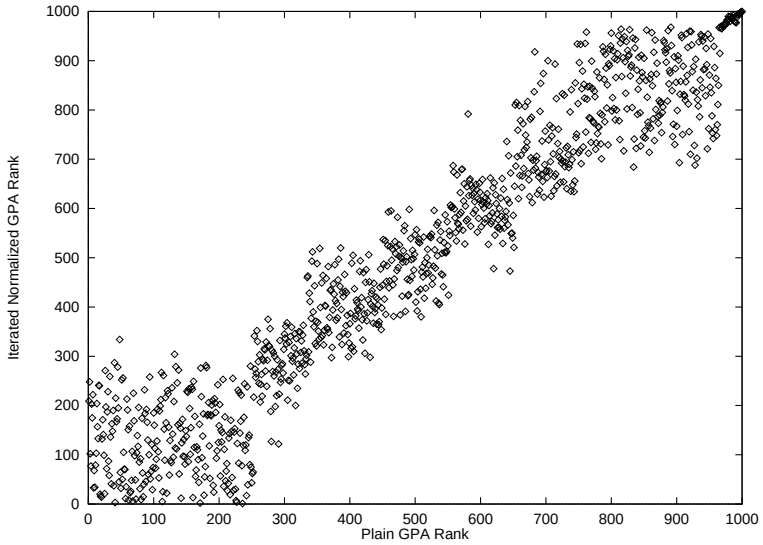

Figure 5: Plain GPA rankings plotted against Iterated Normalized GPA rankings, using simulated students, and no plus or minus grades.

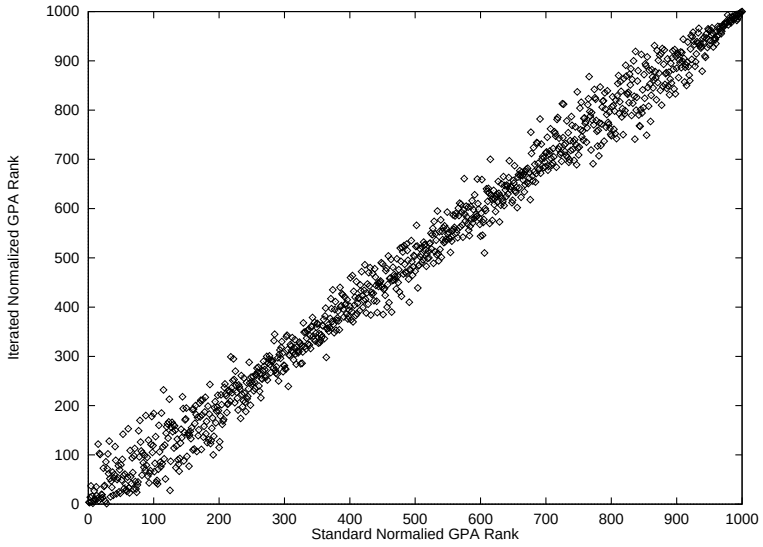

Figure 6: Standard Normalized GPA rankings plotted against Iterated Normalized GPA rankings, using simulated students, and no plus or minus grades.

student's position is determined relative to the other students, and one changed grade might trigger a chain of rank changes. To test the sensitivity, the sample population was modified slightly, and student Q's grade of A+ in Music History was changed to a C-, a very drastic change.

The change was tested including plus and minus grades and using only whole letter grades. When only letter grades are considered, the change is to a C. Using the GPA ranking and plus and minus grades, Q dropped from 1st to 14th, and there were no other changes. With only whole letter grades, Q dropped from 2nd to 16th, and there were no other changes except to make room for Q.

For the Standard Normalized GPA with plus and minus grades, Q dropped from 3rd to 12th. L, N, and J improved several places, apparently because they also took Music History and benefitted from the drop in mean grade. R improved one spot, apparently for the same reason. K dropped by one. Without plus and minus grades, Q dropped 8 places, and J, L, and N improved one position each. Student I dropped three places, perhaps because of how N and K benefitted from Music History.

The Iterated Normalized GPA including plus and minus grades was rather stable. Q dropped 9 places, and J and L improved a couple of places each, benefitting from the apparent increase in the difficulty of Music History. G dropped two places, possibly because he scored lower in Health. When only whole letter grades are used, Q drops from 13th to 16th. J and K improve a couple of places, benefitting from the increased difficulty of Music History, while G drops again, three places this time. O and F switched places for no obvious reason.

Using Least-Squares and plus and minus grades, Q drops 9 places. Other members of the Music History course J and K improve a bit, and L improves a lot. With letter grades only, Q drops from 15th to 16th, and J, K, and L improve. For no obvious reason, E and C switched places. O dropped by two because of improvements by K and L.

Thus, it would seem that plain GPA ranking is the most stable, since at most one person changes position and the rest move up or down at most one to compensate. The next most stable seems to be Least-Squares, followed by Iterated Normalized, and finally Standard Normalized. In each scheme, the classmates of the person whose grade changed are most likely to change rank. There were a few chain-reaction re-orderings, which are harder to explain. Also, having plus and minus grades appears to improve stability in general.

10.3 How Does Course Size Affect the Outcome?

Another simulation was run with 1000 students, 500 courses, and a course load of 6 per student. Courses came out smaller, and the correlation between the Standard Normalized ranking and the Iterated Normalized ranking was weaker. This is probably due to the fact that standard deviations computed on smaller data sets tend to be less reliable, as are average grades and average GPA's.

11 Strengths and Weaknesses of Each Model and Recommendations

Our recommendation hinges on the answers to two questions, which will vary college by college.

If the college wishes to promote well-roundedness over specialization (which we suggest), and has a fairly small population (less than about 6000 students), we recommend the Least-Squares method.

Otherwise, we recommend the Iterated Normalized GPA method.

We feel that the Least-Squares method is superior to the other two because:

- It does not punish students for attempting to expand their horizons.
- It produces results more consistent with intuitive observation than do the Iterated or Standard Normalized GPA.
- It is more flexible than either the Iterated or Standard Normalized GPA.
- It is clear and easily understood.

The Iterated Normalized GPA method has a few definite advantages as well:

- It is significantly faster than the Least-Squares method.
- If the well-roundedness of students is not a consideration, it produces results which are roughly as consistent with intuitive observation as the Least-Squares method.

We feel that the Standard Normalized GPA method is decidedly inferior, and should not be recommended, because:

- It makes no attempt to correct for schedule difficulty or well-roundedness.
- It assumes that all courses have the same range of ability among their students.
- It produces results which are no more consistent with intuitive observation than those produced by the traditional GPA.
- It forces all grades to be interpreted as part of a bell curve, which is inappropriate for small and highly specialized courses.

12 Further Recommendations

12.1 Transition from GPA Ranking

The three methods given here all rank an entire student body for one semester of courses. Thus, to rank students just within a single class, we must either average their ability scores (revised GPA's) or their ranks within their class over each semester.

The new system could be phased in at any time if grades for enough preceding years are kept on record. The new ranking algorithm could be applied to students who have graduated to determining rankings for the next class. However, we recommend careful testing on several past years of data as well as current grades.

The administration should be prepared for a great deal of student and faculty opposition because it is a new, untested system. The Standard Normalized and Iterated Normalized schemes are most likely to cause opposition because they directly alter the point values of grades during computation. The Least-Squares method simply re-interprets them and is less likely to make teachers feel that their authority has been violated.

12.2 Transfer Students

ABC College will have to come up with its own policy concerning the ranking of transfer students. One option is to translate transferred grades to an equivalent grade in a particular course at ABC. That allows the ranking algorithm to run on the maximum amount of information. However, someone will have to compare all other colleges to ABC very carefully to create the official translation policy. Another possibility is to ignore transferred grades when computing the rankings. That avoids the problem of estimating how grades at other schools compare to ABC's, but at the expense of throwing out a lot of information.

12.3 Importance of Plus and Minus Grades

It seems that plus and minus grades are extremely helpful in determining class rank, especially since grades are so heavily inflated. Without them, ABC has to rank its students primarily on the basis of just two grades, A and B, and a considerable fraction of the students have exactly the same grades. With pluses and minuses, there are six different grades, A+, A, A-, B+, B, and B-, which come into play, thus differentiating students more precisely. All four ranking systems appear to work better when plus and minus grades are used. ABC should encourage its teachers to use them with care.

Appendices

A A Word About Convergence

The Iterated Normalized GPA ranking scheme makes corrections to grades based on its estimate of how hard or easy a course is.

According to numerical experiments, the corrections converge very quickly to zero. Here, for example, are the corrections make for two courses generated by the simulation described in Section E. The parameters h, s, and e are described in that section as well. The important numbers are the last three columns. They demonstrate that the difference between the average corrected grade given in the course and the average corrected GPA of students in the course converges to zero.

The first one is an extreme case of a difficult course:

```
Course15(h=0.222,s=0.778,e=-0.278): avgGrade=3.367 avgGPA=3.671 diff=0.304
Course15(h=0.222,s=0.778,e=-0.278): avgGrade=3.671 avgGPA=3.731 diff=0.060
Course15(h=0.222,s=0.778,e=-0.278): avgGrade=3.731 avgGPA=3.745 diff=0.014
Course15(h=0.222,s=0.778,e=-0.278): avgGrade=3.745 avgGPA=3.749 diff=0.004
Course15(h=0.222,s=0.778,e=-0.278): avgGrade=3.749 avgGPA=3.750 diff=0.001
Course15(h=0.222,s=0.778,e=-0.278): avgGrade=3.750 avgGPA=3.750 diff=0.000
Course15(h=0.222,s=0.778,e=-0.278): avgGrade=3.750 avgGPA=3.750 diff=0.000
Course15(h=0.222,s=0.778,e=-0.278): avgGrade=3.750 avgGPA=3.750 diff=0.000
Course15(h=0.222,s=0.778,e=-0.278): avgGrade=3.750 avgGPA=3.750 diff=0.000
Course15(h=0.222,s=0.778,e=-0.278): avgGrade=3.750 avgGPA=3.750 diff=0.000
```

The second one is an extreme case of an easy course:

```
Course11(h=0.904,s=0.096,e=0.404): avgGrade=3.889 avgGPA=3.480 diff=-0.408
Course11(h=0.904,s=0.096,e=0.404): avgGrade=3.480 avgGPA=3.424 diff=-0.056
Course11(h=0.904,s=0.096,e=0.404): avgGrade=3.424 avgGPA=3.416 diff=-0.008
Course11(h=0.904,s=0.096,e=0.404): avgGrade=3.416 avgGPA=3.414 diff=-0.002
Course11(h=0.904,s=0.096,e=0.404): avgGrade=3.414 avgGPA=3.414 diff=-0.000
Course11(h=0.904,s=0.096,e=0.404): avgGrade=3.414 avgGPA=3.414 diff=-0.000
Course11(h=0.904,s=0.096,e=0.404): avgGrade=3.414 avgGPA=3.414 diff=-0.000
Course11(h=0.904,s=0.096,e=0.404): avgGrade=3.414 avgGPA=3.414 diff=-0.000
Course11(h=0.904,s=0.096,e=0.404): avgGrade=3.414 avgGPA=3.414 diff=-0.000
Course11(h=0.904,s=0.096,e=0.404): avgGrade=3.414 avgGPA=3.414 diff=-0.000
```

Ten iterations are apparently sufficient to bring the difference down to zero, within three decimal places.

B Time and Resources Required for the Least-Squares Algorithm

Assuming 6000 students who each take 4 courses in a semester, and 1200 courses, each course averages $21 \times 10 = 210$ pairs of grades, so there are about 250,000

pairs of grades total. Therefore, the size of the A matrix is 250,000 rows, one for each pair of grades, and 6000 columns, one for each student. The size of b is one column and 250,000 rows.

Without some clever programming, there is no way to manipulate A within the memory of a conventional computer. However, A may be stored as a sparse matrix, where the non-zero entries are stored in linked lists. Each non-zero number takes up 12 bytes (4 for the floating point number, 4 for a pointer to the next non-zero element, and 4 for an index specifying the column number). Each row of A has at most 4 non-zero entries, for a total of about 48 bytes per row, or 12 megabytes total. A^T can be faked by referencing elements of A with the row and column indices switched, thus requiring minimal extra memory. The vector b cannot be "faked," and must take up 4 bytes per row, times 250,000 rows, for another megabyte of memory.

The real memory problem is that $A^T A$ cannot be faked, and must take up 4 bytes per entry times 36,000,000 entries, or 144 megabytes. That is barely within the range of medium-sized current computers. Computing $A^T A$ takes on the order of $250,000 \times 6000^2 = 9 \times 10^{12}$ multiplications. Computing $A^T b$ takes only about 1.5×10^9 multiplications. Solving the linear system $A^T Ax = A^T b$ takes on the order of $6000^3 = 2.2 \times 10^{11}$ operations.

Thus, the time to solve the system is about 10^{13} operations, which takes 50,000 seconds or 14 hours to run on a 200 megahertz computer.

If the population of the school is much larger, the storage requirement for the computation (on the order of the square of the number of students) becomes unreasonably large.

C The Schedules and Grades of Students in the Small Sample Population

Stars indicate the major of each student. "CPS" means Computer Science, and "PhysEd" means Physical Education.

Student	Courses
A	PhysEd 4.3, Health 4.0, *History 3.0, Math 2.3
B	PhysEd 4.3, Health 3.3, *Psychology 2.0, CPS 2.0
C	Math 4.0, *Physics 4.3, CPS 4.0, Philosophy 3.7
D	*Math 4.0, Physics 3.7, CPS 4.0, French 3.0
E	*Math 4.3, Physics 4.0, English 3.3, History 3.7
F	Physics 3.7, *CPS 4.0, French 3.7, History 3.0
G	Math 4.0, *CPS 4.3, Health 4.0, English 3.7
H	CPS 3.0, *Physics 4.0, PhysEd 4.0, Psychology 3.0
I	English 4.0, French 4.3, CPS 3.7, *Philosophy 4.3
J	English 3.7, *French 4.0, Music History 4.0, Math 2.7
K	*English 4.3, Philosophy 4.0, Psychology 4.0, Music History 4.3
L	English 3.7, *History 4.0, Psychology 4.0, Music History 4.0
M	Music History 4.3, Psychology 4.3, *French 4.3, PhysEd 4.0
N	*Music History 4.0, Psychology 4.0, French 4.0, Health 4.0
O	Physics 4.0, English 3.3, *Math 4.0, Philosophy 4.0
P	Physics 3.0, *English 3.7, Math 3.3, Philosophy 4.0
Q	PhysEd 4.0, Health 4.3, Music History 4.3, *Psychology 4.3
R	PhysEd 4.0, Health 4.0, Music History 4.0, *CPS 4.0

This is the same table, only without allowing + and - grades.

Student	Courses
A	PhysEd 4.0, Health 4.0, *History 3.0, Math 2.0
B	PhysEd 4.0, Health 3.0, *Psychology 2.0, CPS 2.0
C	Math 4.0, *Physics 4.0, CPS 4.0, Philosophy 4.0
D	*Math 4.0, Physics 4.0, CPS 4.0, French 3.0
E	*Math 4.0, Physics 4.0, English 3.0, History 4.0
F	Physics 4.0, *CPS 4.0, French 4.0, History 3.0
G	Math 4.0, *CPS 4.0, Health 4.0, English 4.0
H	CPS 3.0, *Physics 4.0, PhysEd 4.0, Psychology 3.0
I	English 4.0, French 4.0, CPS 4.0, *Philosophy 4.0
J	English 4.0, *French 4.0, Music History 4.0, Math 3.0
K	*English 4.0, Philosophy 4.0, Psychology 4.0, Music History 4.0
L	English 4.0, *History 4.0, Psychology 4.0, Music History 4.0
M	Music History 4.0, Psychology 4.0, *French 4.0, PhysEd 4.0
N	*Music History 4.0, Psychology 4.0, French 4.0, Health 4.0
O	Physics 4.0, English 3.0, *Math 4.0, Philosophy 4.0
P	Physics 3.0, *English 4.0, Math 3.0, Philosophy 4.0
Q	PhysEd 4.0, Health 4.0, Music History 4.0, *Psychology 4.0
R	PhysEd 4.0, Health 4.0, Music History 4.0, *CPS 4.0

D The Rankings of the Sample Population

D.1 Including Plus and Minus Grades

D.1.1 In order of plain GPA, using plus and minus grades

```
Student Q (gpa=4.250)
Student M (gpa=4.250)
Student K (gpa=4.167)
Student I (gpa=4.083)
Student R (gpa=4.000)
Student N (gpa=4.000)
Student C (gpa=4.000)
Student G (gpa=4.000)
Student L (gpa=3.917)
Student E (gpa=3.833)
Student O (gpa=3.833)
Student D (gpa=3.667)
Student J (gpa=3.583)
Student F (gpa=3.583)
Student H (gpa=3.500)
Student P (gpa=3.500)
Student A (gpa=3.417)
Student B (gpa=2.917)
```

D.1.2 In order of Standard Normalized GPA, using plus and minus grades

```
Student K (gpa=0.837)
Student I (gpa=0.809)
Student Q (gpa=0.596)
Student M (gpa=0.521)
Student G (gpa=0.385)
Student C (gpa=0.221)
Student E (gpa=0.211)
Student L (gpa=0.157)
Student N (gpa=-0.006)
Student O (gpa=-0.029)
Student R (gpa=-0.197)
Student D (gpa=-0.262)
Student A (gpa=-0.272)
Student F (gpa=-0.281)
Student H (gpa=-0.447)
Student J (gpa=-0.491)
Student P (gpa=-0.594)
Student B (gpa=-1.159)
```

D.1.3 Iteratively Normalized GPA, using plus and minus grades

```
Student K (gpa=4.218)
Student I (gpa=4.169)
Student M (gpa=4.093)
Student C (gpa=4.076)
Student G (gpa=4.071)
Student L (gpa=4.058)
Student E (gpa=4.046)
Student Q (gpa=4.021)
Student O (gpa=3.955)
Student N (gpa=3.904)
Student R (gpa=3.764)
Student D (gpa=3.737)
Student F (gpa=3.689)
Student J (gpa=3.658)
Student P (gpa=3.622)
Student H (gpa=3.408)
Student A (gpa=3.356)
Student B (gpa=2.755)
```

D.1.4 Least-Squares, using plus and minus grades

```
Student E (gpa=2.323)
Student I (gpa=2.264)
Student G (gpa=2.240)
Student C (gpa=2.236)
Student O (gpa=2.180)
Student K (gpa=2.137)
Student M (gpa=2.047)
Student Q (gpa=2.029)
Student F (gpa=2.009)
Student R (gpa=1.991)
Student P (gpa=1.944)
Student D (gpa=1.933)
Student L (gpa=1.924)
Student N (gpa=1.872)
Student J (gpa=1.740)
Student H (gpa=1.602)
Student A (gpa=1.438)
Student B (gpa=0.890)
HH = -0.622
```

D.2 Using Letter Grades Only

D.2.1 In order of plain GPA, using whole letter grades only

```
Student R (gpa=4.000)
Student Q (gpa=4.000)
Student C (gpa=4.000)
Student N (gpa=4.000)
Student M (gpa=4.000)
Student G (gpa=4.000)
Student K (gpa=4.000)
Student I (gpa=4.000)
Student L (gpa=4.000)
Student O (gpa=3.750)
Student J (gpa=3.750)
Student F (gpa=3.750)
Student E (gpa=3.750)
Student D (gpa=3.750)
Student H (gpa=3.500)
Student P (gpa=3.500)
Student A (gpa=3.250)
Student B (gpa=2.750)
```

D.2.2 In order of Standard Normalized GPA, using whole letter grades

```
Student G (gpa=0.528)
Student L (gpa=0.488)
Student C (gpa=0.386)
Student I (gpa=0.363)
Student N (gpa=0.340)
Student K (gpa=0.271)
Student M (gpa=0.238)
Student Q (gpa=0.238)
Student R (gpa=0.228)
Student F (gpa=0.106)
Student J (gpa=0.072)
Student E (gpa=0.071)
Student D (gpa=-0.124)
Student O (gpa=-0.145)
Student H (gpa=-0.297)
Student P (gpa=-0.597)
Student A (gpa=-0.611)
Student B (gpa=-1.556)
```

D.2.3 Iteratively Normalized GPA, using whole letter grades

```
Student L (gpa=4.123)
Student G (gpa=4.074)
Student I (gpa=4.064)
Student C (gpa=4.051)
Student K (gpa=4.034)
Student N (gpa=3.963)
Student E (gpa=3.916)
Student M (gpa=3.890)
Student J (gpa=3.871)
Student D (gpa=3.843)
Student F (gpa=3.836)
Student O (gpa=3.827)
Student Q (gpa=3.805)
Student R (gpa=3.796)
Student P (gpa=3.577)
Student H (gpa=3.389)
Student A (gpa=3.184)
Student B (gpa=2.593)
```

D.2.4 Least-Squares, using whole letter grades

```
Student G (gpa=2.249)
Student I (gpa=2.192)
Student E (gpa=2.172)
Student C (gpa=2.170)
Student F (gpa=2.143)
Student O (gpa=2.042)
Student L (gpa=2.035)
Student D (gpa=2.032)
Student R (gpa=2.016)
Student K (gpa=1.984)
Student J (gpa=1.948)
Student N (gpa=1.900)
Student M (gpa=1.883)
Student P (gpa=1.871)
Student Q (gpa=1.846)
Student H (gpa=1.578)
Student A (gpa=1.264)
Student B (gpa=1.543)
HH = -0.556
```

E Test Data for the Ranking Schemes

The best way to test the ranking schemes is to apply them to an existing population of students. Since that information is not given, the alternative is to invent some reasonable data and see what the ranking schemes do with it.

E.1 Simulating Courses

We want to take the following things into consideration when creating courses:

- Students tend to pick more courses in areas they are comfortable in. In particular, they are required to select courses in their majors.
- Courses vary in subject matter. Some require a lot of math and scientific experience, while others focus more on human nature, history, and literature.
- Courses vary in difficulty. Here, we are not considering the difficulty of the material, but rather how difficult it is to get a good grade in the course. Students generally prefer courses where they expect to get better grades.
- Students are able to estimate their grade in a course fairly accurately.

Each simulated course c therefore has three attributes. The first two are fractions, c_s and c_h, which represent how much the course emphasizes the sciences and the humanities, respectively. Since these are fractions of the total effort required for a course, we have $c_s + c_h = 1$. In the simulation, c_s is determined by generating uniformly distributed random numbers between 0 and 1, and $c_h = 1 - c_s$.

The third attribute c_e is the "easiness" of the course, that is, how easy it is to get a good grade. This number represents the tendency of the teacher to give higher or lower grades. In the simulation, c_e is determined by taking a uniformly distributed random number between -0.5 and 0.5, indicating that teachers may skew their grades by up to half a letter grade up or down. A uniform distribution is used rather than a normal distribution to make the courses vary in difficulty over the entirety of a small range.

E.2 Simulating Students

We want to take the following things into consideration when creating simulated students:

- Students have varying strengths and weaknesses. In particular, some students have different ability levels in the sciences and humanities. Students prefer courses within their comfort zones.
- Students prefer getting higher grades.

Each simulated student s has two attributes, s_s and s_h. Both of these are numbers representing grades which indicate the student's abilities in the sciences and humanities, respectively. Both range from 0 to g_{max}, which is either 4.0 or 4.3 depending on the grading scale.

Given a course c and a student s, the grade for that student in that course is given by

$$g = \min(s_s c_s + s_h c_h + c_e, g_{max}) \tag{1}$$

In the simulation, s_s and s_h are determined by taking random numbers from a normal distribution with mean 3.5 and standard deviation 1.0, with a maximum of g_{max}.

E.3 Generating a Simulated Population

The simulated population is created by first generating a number of courses and a number of students. A course load is selected for each student s by repeating the following: First, a course c is selected at random. If the student is weak in science ($s_s < 2.5$) and the course is heavy in science ($c_s > 0.75$) then the course is rejected. Similarly, if the student is weak in humanities and the course is heavy in humanities, the course is rejected. If the student estimates his or her overall grade at less than 2.5, the course is rejected. This process of selection and rejection is repeated until a course is not rejected, but at most ten times, and then the last course is taken no matter what. The selected course is then added to the student's schedule and the grade computed as stated in Equation 1, rounded to the nearest possible grade.

The rejection process allows for the students' preferences in selecting courses, and the fact that at most ten courses can be rejected allows for distribution requirements.

E.4 Analysis of the Simulated Data

The simulation program was used to create 1000 students and 200 courses, where the course load was six. Thus, there were around $1000 \times 6/200 \approx 30$ people in each course, which is reasonable. Two runs were made, one with only whole grades, and one with + and − grades allowed.

We can determine a lower bound for the average GPA at ABC College. Suppose we have N students, each of which takes M courses. Denote by g_{ij} the grade of student i in that student's jth course. Then the average grade for that entire class is given by

$$\frac{\sum_{i=1}^{N} \sum_{j=1}^{M} g_{ij}}{NM} \tag{2}$$

The average GPA is given by

$$\frac{\sum_{i=1}^{N} \frac{\sum_{j=1}^{M} g_{ij}}{M}}{N} \tag{3}$$

The two are equal, so if the average grade at ABC College is A-, then the average GPA should be no more than 3.5. Any GPA less than 3.5 would be rounded to a B+ or less, and those greater than 3.5 would be rounded to A- or better. In the both data sets, the median GPA was 3.5, which is in agreement with the information given about ABC College.

E.5 Strengths and Weaknesses of the Simulation

The computation runs very quickly – in a few minutes – even though it was written in a high-level interpreted language (Python). It is very flexible, and can be adjusted to reflect different grade distributions as may be found in different colleges. It takes into account variation in student interest and in course material.

However, most of the courses turn out the roughly same size. Many colleges have a high proportion of small, seminar style courses, and there are almost always some very large lectures. The simulation ranks the whole school together and does not distinguish among the classes. There are only two majors in the simulation, sciences and humanities, and while there are forces within the simulation that push students into taking more courses in their preferred area of knowledge, there are no guarantees that the resulting schedules accurately reflect major requirements. There are also no prerequisites enforced, and thus no courses which are predominantly freshmen and seniors. This also means that the simulator cannot realistically create courses for more than one year.

The ideal course of action is to find some real grades and test the ranking schemes on them. Once ABC College has more knowledge of which scheme they wish to use, they should make grades from recent years available for further testing.