Alternatives to the Grade Point Average for Ranking Students

Jeffrey A. Mermin
W. Garrett Mitchener
John A. Thacker
Duke University
Durham, NC 27708–0320
wgm2@acpub.duke.edu

Advisor: Greg Lawler

Introduction

The customary ranking of students by grade point average (GPA) encourages students to take easy courses, thereby contributing to grade inflation. Furthermore, many ties occur, especially when most grades are high. We consider several alternatives to the *plain GPA ranking* that attempt to eliminate these problems while ranking students sensibly. Each is based on computing a revised GPA, called an *ability score*, for each student. We evaluate these alternative methods within the context of the fictitious ABC College, where grades are inflated to the extreme that the average grade is A−.

- The *standardized GPA* replaces each grade by the number of standard deviations above or below the course mean. Students are then ordered by the average of their revised grades.
- The *iterated adjusted GPA* compares the average grade given in a course to the average GPA of students taking it, thereby estimating how difficult the course is. It repeatedly adjusts the grades until average grade equals the average GPA and uses the corrected GPA to determine rank.
- The *least-squares method* assumes that the difference between two students' grades in a course is equal to the difference between their ability scores. It then sets up a large matrix of linear equations, with an optional handicap for courses taken outside a student's major, and solves for the ability scores with a least-squares algorithm.

The UMAP Journal 19 (3) (1998) 279–298.

An acceptable ranking method must reward students for scoring well, while taking into account the relative difficulties of their courses. It must clearly distinguish the top 10% of students. Preferably, the method should make allowances for the fact that students often earn lower grades in courses outside their majors and should not discourage them from taking such courses.

We used a small simulated student body to explore how the different methods work and to test the effects of changing a single grade. The least-squares method gave the most intuitive and stable results, followed by the iterated adjusted, the standardized, and finally the plain GPA. Under the least-squares and iterated adjusted methods, when a certain student's grade was changed in one course, that student and other students in that course changed position but most of the other students moved very little.

We used a larger simulated student body, generated by a computer program, to compare the iterated adjusted and standardized algorithms. They agree on most of students in the top decile, around 89% if plus and minus grades are included. They did not agree well with the plain GPA ranking, due to massive ties in the latter.

All four methods are more reliable when plus and minus grades are included, since a great deal of information is lost if only letter grades are given.

We recommend the least-squares method, since it is not very sensitive to small changes in grades and yields intuitive results. It can also be adapted to encourage well-roundedness of students, if the college chooses.

However, if there are more than about 6,000 students, the least-squares method can be prohibitively difficult to compute. In that case, we recommend the iterated adjusted GPA, which is easier to calculate and is the best of the remaining methods.

We recommend against the standardized GPA, because it does not properly correct for course difficulty, makes assumptions that are inappropriate for small or specialized courses, and produces counterintuitive results. We also recommend against the plain GPA, because it assumes that all courses are graded on the same scale and results in too many ties when grades are inflated.

To avoid confusion, we use the following terminology: 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.

Assumptions and Hypotheses

- It is possible to assign a single number, or "ability score" (this will be the revised GPA), to each student, which indicates the student's relative scholastic ability and, in particular, the student's worthiness for 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 completely order students by rank.

- The performances of an individual student in all courses are positively correlated, since:
  - There is a degree of general aptitude corresponding to the ability score that every student possesses.
  - All instructors, 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 that 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, that is, students tend to select courses at which they are most talented.
  - All students should major in an area of expertise, so that they are most talented at courses within or closely related to their majors.
  - The college may require courses that reflect its emphasis; even if the required courses could be considered "unfair" because they are weighted towards one subject (e.g., writing), that is the college's choice and highly ranked students must do well in such 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., A$-$ ≈ 3.7, while C$+$ ≈ 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 that rounds to an A$-$.
- In general, X should be ranked ahead of Y (we write X $>$ Y) if:
  - X has better grades than Y, and
  - X takes a more challenging courseload than Y, and
  - X has a more well-rounded courseload (we recognize that this point is debatable).

Analysis of Problem and Possible Models

The Problem with Plain 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 course and then comparing these values to determine the student's class rank.

The immediate problem with the plain 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 plain GPA exacerbates its own problems by encouraging the grade inflation that makes it so useless. Since the plain GPA does not correct for course difficulty, students may seek out courses in which it is easy to get a good grade. Faced with the prospect of declining enrollment and poor student evaluations, instructors who grade strictly may feel pressure to relax their grading standards. Instructors who grade easily may be rewarded with high enrollment and excellent evaluations, potentially leading to promotion. The entire process may create a strong push towards grade inflation, since the plain GPA punishes both the student taking a difficult course and the instructor teaching it. Any system intended to replace the plain GPA should address this problem, so that grade inflation will be arrested and hopefully reversed.

Another potential concern is that the plain GPA encourages specialization by students. Since students tend to perform better in courses related to their majors, the GPA rewards 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.

Three Possible Solutions

Several potential alternatives to GPA ranking directly compare grades within each course. Under such a system, the following considerations come into play:

[1]Repeated trials of the process described later yield this result.

- It is not possible to compare students just to others in their own 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 describe them in more detail.

- For the *standardized GPA* each student is given a revised GPA based on the student's grade's position in the distribution of grades for each course.
- The *iterated adjusted GPA* attempts to correct for the varying difficulties of courses. In theory, every grade given to a student should be approximately equal to the student's GPA, so that 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 *least squares* method 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 $X - Y = 4.0 - 3.0 = 1.0$). Since in any nontrivial 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.

Standardized GPA

How It Works

The standardized GPA is perhaps the simplest method and one most in keeping with the dean's suggestion. In each course, we determine how many standard deviations above or below the mean each student's grade is. This *standard score* becomes the student's "grade" for the class, the student's standard scores are averaged for a *standardized GPA*, and students are ranked by standardized GPA. 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.

Strengths

- The standardized GPA is not much more difficult to calculate than the plain GPA measurement.
- Each course can be considered independently. Instead of waiting for all results to come in, the registrar can calculate the standardized scores for each course as grades come in, possibly saving time in sending grades out.
- The standard deviations do correct for differing course averages, for example, getting a B+ when the course average is a C+ looks better than getting an A− when the course average is an A. At the same time, this method 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.

Weaknesses

The standardized GPA suffers from many of the same problems as the plain GPA.

- It does not reward students who have a more well-rounded courseload. Instead, students are punished severely if they perform at less than the course average; for example, a student who takes a course outside his or her major is likely to score worse than students majoring in the course's subject.
- The plain GPA makes no distinction between easy and difficult courses and thus encourages easy courses. The standardized 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 only by students who excel both in the subject of the course and in general, 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 the course have lower scores.
  - Additionally, ability scores may be significantly raised by adding poor students to the course.
- The standardized GPA does *not* assume that instructors assign grades based on a normal curve or to fit any other prespecified distribution. Not all instructors grade on the normal curve or even on any curve. Some courses may require grades to fit some other distribution in order to be fair, for example, if all the students are extraordinarily talented.
- The method 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. The difficulty of a course should be measured not only by the grades of its students but also by the aptitudes of those students.

Consequences

Grading based on deviation from the mean fosters cutthroat competition among students, since any student's ability score may be significantly raised by lowering the ability scores of other students.

Iterated Adjusted GPA

How it Works

Rather than directly comparing students, this method compares courses. Suppose that a course is unusually difficult. Then students should receive lower grades in that course relative to their others, so the average grade in that course should 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 GPAs will be different, and most likely some courses will need further correction. The iterated adjusted GPA method makes ten corrections to all grades, then sorts students in order of corrected GPA. (Our numerical experiments show that ten iterations are sufficient to bring the difference between the average GPA and the average grade down to zero, to three decimal places.)

Strengths

- This algorithm is fairly quick to compute, taking only a couple of minutes for 1,000 students, 200 courses, and 6 courses per student.
- The computation is straightforward to explain and easily understood by non-experts.

Weaknesses

- All grades from all courses must be known in order to run the computation.
- The corrected grades cannot be computed independently by students.

- There is no guarantee that the corrected GPAs will be comparable across semesters; to compute overall class rank at graduation, it will be necessary to average ranks across semesters, rather than average corrected GPAs.

Consequences

This method systematically corrects for instructor 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 grades are unusually high. If students score high in a course relative to their other grades, it could be because the course was easy or because the students put forth 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 is punished by the correction, and if he gets a lower grade, he is rewarded—which is clearly undesirable. Using the sample data set presented later in **Table 1**, we experimented with independent studies and determined that they had minimal impact on the rank order. However, to avoid the possibility of such strange results, independent studies should be ignored in the computation.

The Least-Squares Algorithm

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 X and Y take the same course, and get grades A and B, then we have a difference $X - Y = 4.0 - 3.0 = 1.0$. We further assume that students majoring in natural science fields perform better in natural science courses than in humanities courses, and vice versa, and that the difference is of approximately the same for all students; we call it H_H. Hence, if, in the example above, students X and Y are taking a mathematics course, but X is majoring in physics and Y is majoring in literature, we have $X - (Y + H_H) = 1.0$.

A course with N students generates $N(N+1)/2$ such linear equations; the abilities of each student are the solution to the set of all such equations from every course offered during the semester. In practice, these equations never have a solution. 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 H_H, and b is the right-hand side of each equation. Multiplication by the transpose of A yields the equation $A^T Ax = A^T b$. This matrix equation has a one-dimensional solution set, with nullspace equal to scalar multiples of $(1\,1\,1\,\ldots\,1\,0)^T$, where the 1s 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 plain GPA. These scores can be averaged over all eight semesters to produce a ranking at graduation.

Strengths

- Least squares corrects for the difficulty of every student's courseload.
- Least squares can reward students for carrying a well-rounded courseload. 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, with no consequence except that the ability scores will no longer consider the balance or specialization in each student's courseload.
  - If a school wishes to emphasize several areas of specialization rather than just two, it could do so by replacing H_H with constants representing the difficulty of the transitions between each pair.
  - A school wanting to assure that certain emphasized courses (e.g., a freshman writing course) not unduly benefit students majoring in some departments could categorize such courses 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.
  - Other corrections may be made for students with special circumstances; for example, if a student double-majors in two different areas of specialization, each well-roundedness correction might be replaced by the average of the two corrections from each of the student's major areas.

Weaknesses

The most glaring weakness of this method is that it involves huge amounts of computation and may severely tax computing resources at larger universities.

For a student body of 6,000, with 120 courses of size 20 and each student taking 4 courses, we have $1{,}200 \times 21(21+1)/2 \approx 250{,}000$ pairs of grades. This results in a sparse A with 250,000 rows, 6,000 columns, and only 4 nonzero entries in each column (for the 4 courses that the student took). Then A^TA has 36,000,000 entries; at 4 bytes per entry, keeping it in memory requires 144 MB, barely within range of current medium-size computers. Computing A^TA takes on the order of $250{,}000 \times 6{,}000^2 = 9 \times 10^{12}$ multiplications, computing A^Tb takes only about 1.5×10^9 multiplications, and solving $A^TAx = A^Tb$ takes about $6{,}000^3 = 2.2 \times 10^{11}$ operations. Thus, the time to solve the system is about 10^{13} operations, which would take $50{,}000$ sec $\approx$ 14 hr on a 200 MHz computer.

The memory needed increases with the square of the number of students and quickly becomes infeasible with this approach and current technology.

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 instructors widening their grading scales, in order to reward their best students, thus reducing grade inflation to something more reasonable.

A Small Test Population

We postulate a minicollege, with 18 students (A–R), that offers only the following courses: Math, Physics, Computer Science, Physical Education, Health, English, French, History, Philosophy, Psychology, and 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 **Table 1**. Just looking at these transcripts, without analyzing them numerically, we find that we should have the following, which any valid ranking system *must satisfy* (recall that X > Y means that X should be ranked above Y):

- A > B; C > D; and E > F, and so on, because A, C, etc., carry better grades than B, D, etc., in courseloads of similar difficulty.
- O, D > J because O and D have slightly better grades than J in more difficult courseloads.
- E > D because E has better grades in a more difficult courseload.

We also recognize the following relationships as *desirable*:

- O > Q, R and P > R, because O and P have almost as good grades and much more difficult schedules.

Table 1.
Transcripts of the test population.
A star indicate the student's major. "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

- M > Q and N > R, because M and Q have similar grades but M has a more difficult schedule, and similarly for N and R.
- K > M, N, Q, R because K has similar grades in a much more difficult schedule.
- C, G, and K should be ranked near each other because they have similar grades in similar schedules.
- P > J because P 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 of a student's schedule should affect rank, we also find the following relationships:

- E > C, D because E has almost as good grades in a more difficult, much more well-rounded schedule.
- I > K, M because I has similar grades against a more well-rounded schedule.

The rankings of this sample population are given in the **Table 2**. A comparison of the different methods relative to the criteria that we have set out is in **Table 3**. Least squares does best, followed by iterated adjusted, standardized, and plain.

Table 2.
Rankings of the sample population under the various methods.

	With +/−								Without +/−							
Rank	Plain		Standardized		Iterated		LS		Plain		Standardized		Iterated		LS	
1	Q	4.25	K	0.84	K	4.22	E	2.32	R	4.00	G	0.53	L	4.12	G	2.25
2	M	4.25	I	0.81	I	4.17	I	2.26	Q	4.00	L	0.49	G	4.07	I	2.19
3	K	4.17	Q	0.60	M	4.09	G	2.24	C	4.00	C	0.39	I	4.06	E	2.17
4	I	4.08	M	0.52	C	4.08	C	2.24	N	4.00	I	0.36	C	4.05	C	2.17
5	R	4.00	G	0.39	G	4.07	O	2.18	M	4.00	N	0.34	K	4.03	F	2.14
6	N	4.00	C	0.22	L	4.06	K	2.14	G	4.00	K	0.27	N	3.96	O	2.04
7	C	4.00	E	0.21	E	4.05	M	2.05	K	4.00	M	0.24	E	3.92	L	2.04
8	G	4.00	L	0.16	Q	4.02	Q	2.03	I	4.00	Q	0.24	M	3.89	D	2.03
9	L	3.92	N	−0.01	O	3.96	F	2.01	L	4.00	R	0.23	J	3.87	R	2.02
10	E	3.83	O	−0.03	N	3.90	R	1.99	O	3.75	F	0.11	D	3.84	K	1.98
11	O	3.83	R	−0.20	R	3.76	P	1.94	J	3.75	J	0.07	F	3.84	J	1.95
12	D	3.67	D	−0.26	D	3.74	D	1.93	F	3.75	E	0.07	O	3.83	N	1.90
13	J	3.58	A	−0.27	F	3.69	L	1.92	E	3.75	D	−0.12	Q	3.81	M	1.88
14	F	3.58	F	−0.28	J	3.66	N	1.87	D	3.75	O	−0.15	R	3.80	P	1.87
15	H	3.50	H	−0.45	P	3.62	J	1.74	H	3.50	H	−0.30	P	3.58	Q	1.85
16	P	3.50	J	−0.49	H	3.41	H	1.60	P	3.50	P	−0.60	H	3.39	H	1.58
17	A	3.42	P	−0.59	A	3.36	A	1.44	A	3.25	A	−0.61	A	3.18	A	1.26
18	B	2.92	B	−1.16	B	2.76	B	0.89	B	2.75	B	−1.56	B	2.59	B	1.54

Table 3.
Number of criteria satisfied by each method on the minicollege data set, for +/− grades.

	Plain	Standardized	Iterated	Least Squares
Required (20)	all	all	all	all
Desirable (13)	5	6	8	9
Well-roundedness (4)	1	2	2	all

Test Population Redux (No +/− Grades)

We now take the test population and drop all pluses and minuses from the grades. Again, we determine some basic *required* relationships that any valid ranking system must satisfy:

- A > B; C > D; and G > H since A, C, and G have better grades in similar courses.

We also recognize the following relationships as *desirable*:

- O > P because O has slightly better grades in the same courseload.
- E > F because E has the same grades in a more difficult courseload.
- O > Q, R because O has almost equivalent grades in a much more difficult courseload.
- C > I, G because C has the same grades in a more difficult courseload.
- I > K, L because I has the same grades in a more difficult courseload.
- K, L > M, N because K and L have the same grades in a more difficult courseload.

- M, N > Q, R because M and N have the same grades in a more difficult courseload.

If we postulate that the well-roundedness of a student's schedule should affect 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; and
- 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 the right-hand half of **Table 2**. **Table 4** gives a comparison of the methods.

Table 4.
Number of criteria satisfied by each method on the minicollege data set (no $+/-$ grades).

	Plain	Standardized	Iterated	Least Squares
Required (3)	all	all	all	all
Desirable (12)	1	6	9	9
Well-roundedness (6)	3	1	3	4

Stability

How Well Do the Models Agree?

We have four ways of ordering students: plain GPA, standardized GPA, iterated adjusted 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, we created by means of a computer simulation a population of 1,000 students and 200 courses, with 6 courses per student. The details of the simulator are explained in the **Appendix**. We implemented all of the algorithms except least squares, which was too difficult for the available time. A single run of the simulation is analyzed here, but these results are typical of other runs.

With Plus and Minus Grades

See **Figures 1–3** 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 adjusted or the standardized rankings. There are lots of scattered points, which is due mostly to the facts that there are lots of ties in plain GPA rankings (especially near the top of the class) and that tied students are ordered more or less at random. Very few ties are present in any of the other methods.

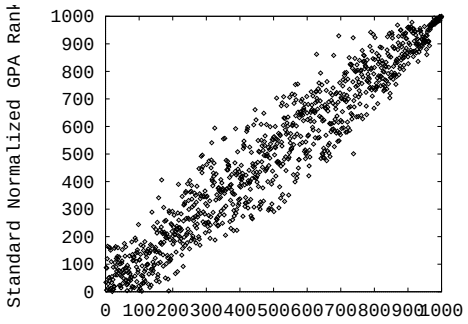

Figure 1. Plain GPA rankings vs. standardized GPA rankings, using simulated students.

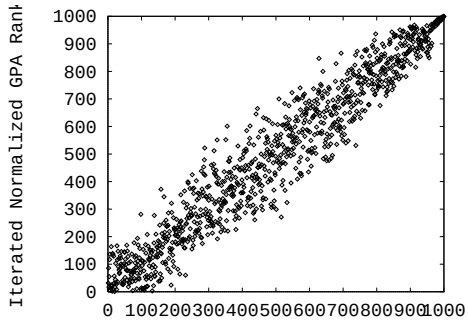

Figure 2. Plain GPA rankings vs. iterated adjusted GPA rankings, using simulated students.

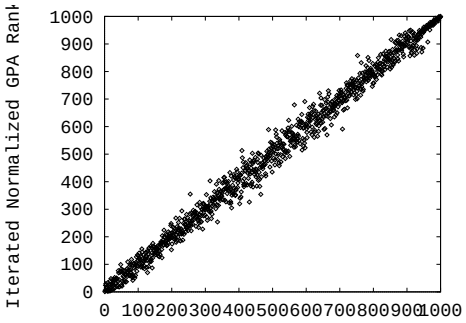

Figure 3. Standardized GPA rankings vs. iterated adjusted GPA rankings, using simulated students.

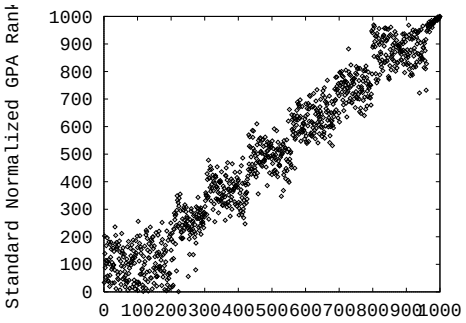

Figure 4. Plain GPA rankings vs. standardized GPA rankings, using simulated students, with no plus or minus grades.

The iterated adjusted and standardized rankings are in better agreement, with fewer outlying points. These two methods agree on 89 of the 100 students in the top decile.

Without Plus and Minus Grades

See **Figures 4–6** 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 nontraditional methods due to these

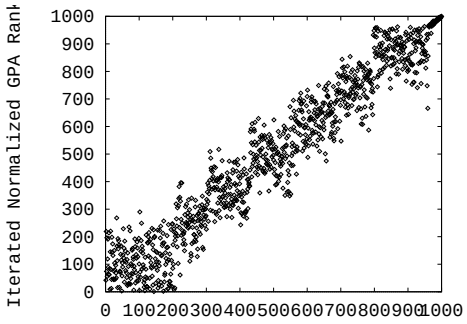

Figure 5. Plain GPA rankings vs. iterated adjusted GPA rankings, using simulated students, with no plus or minus grades.

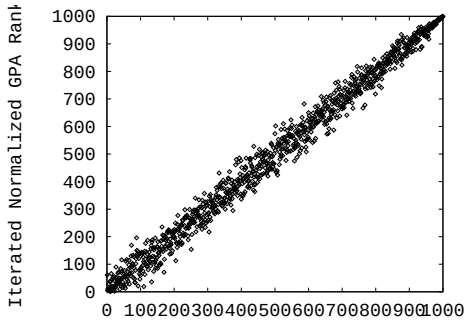

Figure 6. Standardized GPA rankings vs. iterated adjusted GPA rankings, using simulated students, with no plus or minus grades.

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.

How Much Does Changing One Grade Affect the Outcome?

If one grade of one student is changed, the student's 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 nontraditional rankings, each student's rank is determined relative to the other students, and one changed grade might trigger a chain of rank changes.

To test sensitivity, the sample population was modified slightly: 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; with only whole letter grades, Q dropped from 2nd to 16th. In both cases, there were no changes in the order of other students except to make room for Q.

- For the standardized 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 rank each. Student I dropped three places, perhaps because of how N and K benefitted from Music History.

- The iterated adjusted 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 dropped from 13th to 16th. J and K improved a couple of places, benefitting from the increased difficulty of Music History, while G dropped again, three places this time. O and F switched places for no obvious reason.

- Using least squares and plus and minus grades, Q dropped 9 places. Other members of the Music History course J and K improved a bit, and L improved a lot. With letter grades only, Q dropped from 15th to 16th, and J, K, and L improved. 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 rank and the rest move up or down at most one rank to compensate. The next most stable seems to be least squares, followed by iterated adjusted, and finally standardized. In each scheme, the coursemates of the person whose grade changed are most likely to change rank. There were a few chain-reaction reorderings, which are harder to explain. Also, having plus and minus grades appears to improve stability in general.

How Does Course Size Affect the Outcome?

Another simulation was run with 1,000 students, 500 courses, and 6 courses per student. Courses came out smaller, and the correlation between the standardized ranking and the iterated adjusted 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 GPAs.

Strengths and Weaknesses of Each Model and Recommendations

If the college wishes to promote well-roundedness over specialization (we would suggest this), and has a fairly small population (fewer than about 6,000 students), we recommend the least-squares method. Otherwise, we recommend the iterated adjusted 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 standardized GPA.
- It is more flexible than either the iterated or standardized GPA.

- It is clear and easily understood.

The iterated adjusted 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 that are roughly as consistent with intuitive observation as the least-squares method.

We feel that the standardized 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 that are no more consistent with intuitive observation than those produced by the plain GPA.

Further Recommendations

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 GPAs) 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 standardized and iterated adjusted schemes are likely to encounter opposition because they directly alter the point values of grades during computation. The least-squares method simply reinterprets them and is less likely to make instructors feel that their authority has been violated.

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.

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 instructors to use them with care.

Appendix: Details of the Simulation

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 instructor 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

instructors may skew their grades by up to half a letter grade up or down. We use a uniform distribution rather than a normal distribution so as to make the courses vary in difficulty over the entirety of a small range.

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 that 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

$$g = \min\left(S_s c_s + S_h c_h + c_e, g_{\max}\right). \tag{1}$$

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 $g_{\max}$.

Generating a Simulated Population

The simulated population is created by first generating a number of courses and a number of students. A courseload 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 **(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.

Analysis of the Simulated Data

The simulation program was used to create 1,000 students and 200 courses, where the courseload was six. Thus, there were around $1{,}000 \times 6/200 \approx 30$ 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 whom 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

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

and the average GPA is given by

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

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 agrees with the information given about ABC College.

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 roughly the 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 that are predominantly populated by freshmen and seniors. This also means that the simulator cannot realistically create courses for more than one year.