Judge's Commentary: The Outstanding Grade Inflation Papers

Daniel Zwillinger
Waltham, MA 02453
zwillinger@alum.mit.edu

Grade point average (GPA) is the most widely used summary of undergraduate student performance. Unfortunately, combining student grades using simple averaging to obtain a GPA score results in systematic biases against students enrolled in more rigorous curricula and/or taking more courses. Here is an example [Larkey and Caulkins 1992] of four students (call them I–IV) in which Student I always obtains the best grade in every course that she takes and Student IV always obtains the worse grade in every course that he takes, yet Student I has a lower GPA than Student IV does:

	Student I	Student II	Student III	Student IV	Course GPA
Course 1	B+			B−	3.00
Course 2	C+		C		2.15
Course 3			A	B+	3.65
Course 4	C−	D			1.35
Course 5		A		A−	3.85
Course 6	B+			B	3.15
Course 7		B+	B		3.15
Course 8	B+	B	B−	C+	2.83
Course 9		B	B−		2.85
Student GPA	2.78	2.86	2.88	3.0	

The MCM problem was to determine a "better" ranking than one using pure GPAs; this problem has no simple "solution." Johnson [1997] refers to many studies of this topic and suggests a technique that was considered—but not accepted—by the faculty at Duke University.

Each participating team is to be commended for its efforts in tackling this problem. As in any open-ended mathematical modeling problem, there is not only great latitude for innovative solution techniques but also the risk of finding

The UMAP Journal 19 (3) (1998) 323–327.

no results valuable in supporting one's thesis. Solutions submitted contained a wide variety of approaches, including graph theory and fuzzy logic.

Unfortunately, several teams were confused as to the exact problem that the dean wanted solved. Assigning students to deciles, by itself, was not the problem; for example, deciles could be assigned from any list of student names by choosing the first 10% of the students to be in the first decile, etc. The dean wanted meaningful deciles, based on students' relative course performance. Simply re-scaling GPAs so that the average became lower (and the top 10% became more spread out) would not change the inherent problem.

The problem statement suggested that relative rankings of students within courses should be used to evaluate student performance. With this assumption, possible approaches include:

- using relative ranking *with* grade information

  (A useful additional assumption might be that faculty would give grades based on an absolute concept of what constitutes mastery of a course.)

- using relative ranking *without* grade information

In the latter approach (chosen by most teams), an instructor who assigns As to all students in a course provides exactly the same information as an instructor who assigns all Cs to the same students in another course.

Specific items that the judges looked for in the papers included:

- Reference to ranking problems in other fields that use relative performance results, such as chess and golf.
- A detailed worked-out example, illustrating the method(s) proposed, even if there were only 4 students in the example.
- Computational results (when appropriate) and proper consideration of large datasets. Teams that used only a small sample in their computational analysis (say 20 students) did not appreciate many of the difficulties with implementing a grade adjustment technique.
- Mention (if not use) of the fact that even though the GPA may not be as "good" a discriminator as the various solutions obtained by the teams, it seems reasonable that there be some correlation between the two.
- A response indicating understanding of the question about the changing of an individual student's grade. Such a grade change could affect that student's ranking, but if it affected many other students' ranks then the model is probably unstable.
- A clear, concise, complete, and meaningful list of assumptions. Needed assumptions included:

  - The average grade was A$-$. (This must be assumed, as it was stated in the problem statement—amazingly, several teams assumed other starting averages!)
  - in an {A+, A, A$-$, ...} system, not *all* grades were A$-$. (Otherwise, there is no hope for distinguishing student performance.)

  Many teams confused assumptions with the results that they were trying to obtain. Teams also made assumptions that were not used in their solution, were naive, or needed further justification. For example,

  - Many teams assumed a continuous distribution of grades. As an approximation of a discrete distribution, this is fine. However, several teams allowed grades higher than A+, and other teams neglected to convert to a discrete one when actually simulating grades.
  - Several teams assumed that teachers routinely turn in a percentage score or course ranking with each letter grade. This, of course, would be very useful information but is not realistic.
  - Low grades in a course do not necessarily imply that a course is difficult. A course could be scheduled only for students who at "at risk." Likewise, a listing of faculty grading does not necessarily allow "tough" graders to be identified: an instructor may teach only "at risk" students.

- The most straightforward approaches to solving this problem were:
  - Use of information about how a specific student in a course compared to the statistics of a course. For example, "Student 1's grade was 1.2 standard deviations above the mean, Student 2's grade was equal to the mean, ..." The numbers $\{1.2, 0, \ldots\}$ can be used to construct a ranking.
  - Use of information about how a specific student in a course compared to other specific students. For example, "in Course 1, Student 1 was better than Student 2, Student 1 was better than Student 3, ..." This information can be used to construct a ranking.

  The judges rewarded mention of these techniques, even if other techniques were pursued.

- Other features of an outstanding paper included:
  - Clear presentation throughout
  - Concise abstract with major results stated specifically
  - A section devoted to answering the specific questions raised in the problem statement or stating why answers could not be given.
  - Some mention of whether the data available (i.e., sophomores being ranked with only two years worth of data) would lead to statistically valid conclusions.

None of the papers had all of the components mentioned above, but the outstanding papers had many of these features. Specific pluses of the outstanding papers included:

- Duke team
  - Their summary was exemplary. By reading the summary, you could tell what they were proposing and why, what the issues they saw were, and what the models they produced were.
  - Their use of least squares to solve an overdetermined set of equations was innovative.
  - Their figures of raw and "adjusted" GPAs clearly and visually showed the correlation between the two and also the amount of "error" caused by exclusive use of GPAs.
- Harvey Mudd team
  - Their sections on "Practical Considerations" and "What Characterizes a Good Evaluation Method" demonstrated a clear understanding of the problem.
  - Their figures of "Raw GPA" versus "Student Quality" clearly and visually showed the correlation between the two and also the amount of "error" caused by exclusive use of GPAs.
- Stetson team
  - Their use of the median, as well the mean, in comparing a specific student to the statistics of a course was innovative. (Use of the median reduces the effects of outliers.)
  - They interpreted the results of rank adjustment for specific individuals in their sample.
  - An awareness of the problem as indicated by literature references in their "Background Information" section.

Reference

Johnson, Valen E. 1997. An alternative to traditional GPA for evaluating student performance. *Statistical Science* 12 (4): 251–278.

Larkey, P., and J. Caulkins. 1982. Incentives to fail. Working Paper 92–51. Pittsburgh, PA: Heinz School of Public Policy and Management, Carnegie Mellon University.

About the Author

Daniel Zwillinger attended MIT and Caltech, where he obtained a Ph.D. in applied mathematics. He taught at Rensselaer Polytechnic Institute, worked in industry (Sandia Labs, Jet Propulsion Lab, Exxon, IDA, Mitre, BBN), and has been managing a consulting group for the last five years. He has worked in many areas of applied mathematics (signal processing, image processing, communications, and statistics) and is the author of several reference books.