Practitioner's Commentary: The Outstanding Grade Inflation Papers

Valen E. Johnson
Institute of Statistics and Decision Sciences
Duke University
Box 90251
Durham, NC 27708-0251
valen@isds.duke.edu

Introduction

I would like to begin my comments by congratulating all three teams on their innovative solutions to what is both a difficult and important societal problem. The depth of thought given to this problem in the short time each team had to generate a solution is very impressive, and all teams present solutions that are surprisingly close to proposals that have appeared in the educational research literature. In fact, the three proposed solutions span the range of previously proposed alternatives to GPAs in terms of model complexity, ranging from relatively simple to highly complex. In the discussion that follows, I focus on the problems associated with each proposal rather than their strengths. I chose this course not because the proposals are weak but instead because they are of sufficient quality to merit serious criticism.

As all three teams note, GPA plays an important role in our educational system. In the particular scenario presented, an adjusted GPA is needed to more equitably allocate scholarships to students at ABC College. More generally, however, GPA is arguably the single most important summary of a student's academic performance while in college. It plays a critical role in determining the success of a student in the job market and is influential in determining whether or not a student is admitted to professional or graduate school.

A more subtle influence of GPA is the impact that GPA has on student course selection. Because GPA is perceived to play such a critical role in a student's career, it is now common for students to select courses based on their expectations of how courses will be graded. In a recent survey of Duke University

The UMAP Journal 19 (3) (1998) 329–336.

undergraduates, 69% of participating students indicated that expected grading policy had some influence in their decision to enroll in courses taken to satisfy a distributional requirement! This fact suggests that fewer "hard" courses are taken by undergraduates as a result of differential grading policies and probably causes a net decrease in the number of science and mathematics courses taken. To a large extent, it also explains the spiraling assignment of grades that has taken place over the last decade. Students gravitate towards courses that are graded leniently, and professors soften grading standards to ensure adequate course enrollments and favorable course evaluations.

Changing the way GPA is computed can solve all of these problems, but implementing such a change is a difficult proposition. Any change to the current GPA system will be opposed by faculty and students who do not benefit from the change, and every meaningful modification of the GPA will produce a sizable proportion of each. For this reason, it is crucial that any modification to the current GPA system be both logically consistent and fair. Additionally, any change to the GPA must be understandable by nonstatisticians, at least at a rudimentary level. Simplicity is a benefit. Unfortunately, the fact that essentially all American universities report traditional GPA attests to the fact that no alternative is available that is

- simple,
- fair, and
- logically consistent.

Compromises in one or more of these three criteria are therefore inevitable. In evaluating the proposed solutions to ABC College's grade inflation problem, I will attempt to consider all three of these criteria and indicate the extent to which I feel each is compromised.

Stetson University

The proposal by the team from Stetson University represents the simplest solution to the problem. Their proposal essentially is to standardize grades in each class using either the median or mean grade, depending on the value of the coefficient of skewness.[1]

From a statistical standpoint, the use of the median in the standardization offers robustness to outliers, or extreme observations. For grade data, extreme observations are usually Ds, and Fs. One or two Fs in a class can significantly affect the mean grade assigned but does not affect the median grade any more than one or two "below average" grades would.

[1] The formula provided for the standardization appears to have a minor error. As presented, each student's grades would always be standardized by either the mean grade or median grade. Presumably, it was the intent of this group to use either the mean or median grade on a course-by-course basis.

Unfortunately, the use of the median grade in the standardization process introduces several potential difficulties.

- First, for highly discrete data (i.e., data taking on only two or three values) the median value can be very uninformative. As an illustration of how bad the median can be, consider the problem of estimating the probability of success for a new cancer treatment. If cures are coded as 1s and deaths as 0s, the median estimate for the probability of a cure in a clinical trial conducted with an odd number of patients must be either 0 or 1! A similar difficulty arises when analyzing student grades when only two or three unique grades are assigned. When grades are inflated, the median grade is likely to be either an A or A−, but this says little about the relative proportions of As and A−s that were awarded, and less still about the proportion of Bs, Cs, Ds and Fs.

- Next, if skewness and outliers are considered problematic, then should not a robust estimate of the variance also be used? The Stetson team uses the sample variance to standardize the grades, but this estimate of the spread of a distribution is more sensitive to outliers than is the sample mean when estimating location. As an alternative to the sample variance, I would recommend that the interquartile range (IQR) be used as a measure of distributional spread when the median is used as a measure of centrality. Like the median, the IQR is robust to outliers; for normally distributed data, it is nominally equal to 1.35 standard deviations.

A compromise between the median (which is robust against outliers) and the mean (which is often statistically efficient) would be a *trimmed mean*. To compute a trimmed mean, a fixed proportion of the extreme values of the data is ignored. Thus, to compute the 10% trimmed mean, the lowest and highest 5% of grades are thrown out and the mean of the remaining data is computed. Trimmed means have the advantage of offering some robustness against outliers while at the same time maintaining good statistical efficiency. Use of the trimmed mean in the Stetson team's standardization procedure might also eliminate the problem of deciding when to switch between mean and median, which in itself introduces some potentially large jumps in the standardized values for small changes in skewness.

Although the Stetson team's proposal wins in terms of simplicity, it is somewhat weaker in terms of fairness and consistency.

- In my opinion, fairness is compromised because the standardization procedure does not account for the quality of students within a class, as the team members themselves comment. At my institution (Duke University), there are many courses known to be populated by top students, and if implemented, this proposal would encourage students to opt out of these courses. Accounting for the quality of students in a course is an important facet of GPA adjustment, and I would hesitate to recommend any method that didn't account for this aspect of classroom grading. Implementing such a method would encourage students to register for lower-level classes with less talented students.

- Consistency is also a problem. To see why, consider two students who take identical courses through their senior years, and receive A+ in all of their courses. In the last semester of their senior year, the second student develops an interest in art history and takes an introductory course in that subject (in addition to the other courses that both he and the first student take). Both students again receive A+s in all of their courses, but unfortunately everyone in the art history course also receives an A+. Which student graduates on top?

  According to the Stetson team's adjustment method, the first student beats out the second student for valedictorian, even though the second student tied the first student in all of the courses they took together and got an A+ in the one course he took above their normal course load. Why? The standardized grade for an A+ in the art history course is 0, which when averaged into the other A+s the student received would lower his adjusted GPA. It is interesting to note that the same problem exists if the art history course is replaced with an independent study course, though in that case it is not clear how the estimate of the standard deviation would be determined.

Duke University[2]

The team from Duke University discusses three proposals for adjusting GPA, the first of which is a nonrobust version of the standardization scheme proposed by the Stetson team. Their other proposals are based on regression-type adjustments to traditional GPA. In their iterated GPA adjustment, the grades from each class are adjusted for the difference between the mean grade assigned in a class and the mean adjusted GPA of students in that class. This difference is then used to compute new adjusted GPAs, which led to new adjustments to the class grades. The team's least-squares estimate of the adjusted GPA is based on the assumption that students tend to receive higher grades in classes taken in their academic majors. As I understand this proposal, they estimate the adjustments for each combination of major and course, conditionally on observed grades.

Both adjustment schemes are quite similar to an adjusted GPA proposed in a more formal framework by Caulkin et al. [1996], though the "least-squares" proposal is also similar to the pairwise course/department differences estimated by Strenta and Elliot [1987], Goldman and Widawski [1976], and Goldman et al.

[2] EDITOR'S NOTE: When he prepared his comments, Prof. Johnson was not aware that a paper from a Duke University team was among the papers that he reviewed. He is not affiliated with the Duke University Mathematics Dept. nor with any members of the Duke team. The copies of the Outstanding papers that he read were the same as those read by the judges; they contained code numbers but no identification of the papers' authors or their institutions.

[1974]. For comparison, the models proposed by Caulkin et al. [1996] have the general form

$$\text{Grade}_{ij} = \text{True GPA}_i + \text{Course effect}_j + \epsilon_{ij}.$$

Under this model, the grade received by student i in course j is assumed to be an additive function of their "true GPA" plus a course effect for the jth course, plus a (normally distributed) random error. Like the Duke team, Caulkin et al. [1996] also propose an iterative procedure for estimating all student's "true GPA" along with all course effects.

From a technical standpoint, I regard this modeling approach as a significant improvement on simple standardization schemes. This approach implicitly accounts for both the grading policies of individual instructors and the quality of students within each class. Algorithmically, such models are comparatively simple to estimate and can be computed in less than two minutes on a PC for datasets containing 12,000 students and 17,000 classes.

The primary drawback of these regression-type models is the assumption that grades are intervally scaled. In other words, it usually does not make sense to assume that the difference between an A and A− is the same as the difference between a C and C−, or that a D added to a B is equal to an A. Typical grading scales assign more probability to As and Bs than to Cs and Ds, and by not taking the ordinal nature of grade data[3] into account, substantial statistical efficiency is lost. These models also suffer from the paradox presented above for standardized GPAs but to a lesser extent. The art history course would lower the second student's "true GPA," but an independent study course would leave it unaffected.

Substantively, I have several minor objections to the modeling assumptions made by the Duke team. Perhaps most important, they premise their model on the assumption that "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." A similar assumption is made by the team from Harvey Mudd College. In fact, this assumption is not necessary or appropriate. Each student's true GPA can instead be interpreted as the ability score for the student *in courses that that student chose to take.* Some courses are required by the university to satisfy distributional requirements, but most will be courses that students choose to take in their areas of interest and competence. The model proposed by Caulkin et al. [1996] and the variation of it proposed by the Duke team can be applied without difficulty to colleges in which, say, humanities students are completely separated from engineering students in the sense that they have no common classes. In such cases, "true GPA" corresponds to the ability of students in the classes that they took. The grades assigned to engineering students in engineering classes should not be used to estimate the abilities of engineering students in humanities classes.

I also feel that it is important to distinguish difficult courses from courses that are graded stringently. They are not (always) the same, so it does not follow

[3] Ordinal data are data that are recorded in ordered categories.

that students should be penalized for taking courses that are graded leniently. In fact, I would argue that any student who receives the highest grade in all classes that she takes should be awarded a high adjusted GPA.

As a final comment on this proposal, I think it is dangerous to devise a ranking algorithm that rewards students for the type of curriculum chosen. For some students and some majors, a "well-rounded" courseload is desirable. For others, a more concentrated curriculum is apropos. A student who has satisfied the relevant distributional requirements for their university should be free to choose whatever courses they wish—without penalty. Indeed, I lament the fact that American mathematics undergraduates are often ill-prepared for graduate studies in statistics programs because they have taken so few mathematics courses.

Harvey Mudd College

The proposal by the team from Harvey Mudd College is also very interesting. Though there are several technical problems in their model specification, the paradigm proposed by this team is surprisingly close to a statistical model called the *Graded Response Model* (GRM). GRMs are normally introduced in advanced graduate-level statistics courses, and I was impressed by the extent to which this team exposed the underlying assumptions of these models.

As suggested by the Harvey Mudd College team, the basic assumption of a GRM is that instructors choose thresholds on an underlying achievement scale and assign grades to students based on the grade intervals into which their classroom achievement is observed to fall. Letting z_{ij} denote the classroom performance of student i in class j, and $\gamma_F^j, \gamma_D^j, \ldots, \gamma_B^j$ the upper cutoffs for each grade on the ability scale, the GRM assumes that student i receives a grade of, say, C in class j if

$$\gamma_D < z_{ij} < \gamma_C.$$

A further assumption of the model is that the mean ability of student i, say z_i, for the courses that student i chooses to take, is related to his performance in class j according to

$$z_{ij} = z_i + \epsilon_{ij}.$$

Here, ϵ_{ij} denotes a random deviation. In the GRM, grade cutoff vectors, mean student achievements, and the distribution of the error terms are estimated jointly. Details of this model are discussed in further in Johnson [1997].

In terms of the GRM, the Harvey Mudd team's q_i is roughly equivalent to z_i, while c_{ij} is comparable to ϵ_{ij}, and d_j plays a role similar to the variance of the distribution of the error term ϵ_{ij} if this distribution is assumed to be the same for all students in a given class. The primary difference between the proposed model and the GRM is that the grade-cutoff vectors γ^j are fixed in the former and estimated in the latter. Although the team takes a reasonable approach toward fixing these cutoffs, doing so leads to several inconsistencies in the resulting model. Partially to overcome these difficulties, the team introduces a harshness term h_k. This harshness term models a uniform shift of the grade-cutoff values for all classes taught by professor k. The team estimates values of h_k from the mean grades assigned by each professor.

The most important technical defect in this model is caused by the assumption that the grade-cutoffs for each class can be obtained by a contraction and shift of the baseline cutoffs. When harshness is 0, it follows that large increases in d_j result in more extreme grades (that is, more As and Fs) and fewer middle grades. Shifts in harshness can change the As to Bs or Cs, but there is still a gap between the high and low marks. This is not typical of the grading patterns observed in actual grade data. To accommodate the distribution of grades actually observed, it is usually necessary to adjust the relative widths of the intervals associated with each grade on a class-by-class basis.

By estimating the grade-cutoffs separately for each class, several of the more controversial assumptions made by this team regarding the properties of undergraduate grades can be eliminated. For example, if grade-cutoffs are estimated individually for each class, it is not necessary to assume that one letter grade corresponds to one standard deviation in student achievement, or that professors do not grade on curves or compare performances of students within classes, or that professors uniformly adjust for course difficulty.

Other questionable model assumptions include the statements that

- students select courses randomly,
- students do not gravitate to courses that are graded leniently, and
- professors have accurate perceptions of student achievements.

None of these assumptions is required for the GRM; by liberalizing the interpretation of this model's parameters, they could be eliminated here as well.

Conclusion

Of the three models proposed, only the model proposed by the team from Harvey Mudd College seems to handle the two-student paradox mentioned above. Their model also attempts to combine information about the grading patterns of instructors across classes, which is an aspect of model fitting not normally included even in GRMs. The primary substantive disadvantage of this team's proposal is its complexity. It is clearly the most difficult model to explain.

In summary, all three teams proposed models that would improve the rankings of students within most undergraduate institutions. Importantly, each proposal would also reduce the incentives introduced by traditional GPA for students to enroll in "easy" classes, and would therefore improve the academic environment at colleges where they were applied. Of course, the greatest weakness of each proposal is that it is only a proposal! I encourage each team truly to make their models an *application* of mathematics by lobbying for the adoption of an adjusted GPA at their institution.

References

Caulkin, J., P. Larkey, and J. Wei, J. 1996. Adjusting GPA to reflect course difficulty. Working paper, Heinz School of Public Policy and Management, Carnegie Mellon University.

Goldman, R., D. Schmidt, B. Hewitt, and R. Fisher, R. 1974. Grading practices in different major fields. *American Education Research Journal* 11: 343–357.

Goldman, R., and Widawski, M. 1976. A within-subjects technique for comparing college grading standards: implications in the validity of the evaluation of college achievement. *Educational and Psychological Measurement* 36: 381–390.

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

Strenta, A., and R. Elliott. 1987. Differential grading standards revisited. *Journal of Educational Measurement* 24: 281–291.

About the Author

Valen E. Johnson is Associate Professor of Statistics and Decision Sciences at Duke University. His research interests include statistical image analysis, ordinal data modeling, and Markov Chain Monte Carlo simulation methods.