1996

Adjusting GPA to Reflect Course Difficulty

Jonathan P. Caulkins
Carnegie Mellon University

Patrick D. Larkey
Carnegie Mellon University

Jifa Wei
Carnegie Mellon University

Published In
.

V. 1.1

Adjusting GPA to Reflect Course Difficulty

Johnathan P. Caulkins Patrick D. Larkey Jifa Wei

The Heinz School of Public Policy and Management

Carnegie Mellon University

Abstract

The computation of Graduate Point Average (GPA) incorrectly assumes that grades are comparable across courses and instructors. GPA overstates the performance of students who elect an "easier" course of study relative to those who choose a more "difficult" course of study. This paper proposes a method of adjusting GPA and applies it to data from one cohort of undergraduates at Carnegie Mellon University. Adjusted GPAs are more highly correlated with students' high school Grade Point Average and with SAT scores than are the raw GPAs or GPAs adjusted using a prominent alternative method, Item Response Theory. A survey of students finds that the new methods' estimates of relative course difficulty are consistent with students' perceptions of relative course difficulty.

Acknowledgment

This research was supported in part by the National Science Foundation under Grant No. SES-9122244. We are grateful to Bill Elliott, Darleen Labarbera, John Papinchak, Richard Rainsberger, and Joy Sirney for their help in collecting the data and to Janusz Szczypula for his helpful comments on an earlier version of this paper. The opinions, findings, and conclusions in this paper are the responsibility of the authors and not NSF or anyone else.

Introduction

Grade point average (GPA) is the dominant measure of student quality in educational systems throughout the world. Students at all levels strive for a higher GPA because there are so many important consequences associated with it: parental and peer approval, admission to undergraduate and graduate programs, financial aid, graduation, graduation standing and honors, and access to selective employment interviews.

Unfortunately, GPA is an extremely flawed measure of student performance (Milton et al. 1988; Elliott & Strenta 1988; Chance 1988; Ramist et al. 1990; Young 1990; Larkey & Caulkins 1992) Like any scalar measure of a complex phenomenon like educational performance, GPA inevitably omits a lot of important information. GPA's weaknesses, however, exceed those inherent to scalar measures. GPA assumes incorrectly that all course grades mean essentially the same thing; for comparing students, GPA assumes that an A is an A, a B is a B. There is, however, substantial variation among majors, courses, and instructors in the rigor with which grades are assigned. Some courses give grades to students that are systematically lower or higher, than the grades obtained by those same students in other courses (Young 1990; Elliott & Strenta 1988; Strenta & Elliott 1987; Goldman & Widawski 1976; Goldman & Slaughter 1976; Goldman & Hewitt 1975; Goldman et al. 1974). The same GPA for different students can represent different levels of achievement and performance. A lower GPA for a student may not necessarily mean that the student performs less well than students who have higher GPAs; the student may simply be studying in fields and taking courses or instructors with more stringent grading standards.

Table 1[1] shows just how perverse unadjusted GPAs can be. Four students (A, B, C, and D) each take five classes of the nine available. Student A receives the highest score in every class she takes. Student D obtains the lowest score in every class he takes. Students B and C fall in between; student B is always better than student C when they take the same class. Given this information, we expect a measure that fairly represents students' performance to rank students in the following order:

[1] This table is taken from Larkey and Caulkins (1992),

A>B>C>D. Unfortunately, GPA, as conventionally computed, yields exactly the opposite order: D>C>B>A. This result is caused by the widely different grading standards among the nine classes in terms of class average grades.

[Table 1 goes about here]

The above example is clearly contrived but real examples of widely different grading standards with perverse implications for GPA are common. For example, 253 students at Carnegie Mellon University took both Calculus and Computing Skills Workshop classes in the fall of 1990. The mean grade of these students was 1.87 in Calculus and 3.45 in the Computing Skills Workshop on a 4.0 scale. This difference in standards is not a problem for comparing the 253 students who took both courses but it is for comparing students who took Calculus but not the Computing Skills Workshop with students who took the Computing Skills Workshop but not Calculus.

The problem that different grading standards poses for GPA as a student performance measure is venerable (Meyer 1909). The grading standards problem can be solved by (1) getting the graders to adhere to the same standard or (2) correcting for differences in grading standards after the grades have been assigned. The most frequently used method for attempting to get graders to adhere to a common standard is to require graders to use a common "curve" in which grades are assigned in fixed proportions (Aiken 1963). Curving introduces a new potential bias in GPA, bias against classes consisting of above average students in ability or performance. The assignment of students to classes is nonrandom. There is overwhelming evidence that student abilities are not evenly distributed across courses and departments. Different fields have different "curves." For example, Sabot and Wakeman-Linn (1991) found that the mean grade in Mathematics courses was 2.53 out of 4.0, 0.59 lower than that for English courses in a seven college study. Low ability students tend to gravitate toward fields with lax grading standards (Willingham 1985; Goldman & Hewitt 1975). The variety in fields, subjects and faculty personalities probably makes coherent, enforceable grading standards unachievable; seeking evaluative consistency among diverse faculty is like "herding squirrels."

A second and perhaps more promising approach to reducing the bias in GPA is to permit the assignment of grades on a largely unregulated basis by individual faculty and then adjust the grades for

construction of a corrected GPA. Researchers have proposed many methods to correct GPA so that it reflects the differential difficulty of grade accomplishments (Linn (1966) and Young (1993)). Two of the more visible proposals have been a pairwise comparison method (Elliott & Strenta 1988; Goldman & Widawski 1976) and an Item Response Theory based method (Young 1990 & 1989). No grade correction method has been used much in practice either because they had no significant effect or because they were difficult to implement (Young 1993).

No adjustment can make GPA a perfect measure of students' performance because of inherent deficiencies in the grades themselves. More important, student quality, performance, and aptitude can never be captured adequately by any scalar measure. Nonetheless, the impossibility of achieving a perfect scalar measure of student performance is hardly an excuse for continuing to live with the correctable flaws in unadjusted GPAs and the improper student rankings and course selection incentives that result from these flaws. We propose alternative methods for adjusting GPAs to reflect variation in the stringency of grading standards across courses and apply them to data on a cohort of undergraduates at Carnegie Mellon University.

Alternative Methods for Adjusting GPA

This section examines five methods for adjusting GPAs to reflect variation in grading standards. Two methods are naive in that they ignore differences in the quality of students across classes. These are foils against which the other methods can be compared.

Deviation method (Naive method I)

The simplest adjustment is to look at individual course grades relative to the average grades awarded in that course by subtracting the mean grade attained in the course from the raw grades. The adjusted GPA would be the average of these adjusted grades. More formally, the adjusted GPA, q_i, is:

$$q_i = \frac{\sum_j \left(r_{ij} - \frac{\sum_i r_{ij}}{N_s^j} \right)}{N_c^i}; \qquad for\ i = 1, 2, 3, \ldots N; \qquad \textbf{(1)}$$

where, N_c^i denotes the number of courses student i takes, N_s^j denotes the number of students in course j, and r_{ij} is student i's grade in course j.

Standardization method (Naive method II)

The deviation method described above may not be satisfactory if the variance in grades is different across courses. This potential problem can be addressed by dividing the grade deviation by the standard deviation of grades in the course. Formally,

$$q_i = \frac{\sum_j \left[\left(r_{ij} - \frac{\sum_i r_{ij}}{N_s^j} \right) / STD_j \right]}{N_c^i}; \qquad for\ i = 1, 2, 3, \ldots N; \qquad \textbf{(2)}$$

where, STD_j is the standard deviation of grades in course j.

The need to adjust for student quality

The naive adjustment methods fail to account for variation across courses in students' ability. Suppose two students received grades equal to the class mean in every course taken and that one of them took classes exclusively with the strongest students in the university while the other took classes with the weakest students in the university. The naive methods would assign them the same adjusted GPA even though the mean grade is tougher to achieve against the strongest than against the weakest.

There are methods that can account for both students' grades and students' ability in courses. Larkey (1991, 1992) developed methods for fairly comparing professional golfers where they had performed in circumstances of varying difficulty. For example, the Professional Golf Association of America awards the prestigious Vardon Trophy each year to the professional golfer with the lowest scoring average and must choose between golfers who have played under scoring conditions that vary in difficulty. Just as easy courses raise students' GPAs, easy golf courses lower the professionals' scoring average. Caulkins et al. (1993) found the same difficulty in comparing airlines' in terms of their on-time performance. Some airports are much more difficult to arrive in on-time than others and airlines that fly more frequently into "difficult" airports are disadvantaged in terms of average on-time performance.

The generic problem for rating golfers, airlines, students, and many other performers is: ***Given N players who compete and receive scores in varying subsets of M tournaments, and given no***

exogenous information about the difficulty of those tournaments or the quality of the players, rate each player's performance relative to that of the other N-1 players (Caulkins et al. 1993)**.** For ranking students in educational environments, students are players and courses are tournaments with students competing for higher grades and GPA.

We derive three methods that adjust students' GPA to reflect course difficulty. Consider a general grade model given by Equation (3). A student's grade in a specific course is a function of the student's attributes and the course's attributes.

$$r_{ij} = f(q_i, b_j) + e_{ij} \qquad \textbf{(3)}$$

Where r_{ij} is student i's grade in course j,

$\mathbf{q_i}$ is a vector of attributes for student i,

$\mathbf{b_j}$ is a vector of attributes for course j, and

ε_{ij} is a random error term.

This function could be very complex since many factors affect students' course grades. For example, student i's attribute vector $\mathbf{q_i}$ could include the student's aptitude in specific areas and how hard this student works. A course's attribute vector could include the subject of the course, the instructor of the course, and the abilities of other students in the course.

Here we consider only the very special case in which the attribute vectors contain only one or two elements. The first method is an additive method in which one scalar is employed to represent a student attribute that is called adjusted GPA and another scalar is used to represent a course attribute that is called the grading standard index. A student's observed grade in a course is the difference between the student's adjusted GPA and the course's grading standard index plus an error term (Equation 4 below). Previous studies (Strenta & Elliott 1987, Goldman & Widawski 1976) have found that the index of grading standards for different departments is nearly additive in predicting the outcomes of comparisons of grading standards in pairs of departments.

For the second method, instead of using the difference between the adjusted GPA and the course difficulty index, we use a ratio of the student's adjusted GPA over the course difficulty index plus an error term to represent the student's observed grade. The third method combines the two

methods and uses one scalar to describe the course's discrimination power and another to describe the course's grading stringency.

The goal is to find parameters for these simple models that minimize the difference between students' observed course grades and the grades projected by these models. The estimated parameters do not necessarily fall between 0 and 4, the conventional range for the GPA measure, but the parameters corresponding to GPA can be linearly scaled to 0-4 or any other interval.

The additive method.

The additive method assumes that student i has an ability to earn an average grade, q_i. Thus absent any information about course difficulty, one would predict that student i's grade would be q_i. Similarly, courses are assumed to differ in grading stringency by shifting all students' grades up or down by a constant β_i. Thus one would predict that student i's grade in course j would be the difference between the student's q_i and the course's constant β_j. q_i is called student i's adjusted GPA, and β_j is called course j's grading standard index here and for the balance of this paper.

$$r_{ij} = q_i - b_j + e_{ij}\ ;\ for\ i = 1, 2, 3,\ldots N;\ j = 1, 2, 3,\ldots M; \qquad \mathbf{(4)}$$

This method predicts that students with any specified ability will receive higher grades from courses with laxer grading standards (small βj) and lower grades from courses with more stringent grading standards.

Rating students' performance and courses' difficulty can be viewed as a parameter estimation problem. One way to estimate the N q_is and M β_js is to use the least squares method, that is, to solve the following minimization problem:

$$\text{Min} \sum_i \sum_j [\ r_{ij} - (\ q_i - b_j\ )\ ]^2 \qquad \mathbf{(5)}$$

This yields the following first order conditions (FOC):

$$q_i = \frac{\sum_j (\ r_{ij} + b_j\ )}{N_c^i}; \qquad for\ i = 1, 2, 3,\ldots N; \quad \mathbf{(6)}$$

$$b_j = \frac{\sum_i (\ q_i - r_{ij}\ )}{N_s^j}; \qquad for\ j = 1, 2, 3,\ldots M; \quad \mathbf{(7)}$$

Notice that the above summations should be interpreted as summing only over those i and j for which a grade exists.

The additive method has an appealing interpretation. Student i's adjusted GPA is an average of the adjusted course grades, $(r_{ij}+\beta_j)$, that student i gets from all courses he or she takes. Equation 6 represents this calculation. Student i's grade in course j is increased by adding course j's difficulty index b_j in this calculation. A course's difficulty index is a relative grading standard of this course versus all other courses and can be estimated. If, for example, most students in a course get lower grades than their GPAs then this course is considered to have a stricter grading standard and, hence, a higher difficulty index. A superior approach is to substitute the adjusted GPA, $q_{i,}$ for students' GPA because it better reflects student performance. Equation (7) calculation calculates the difficulty index using q_i.

Equations (6) and (7) are a system of N+M linear equations that can be easily solved.

Multiplicative method

A parallel approach is multiplicative in the sense that it projects students' GPA in a ratio format:

$$r_{ij} = q_i / a_j + e_{ij}\ ;\ for\ i = 1, 2, 3,\ldots N;\ j = 1, 2, 3,\ldots M; \qquad \textbf{(8)}$$

where parameter α_j is a course difficulty index. The bigger the α_j (>1), the more difficult the course grading standard.

The minimization problem for this method is:

$$\text{Min} \sum_i \sum_j [\ r_{ij} - \frac{q_i}{a_j}\ ]^2, \qquad \textbf{(9)}$$

which yields first order conditions:

$$q_i = \frac{\sum_j (\ r_{ij} / a_j\ )}{\sum_j 1 / a_j^2}\ ;\ for\ i = 1, 2, 3,\ldots N, \quad \text{and} \qquad \textbf{(10)}$$

$$a_j = \frac{\sum_i q_i^2}{\sum_i (\ r_{ij} * q_i\ )}\ ;\ for\ j = 1, 2, 3,\ldots M. \qquad \textbf{(11)}$$

A simple iterative procedure can solve these two sets of equations.

The combined method

The additive and multiplicative approaches fail to consider that courses not only have different grading stringency but also different discrimination power. Some courses give better students grades that are substantially higher than those obtained by weaker students. Other courses give all students nearly the same grade and are not very useful in differentiating between good performers and bad performers. A combined method that uses two parameters to describe course grading partly addresses this problem.

The formula for this method is:

$$r_{ij} = q_i / a_j - b_j + e_{ij}\ ;\ for\ i = 1, 2, 3,\ldots N;\ j = 1, 2, 3,\ldots M; \qquad \textbf{(12)}$$

$$\text{Min} \sum_i \sum_j [\, r_{ij} - (\frac{q_i}{a_j} - b_j) \,]^2; \qquad \textbf{(13)}$$

and the first order conditions are:

$$b_j = \frac{\sum_i (q_i / a_j - r_{ij})}{N_s^j}; \; for\; j = 1, 2, 3, \ldots M, \text{ and} \quad \textbf{(14)}$$

$$a_j = \frac{\sum_i q_i^2}{\sum_i (r_{ij} + b_j) * q_i}; \; for\; j = 1, 2, 3, \ldots M. \qquad \textbf{(15)}$$

$$q_i = \frac{\sum_j (r_{ij} + b_i) * (1 / a_j)}{\sum_j 1 / a_j^2}; \; for\; i = 1, 2, 3, \ldots N; \quad \textbf{(16)}$$

Again, an iterative procedure can be employed to solve these equations.

Each of the three new methods give solutions which exist and are unique up to a scaling factor (Caulkins et al. 1993). All three are equivalent to the conventional GPA calculation if all courses are equally difficult. For example, with the additive method, if b1=b2=b3=...=bM=0, then equation (6) reduces to the formula for calculating the conventional GPA.

Evaluating the New Methods

It is easy to propose methods for adjusting GPA; it is harder to determine if the proposals have any merit. We consider three criteria. First, an adjustment method should generate an accurate measure of students' performance. There is, of course, no observable and absolute measure of performance—if there were then the whole adjustment exercise would be irrelevant—and so we must evaluate accuracy indirectly. One approach in the literature (Elliott & Strenta 1988; Strenta & Elliott 1987) is to see if students' adjusted freshman year GPA is more highly correlated with their high school GPAs and SAT scores than is their unadjusted GPA. We also examine correlations between students'

adjusted GPAs across different years in college on the assumption that student ability and performance is fairly stable over time so "better" measures will give higher correlations.

Second, we can see if the new methods produce reasonable course difficulty indices. Again, there is no absolute standard; but we can ask students to rate courses in terms of grading stringency and compare the survey results with the estimated course difficulty indices.

Finally, how do the new methods compare with existing methods? The Item Response Theory (IRT) based method developed by Young (1989 and 1990) is the most theoretically defensible and well-developed extant adjustment method so we compare the new methods to it.

Empirical Results

The registrar and enrollment offices at Carnegie Mellon University (CMU) provided data for 1197 students who entered CMU in Fall 1990. These data include gender, ethnicity, high school grade point average (HGPA), SAT mathematical score (SAT_M), SAT verbal score (SAT_V), various achievement test scores, grades on courses they took in their first three years in CMU, and college of enrollment. Students were enrolled in four colleges: College of Fine Arts (CFA), Carnegie Institute of Technology (CIT), College of Humanities and Social Science (HSS), and Mellon College of Science (MCS).

We only included courses that give a letter grade because it is not clear how to compare letter grades with PASS/FAIL grades and PASS/FAIL courses are not currently included in calculating students' GPA. We converted the letter grades to numerical values for this study: A=4, B=3, C=2, D=1, and R=0. (R means failure at CMU.) All the other letter grades, such as I (incomplete) and W (withdraw), were coded as missing values; this is equivalent to assuming that the students receiving those grades did not attend those classes. Where the same course was offered in different semesters, they were considered to be different courses since they had different students, and may have been taught by different professors. Different sections of one course taught in one semester were considered to be one course since the data used do not provide course section information.

Students in the College of Fine Arts (CFA) were excluded primarily because there is insufficient reciprocal course-taking between CFA and the rest of CMU to have a reliable basis for comparing

students (Wei 1994). Courses that had an enrollment of fewer than five students from this cohort were excluded. The methods per se do not require a minimum number of students per course, but courses with less than five students may give unreliable course difficulty indices. We calculated students' unadjusted GPA in a slightly different way than is employed by the registrar's office. Instead of weighting course grades by course credits and dividing the product of these two by the total credits, the unadjusted GPA is simply the average of course grades that students obtained; the correlation between the two is very high.

Effects of the new methods for all students

All five adjusted GPAs have higher correlations (in absolute value) with the pre-admission variables (HGPA, SAT_M, SAT_V, and students high school rank) than does the unadjusted GPA (See Table 2). All of the correlation differences between the adjusted GPAs from the three new methods and the raw GPA are statistically significant at the 0.01 level using the Z1* statistic (Steiger 1980). The adjusted GPAs from the combined method have the highest correlation with students' high school rank (HRANK), SAT_M, and SAT_V among all six GPAs. The adjusted GPAs from the additive method have the highest correlation with students' high school grade point average. The naive adjustment methods, deviation and standardization, improve on unadjusted GPA but have lower correlations with pre-admission variables than the three more sophisticated adjustment methods.

[Table 2 goes here.]

Of the three traditional predictors, HGPA has the highest correlation with students' freshman GPA. This result is consistent with previous studies (Willingham et al. 1990; Linn 1966). Note that SAT_M is more highly correlated with students' freshman year GPA than is the SAT_V. This may be peculiar to Carnegie Mellon University where courses in many disciplines are quantitative.

For the balance of the paper, we will mainly report results from the additive method because it is simple, has an appealing interpretation, and produces GPA results very highly correlated ($r = 0.993$) with the GPAs from both the multiplicative and combined methods.

Tables 3 and 4 report course difficulty indices from the additive method by college and department. Table 4 only includes those departments that offered more than five courses to this cohort

of students. The CMU courses, which are courses that do not belong to any college (e.g. fundamental computing skills), have the least stringent grading standards according to the course difficulty indices from the additive method. The Mellon College of Science (MCS) courses have the most stringent grading standards. These results are consistent with previous studies (Elliott and Strenta (1988) and Ramist et al. (1990)).

[Tables 3 and 4 go about here.]

Students were asked their perceptions of course grading standards. A questionnaire that included 20 courses with enrollment of at least 60 students was sent through electronic mail to the 930 of the 1197 students who were still enrolled in CMU in their senior year. Students were asked to rate the 20 courses in terms of the difficulty of getting higher grades on a 1-7 scale with 1 indicating it was very easy to get a higher grade in that course. Students were asked to rate only those courses about which they had some knowledge, either from taking the course or by reputation, and to indicate whether they had taken the courses themselves. We got 122 usable responses, a 13% response rate. A course's surveyed difficulty index is the average of all responses for that course. The correlation between the survey results and course difficulty indices from the additive method is 0.845, suggesting that the course difficulty indices produced by the additive method are consistent with students' perceptions about course grading standards.

Results of adjusted GPA on predictive validity in college admission.

Adjusted GPA's can be predicted substantially better from pre-admission variables than can conventional GPA (see Table 5). Substantially more variance is explained by the three predictors of college grade performance when the dependent variables are adjusted GPAs instead of unadjusted GPAs; the improvement in variance explained with the combined method is 0.132, a 102 percent increase.

[Table 5 goes about here.]

The adjusted students' GPAs are more consistent than the unadjusted GPA overtime during college. Table 6 shows that the correlations of adjusted GPAs between years are higher than those of the unadjusted GPA. The yearly adjusted GPAs are statistically significantly (at the 0.01 level) more highly correlated than are the unadjusted GPAs, using Z2* statistic for testing two dependent correlations described by Steiger (1980).

[Table 6 goes about here.]

Results of the adjusted GPA on gender bias.

Previous studies (McCornack & McLeod 1988; Young 1992) have shown that using a single equation to predict GPA for both men and women often leads to systematic under-prediction of women's GPA. This is also true for this cohort of students. Table 7 shows the descriptive statistics of the three pre-admission variables and their freshman year raw GPA and adjusted GPA by gender. To compare the adjusted GPA with the raw GPA, we transformed the adjusted GPAs linearly to have the same mean and standard deviation as the raw GPA.

[Table 7 goes about here.]

All three pre-admission variables of the female students are lower on average than those of the male students. However, the female students have higher freshman year GPA than the male students. Using one equation to predict students' freshman year GPAs for both female and male students would result in a significant underprediction for female students particularly since the equation would be largely determined by data from the larger number of male students.

Using the adjusted students' GPA as a criterion reduces the underprediction since male students have higher adjusted GPAs than do the female students. The gender difference in adjusted GPAs for this cohort students can be understood better by examining patterns of course selection. The mean proportion of female students in social science courses (0.43) is higher than in natural science (0.23) and engineering (0.12) courses. Since the grading standards in the natural science and engineering courses were found to be generally more stringent than those of the social science courses, the additive method, on average, increases grades for natural science and engineering courses and depresses grades for social science courses. This results in the relatively lower adjusted GPAs for female students.

We also examined the correction effects on various ethnic groups, e.g., black, asian, and white. There were no clear differential effects of the correction along ethnic lines in this sample.

Comparison between the additive and Item Response Theory based methods.

To compare the new methods with the Item Response Theory based method, we need to apply all methods to the same data. The Item Response Theory based method separates courses into categories and applies the method to each category separately. Young divided his course data into three categories: natural science and engineering courses, social science courses, and humanities courses. He found that the Item Response Theory based method performs best on the natural science and engineering courses in the sense that the IRT-based GPA is more highly correlated with the pre-admission variables than the raw GPA (Young 1989). Here, we compare the new methods with the Item Response Theory based method on the natural science and engineering courses offered by CIT, MCS, and the Department of Statistics in HSS.

A PARSCALE[2] program, used by Young in his study, was employed to estimate parameters for the Item Response Theory based method. Two adjusted GPAs, denoted Add-GPA and IRT-based GPA, were generated in addition to the conventional GPA. Table 8 reports multiple regression results with the three pre-admission variables as independent variables and three GPAs as dependent variables.

[2]. PARSCALE[2] is available through Scientific Software, Inc., Chicago, Il.

[Table 8 goes about here.]

The amount of variance in the IRT-based GPA which is explained by the three pre-admission variables is greater than the amount of variance in unadjusted GPA similarly explained (R^2: 0.264 vs. 0.230), but the amount of variance in the Add-GPA is larger still (R^2: 0.321). Note that the SAT_V is not statistically significant in any of the regression equations; this may be because we only included natural science and engineering courses.

Discussion

Conventional GPA is a flawed measure of students' performance and ability. Any scalar measure will have serious shortcomings because performance and ability are not unidimensional. There are, however, persistent demands for such a summary measure. It is important to find a better scalar measure than conventional GPA. This paper proposed three alternatives for GPA. All three are similar in concept, calculation, and outcome. They compute the adjusted GPAs as the least squares parameter estimates for elementary models for how students and course attributes combine to yield grades.

Since there is no definitive, observable standard of student quality, it is impossible to determine absolutely whether these adjusted GPAs are "better" than conventional GPA. However, the various pieces of circumstantial evidence suggest that the adjusted GPAs are both reasonable and probably reflect student ability more accurately than does the unadjusted GPA. In particular, the adjusted GPAs were shown to be more highly correlated with pre-admission variables and more highly correlated across years within college. The adjusted GPAs were also more highly correlated with pre-admission variables than were either unadjusted GPAs or GPA adjusted by an Item Response Theory based method. Furthermore, the course grading difficulty indices produced by the method were consistent with students' perceptions. Finally, there is a certain intuitive appeal to the new methods, particularly the additive method. It makes sense that grades be adjusted to reflect variation across courses in grading stringency and that a course grading stringency should be assayed by considering the quality of the students in that class as well as the grades given.

The replacement of GPA with a better measure must be a gradual process. GPA has been used by many people for a long time and we need to know more about both the effects of GPA and the effects of the alternatives before advocating wholesale change. The problems with conventional GPA are, however, sufficiently obvious that schools should consider producing both unadjusted and adjusted GPAs. More research is essential; the research should: (1) replicate these results in other contexts, (2) refine the methods proposed here (e.g., by developing ways of systematically determining when grades from a college such as CFA should and should not be analyzed separately from grades from other colleges), and (3) search for still better adjustment procedures.

This research is important because improving GPA as a measure of students ability and performance could bring significant benefits both at the individual and societal level. At the individual level, injustices such as those depicted in Table 1 might be mitigated and/or avoided (all of the new methods rank the four students in the desired order: A>B>C>D). Adjusting grades for grading stringency should reduce the incentive for students to prefer courses and instructors with easy grading standards and lead them to base course choice more on factors such as the value of what they expect to learn. Larkey and Caulkins (1992) argue that since natural science and engineering courses generally have more stringent grading standards, adjusting GPA to reflect grading difficulty might remove a disincentive to major in those fields. In all fields it could alleviate pressures toward grade inflation.

It is worth observing that the proposed adjustment methods are very simple and, hence, quite flexible. For example, one could constrain the extent to which a course's grades are adjusted simply by adding the appropriate constraint to the least squares optimization problem solved to estimate the student and course parameters.

The methods can also be applied in different contexts, such as incorporating Achievement Test scores into college admission decisions. Many students take Achievement Tests, but this information is typically used less systematically than high school GPA or SAT scores in part because not every students takes every Achievement Test. Wei (1994) applied the methods developed here to the SAT math, Biology, Chemistry, Physics, Math Level I, and Math Level II achievement tests to estimate a measure of mathematics aptitude which is broader than SAT math test alone. Similarly, the SAT verbal,

English Composition, English Composition with Essay, and American History and Social Studies scores were combined to yield a broader measure of verbal skills. These broader measures were more highly correlated with undergraduate grades than were the SAT scores alone.

Also, the measures of course grading stringency can be useful in and of itself. For example, Wei (1994) used them to assess the impact of factors such as college, class size, and homogeneity of students on grading stringency.

Finally, the adjusted GPA can be implemented centrally; it is not necessary to ask faculty to change their grading behavior in any way.

References

Aiken, L. R. 1963 "The Grading Behavior of A College Faculty," Educational and Psychological Measurement, Vol. XIII, No. 2.

Bookhart, S. M. 1993 "Teachers' Grading Practices: Meaning and Value," Journal of Educational Measurement, Vol. 30, No. 2, pp. 123-142.

Braun, H. I. & Szatrowski, T. H. 1984a "The scale-linkage algorithm: Construction of a universal criterion scale for families of institutions," *Journal of Educational Statistics*, 4, pp. 311-330.

Braun, H. I. & Szatrowski, T. H. 1984b "Validity studies based on a universal criterion scale," *Journal of Educational Statistics*, 4, pp. 331-344.

Caulkins, J. P., Barnett, A., Larkey, P. D., Yuan, Y. & Goranson J. 1993 "The On-Time Machines: Some Analysis of Airline Punctuality," *Operations Research*, Vol 41, No. 4, pp. 710-720.

Chance, P. 1988 "Testing Education," *Psychology Today*, May, pp. 20-21.

Elliott, R. and Strenta, A. C. 1988 "Effects of Improving the Reliability of the GPA on Prediction Generally and on Comparative Predictions for Gender and Race Particularly," *Journal of Educational Measurement*, Vol. 25, No. 4, pp. 333-347.

George, K. D. and Joll, C. 1981 Industrial Organizations: Competition, Growth and Structural Change, 3rd Edition, London: George Allen & Unwin (publisher), Ltd.

Goldman R. D., Schmidt, D. E., Hewitt, B. N. and Fisher R. 1974 "Grading Practices in Different Major Fields," *American Educational Research Journal*, Fall 1974, Vol. 11, No. 4, pp. 343-357.

Goldman, R. D. and Hewitt, B. N. 1975 "Adaption-Level as An Explanation for Differential Standards in College Grading," *Journal of Educational Measurement*, Vol. 12, No. 3.

Goldman, R. D. and Slaughter, R. E. 1976 "Why College Grade Point Average is Difficult to Predict," *Journal of Educational Psychology*, Vol. 68, No. 1, pp. 9-14.

Goldman, R. D. and Widawski, M. H. 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, pp. 381-390.

Goldman, R. D. and Hewitt, B. N. 1976 "The Scholastic Aptitude Test 'Explains' Why College Men Major in Science More often Than College Women," *Journal of Counseling Psychology*, Vol. 23, No. 1, pp. 50-54.

Hambleton, R. K. and Swaminathan, H. 1985 **Item response theory: Principles and applications**. Boston, MA: Kluwer Academic Publisher.

Helson, H. 1947 "Adaptation level as frame of reference for prediction of psycho-physical data," *American Journal of Psychology*", Vol. 60, pp. 1-29

Hills J. R. 1964 "The Effect of Admissions Policy On College Grading Standards," *Journal of Educational Measurement*, Vol. 1, No. 2

Hills J. R. and Gladney, M. B. 1968 "Factors Influencing College Grading Standards," *Journal of Educational Measurement*, Vol. 5.

Larkey, P.D. 1991. "A Better Way to Find the Top Scorer." *Golf World*, January 25, pp. 72-74.

Larkey, P. D. and Caulkins, J. P. 1992 "Incentives to Fail," Working paper 92-51, Heinz School, Carnegie Mellon University.

Linn, R. L. 1966 "Grade Adjustments for Prediction of Academic Performance: A Review," *Journal of Educational Measurements*, Vol. 3, No. 4, pp. 313-329.

Lord, F. M. and Novick, M. R. 1968 **Statistical theories of mental test scores**. Reading, MA: Addison-Wesley.

Lord, F. M. 1980 **Applications of item response theory to practical testing problems**. Hillsdale, NJ: Lawrence Erlbaum.

McCornack R. l. and McLeod, M. M. 1988 "Gender Bias in the Prediction of College Course Performance," *Journal of Educational Measurement*, Vol. 25, No. 4, pp. 321-331.

Meyer, M. 1909 "The Grading of Students," *Science*, Vol. XXVII. No. 712.

Milton, O., Pollio, H. R. and Elison, J. A. 1988 "GPA Tyranny," *The Education Digest*, December.

Mislevy R. J., Beaton, A. E., Kaplan, B. and Sheehan, K. M. 1992 "Estimating Population Characteristics From Sparse Matrix Samples of Item Responses," *Journal of Educational Measurement*, Vol. 29, No. 2, pp. 133-161.

Ramist, L., Lewis, C., and McCamley, L. 1990 "Implications of using freshman GPA as the criterion for the predictive validity of the SAT," In W. W. Willingham. C. Lewis, R. Morgan, & L. Ramist (Eds.), **Predicting college grades: An analysis of institutional trends over two decades** (pp. 253-288). Princeton, NJ: Educational Testing Service.

Steiger, J. H. 1980 "Tests for Comparing Elements of a Correlation Matrix, " Psychological Bulletin, Vol. 87, No. 2, pp. 245-251.

Sabot, R. and Wakeman-Linn, J. 1991 "Grade Inflation and Course Choice," *Journal of Economic Perspectives*, Vol. 5, No. 1.

Strenta, A. C. and Elliott, R. 1987 "Differential Grading Standards Revisited," *Journal of Educational Measurement*, 1987, Vol. 24, No. 4, pp. 281-291.

Willingham, W. W., Lewis, C., Morgan,R. and Ramist, L. 1990 **Predicting College Grades: An Analysis of Institutional Trends Over Two Decades**. Princeton, NJ: Educational Testing Service.

Willingham, W. W. 1985 **Success in college**. New York: College Entrance Examination Board.

Young, J, W. 1989 **Developing a universal scale for grades: Investigating predictive validity in college admissions**. Ph.D. dissertation. Stanford University.

Young, J. W. 1990 "Adjusting the Cumulative GPA Using Item Response Theory," Journal of Educational Measurement, Vol. 27, No. 2, pp. 175-186.

Young, J. W. 1990 "Are Validity Coefficients Understated Due to Correctable Defects in the GPA?" Research in Higher Education, Vol. 31, No. 4. pp. 319-325.

Young, J. W. 1991 "Gender Bias in Predicting College Academic Performance: A New Approach Using Item Response Theory," Journal of Educational Measurement, Vol. 28, No.1, pp. 37-47.

Young, J. W. 1993 "Grade Adjustment Methods," Review of Educational Research, Vol. 63, No. 2, pp. 151-165.

Table 1. Example

	Student A	Student B	Student C	Student D	Class Avg.
Class #1	93			90	91.50
Class #2	85		80		82.50
Class #3			100	95	97.50
Class #4	80	75			77.50
Class #5		97		95	96.00
Class #6	93			85	89.00
Class #7		92	89		90.50
Class #8	92	91	90	88	90.25
Class #9		91	89		90.00
GPA	88.60	89.20	89.60	90.60	

Table 2. Correlations Between Pre-admission Variables and Various Criterion Variables

	GPA	Dev-GPA	Std-GPA	Mul-GPA	Add-GPA	Com-GPA
HGPA (926)	0.348	0.386	0.387	0.419	**0.434**	0.432
HRANK(920)	-0.311	-0.368	-0.367	-0.395	-0.414	**-0.425**
SAT – V (1074)	.0104	0.122	0.132	0.138	0.150	**0.150**
SAT-M(1074)	.0179	0.254	0.250	0.311	0.351	**0.398**

Note: HGPA denotes students high school GPA; HRANK denotes students high school rank in their senior class and we standardize the HRANK between 1 to 100 with 1 indicating the best students. SAT-V denotes the SAT verbal score and SAT_M denotes the SAT quantitative score. The number in the parentheses denotes the number of observations used in the calculation. Bold number means that it is the biggest (in absolute value) in a row.

Table 3. Courses Difficulty Index by Colleges

	GPA		Mul-GPA		Add-GPA		Com-GPA	
	Estimate	t-Value	Estimate	t-Value	Estimate	t-Value	Estimate	t-Value
Intercept	0.161	0.656	-0.711	-3.028	-0.963	-4.148	-1.139	-4.879
HGPA	0.544	9.377	0.600	10.841	0.608	11.098	0.584	10.605
SAT_M	0.127	2.070	0.349	5.976	0.422	7.309	0.529	9.123
SAT_V	0.118	2.044	0.102	1.857	0.097	1.787	0.062	1.129
R2	0.130		0.215		0.243		0.262	

Note: The SAT_M and SAT_V in these regressions are rescaled by dividing the original values by 200.

Table 4. Courses Difficulty Index by Departments

	Freshman Year GPA	Sophomore Year GPA	Junior Year GPA
Freshman Year GPA	--	0.628	0.485
Sophomore ear GPA	0.683	--	0.565
Junior Year GPA	0.560	0.625	--

Note: Correlations above the main diagonal are for raw grade point average; correlations below the main diagonal are for the adjusted GPA from additive method.

Table 5. Multiple Regressioin Results fro Various GPAs

	Female Students			Male Students			Difference between Female and Male students
	N	Mean	STD	N	Mean	STD	
SAT-M	313	602	94	767	673	76	-71.000
SAT-V	313	548	87	767	556	70	-8.000
SAT	313	1150	150	767	1229	135	-79.000
HGPA	266	3.525	0.442	664	3.570	0.432	-0.045
GPA	340	2.800	0.832	835	2.793	0.812	0.007
Add-GPA	340	2.674	0.817	835	2.845	0.813	-0.171

Table 6. Correlations of Yearly Calculated GPA

Colleges	No. of Courses	Course Average Grade	Course Difficulty Index
Mellon College of Science (MCS)	48	2.691	0.174
Carnegie Institute of Technology (CIT)	10	2.995	0.083
College of Humanities and Social Science (HSS)	58	2.855	-0.141
Carnegie Mellon University (CMU)	11	3.623	-0.681

Table 7. Descriptive Statistics of Pre-admission Variables and Students' Freshman Year GPA by Gender

Departments	No. of Courses	Course Average Grade	Course Difficulty Index
Mathematics	19	2.565	0.313
Physics	13	2.686	0.172
Chemistry	7	2.832	0.095
Psychology	7	2.779	0.002
Philosophy	6	2.807	-0.064
History	16	2.791	-0.165
English	8	2.914	-0.209
Computer Science	7	3.037	-0.231
Modern Languages	12	2.987	-0.343

Table 8. Comparison Between the Item Response Theory Based Adjustment Method and the New Methods

	Dependent Variable (N=784)					
	GPA		Irt-GPA		Add-GPA	
	Estimate	T Value	Estimate	T Value	Estimate	t-Value
Intercept	-3.083	-8.043	-8.620	-22.752	-4.095	-11.258
HGPA	0.740	9.310	0.773	9.839	0.772	10.227
SAT_M	0.004	9.207	0.005	10.091	0.006	13.060
SAT-V	0.000	0.684	0.001	1.665	0.000	0.348
R2	0.230		0.264		0.321	