

Average Rank and Adjusted Rank Are Better Measures of College Student Success than GPA

Donald Wittman,  University of California, Santa Cruz

Abstract: *I show that there are better measures of student college performance than grade point average (GPA) by undertaking a fine-grained empirical investigation of grading within a large public university. The value of using GPA as a measure of comparative performance is undermined by academically weaker students taking courses where the grading is more generous. In fact, college courses composed of weaker performing students (whether measured by their relative performance in other classes, SAT scores, or high school GPA) have higher average grades. To partially correct for idiosyncratic grading across classes, alternative measures, student class rank and the student's average class rank, are introduced. In comparison to a student's lower-division grade, the student's lower-division rank is a better predictor of the student's grade in the upper-division course. Course rank and course grade are adjusted to account for different levels of academic competitiveness across courses (more precisely, student fixed-effects are derived). SAT scores and high school GPA are then used to predict college performance. Higher explained variation (R^2) is obtained when the dependent variable is average class rank rather than GPA. Still higher explained variation occurs when the dependent variable is adjusted rank.*

Keywords: GPA, rank, SAT, predicting college performance, student fixed-effects, course fixed-effects

Introduction

Grade point average (GPA) is a ubiquitous measure of college performance. In guiding their admissions, a majority of those colleges predicting student outcomes use first-year college GPA as their measure of student success (National Association for College Admission Counseling, 2016). In addition, colleges use GPA in determining whether a student is under the threat of dismissal. Typically, a student who has less than a 2.0 on a 4-point scale for the previous quarter and/or cumulatively is put on academic probation (Leon et al., 2019). GPA is also an important variable in research. In Google Scholar, there were 49,500 references generated from the following search terms: +regression +coefficients +university +first +year +GPA. Even if this exaggerates the number of research papers by one or even two orders of magnitude, one is still left with a very large number. And GPA is an important factor for those hiring. For college graduates in the class of 2019, 73% of surveyed employers intended to screen job candidates by GPA, with 3.0 being the cutoff (National Association of Colleges & Employers, 2018).

Unfortunately, as will be seen, college GPA (henceforth referred to as GPA without a modifying adjective to distinguish it from HSGPA, high school GPA) is an inaccurate and biased measure of student success. To understand why, it is useful to first consider a situation where first-year GPA works—the United States Air Force Academy. There students are randomly assigned to *classes* within the same set of core *courses*

and the common grading scheme across classes within a course is rigidly adhered to (Carrell & West, 2010). However, for most universities, using GPA is problematic because (1) it does not account for the different grading policies across courses and majors and (2) more significantly, it does not account for the differential *selection by students* across courses and majors. To illustrate point 1, if students did choose courses randomly, differing grading policies across courses would increase the variance in any estimated relationship between a set of independent variables and GPA, but no bias would be present. Point 2 is more serious as it creates biases in estimation. For example, if academically weaker students, as measured by their SAT scores, are more likely to *select* college courses where the grading is easier, then regressing GPA against SAT scores would downward bias the coefficient of SAT scores. This is not just a hypothetical possibility. There has been considerable research on the role of harsher grading in science, technology, engineering and math courses that encourages weaker students to either drop out of STEM majors or not consider a STEM major in the first place (see, for example, Ost, 2010; Minayal, 2020).

In the following pages, I review how the literature has responded to the above issues and introduce two new measures of student success based on rank. Making use of a large public university data set, I show that using rank reduces some of the noise associated with using raw grades. I also explain how and why “adjusted rank” is able to overcome the selection

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Educational Measurement: Issues and Practice* published by Wiley Periodicals LLC on behalf of National Council on Measurement in Education.

problem and resulting biases alluded to above and provide some evidence that adjusted rank is superior to an alternative measure of student success that adjusts raw grades. In essence, I undertake a variety of analyses based on different data subsets with different questions in mind. All of these analyses show the inherent biases in using GPA and/or the weakness of GPA relative to the other measures of student success.

Literature Review

The above selection problem is well understood, but dealing with it has been haphazard at best. Here are the most common responses found in my literature review: (1) ignore the problem and continue to use GPA; (2) look only at first year GPA; (3) just consider grades for a particular major or closely related set of majors; (4) employ grade relative to the class average; and much more rarely, (5) adjust the grades. At times, various combinations of 2, 3, 4, and 5 are employed.

Response 1 is often employed when the limited information available (in particular, how well other students did in the courses taken by the individual student is missing) prevents any adjustments for the selection effect (see, for example, Zwick & Himmelfarb, 2011; Zwick, 2013; Beard & Marini, 2013).

To make for an easy comparison to other approaches, it is useful to formally define GPA at this time. Let G_{ij} be student i 's grade (defined here as the numerical version of the letter grade) in class j . Let U_j be the number of units that class j is worth, and let $\{S_i\}$ be the set of classes under consideration that are taken by i (for example, those classes taken by i in i 's freshman year, FY). Then i 's average weighted grade in college equals GPA equals

$$GPA_i = \frac{\sum_{j \in \{S_i\}} U_j G_{ij}}{\sum_{j \in \{S_i\}} U_j}. \quad (1)$$

Using only first-year grades (response 2) is a move in the right direction because there is likely to be more overlap in courses taken in the first year, when students are fulfilling general education requirements and taking introductory courses, than in later years when most students are taking courses in their respective majors. This is the argument provided by those who use first-year grades (see for example, the just mentioned authors, Zwick & Himmelfarb, 2011; Beard & Marini, 2013).

Response 3 does away with trying to compare performance across all courses, and instead focuses on a narrower set of majors. Winter and Dodou (2011) consider specific majors within engineering (electrical, mechanical, etc.). Within each major, the overlap in courses taken is much higher. But the selection problem, while attenuated, still remains when students take elective courses. Of course, this method completely gives up on a general measure of relative performance over all majors. Note that Winter and Dodou also focus on specific courses in the high school curriculum rather than just using HSGPA.

Because GPA is such a widely used measure of student success and a tool employed in answering other issues, such as the relative importance of HSGPA versus the SAT in predicting GPA, researchers using GPA tend to spend little time discussing its merits and demerits. More robust arguments arise when authors make use of the alternatives.

Felton and Koper (2005) focus on the distortion in behavior when raw grades are used in determining GPA. They argue that, other things being equal, the lower the average grade for a class, the fewer the number of students taking the class and the more critical the students are of the professor's teaching. It is therefore in the interest of the faculty member to be a more lenient grader. Felton and Koper suggest using *relative grade* for each class and relative GPA (response 4) to reduce both grade inflation and the distortion in classes chosen when raw grades are used to determine GPA. More formally, let $\{T_j\}$ be the set of n_j students in class j . Then the average grade in class j is

$$\sum_{i \in \{T_j\}} \frac{G_{ij}}{n_j} = \bar{G}_j. \quad (2)$$

Then student i 's relative grade in class j equals i 's grade minus the class average grade equals

$$relg_{ij} = G_{ij} - \bar{G}_j. \quad (3)$$

In turn, i 's average weighted relative grade equals i 's relative GPA equals

$$relGPA_i = \frac{\sum_{j \in \{S_i\}} U_j relg_{ij}}{\sum_{j \in \{S_i\}} U_j}. \quad (4)$$

Brown and Van Neil (2012) go one step further by dividing relative grade by the standard deviation. More formally, they propose the following measure: i 's relative grade in class j divided by the estimated standard deviation of grades in class j equals

$$\frac{relg_{ij}}{\hat{\sigma}_j} = \frac{G_{ij} - \bar{G}_j}{\hat{\sigma}_j}. \quad (5)$$

Thus, the distance of i 's grade from the average grade in the class takes into account the standard deviation of grades within the class, making all measures in terms of standard deviations.

By construction the average relative grade (either Equation 3 or 5) for students in any *given* class is zero. Hence, the average relative grade in sciences classes is the same as the average relative grade in social science classes. Relative grade implicitly assumes that the variation in the average grade across classes is due to teacher idiosyncrasies regarding what grade should be assigned for average performance in the class. Unfortunately, relative grade does not take into account the average academic characteristics of the students taking the class and thus does not sufficiently deal with the selection issue raised in the introduction.

Response 5 uses same-student performance across a set of courses as a method of comparison to directly confront the selection issue raised in the introduction. Caulkins et al. (1996), Arcidiacono et al. (2012), and Vanderbei et al. (2014) adjusted *grades* to account for the differential selection of classes by students. Essentially, they determined the latent values for both individual classes and individual students by accounting for the level of academic performance of the students in their other classes (to be explained in greater detail later).¹

The just mentioned authors, along with Strenta and Elliot (1987), Young (1990), and Johnson (2003), are the most vociferous in their complaints about using GPA as a measure of success. Their arguments proceed along the following lines. Many students have an incentive to take courses where the grading is easy so that they will have a higher GPA, which will either improve their likelihood of getting into law school or prevent them from being on academic probation. In turn, there is an incentive for faculty members to grade more generously to avoid complaints and bad evaluations of their teaching by students who want a higher grade. Naturally, these incentives are not uniform across students and faculty members, but it does mean that stricter grading in STEM courses deters many students from taking such courses. It also weakens GPA as a measure of student success and makes college GPA harder to predict.

Method

Rank and Adjusted Rank as Alternative Measures of Comparative Performance

In this article, I directly confront the selection issue. I first introduce a third measure (beyond average grade, GPA, and average relative grade, relGPA), which is based on rank. Rank considers a student's percentile performance in each class that a student has taken. The term RANK in capitals is then a student's weighted *average* rank across classes. This is to be distinguished from converting GPA into rank as done, for example, by Bowen and Bock (1998) and Arcidiacono et al. (2012). I use the word rank instead of percentile because it is shorter in length and its abbreviation (rk) is easier to grasp than any possible abbreviation of percentile. The percentile is calculated so that average percentile in each class is 50%. In this way, class averages are again ignored. Looking at those classes in which i was enrolled, let $-ij$ refer to students in class j other than student i . Let the number of students in class j who have a grade strictly greater than i 's grade in class $j = \text{num}\{G_{-ij} > G_{ij}\}$. Let the number of students in class j who have a grade strictly less than i 's grade in class $j = \text{num}\{G_{-ij} < G_{ij}\}$. The number of students in class j has been previously defined as n_j . Finally, let i 's rank in class j equal

$$\text{rank}_{ij} = \frac{1}{2} \left[\frac{\text{num}\{G_{-ij} < G_{ij}\}}{n_j} + 1 - \frac{\text{num}\{G_{-ij} > G_{ij}\}}{n_j} \right]. \quad (6)$$

For example, if in class j , there are 4 students who have grades strictly lower than i 's grade, 2 students with grades strictly higher than i 's grade, and 3 students who have the same grade as i , then i 's rank in class is $.5[4/10 + 1 - 2/10] = 6/10$.

Student i 's average weighted rank equals

$$\text{RANK}_i = \frac{\sum_{j \in \{S_i\}} U_j \text{rank}_{ij}}{\sum_{j \in \{S_i\}} U_j} \quad (7)$$

Although students in particular classes are often ranked, I have not been able to find any evidence that average weighted rank over all courses (RANK_i) has previously been used as a measure of comparative student performance.

The difference between rank and relative grade can best be illustrated by considering two different symmetric distributions. In one, 10% of the students receive an A, 80% receive a B, and 10% receive a C, while in the other distribution, 20% of the students receive an A, 60% receive a B, and 20% receive a C. Using relative grade, an A (and a C) is valued the same in both distributions. But using rank, the first distribution places an A student in the top 95th percentile (and places a C student in the bottom 5th percentile), while the second distribution places an A student in the top 90th percentile (and places a C student in the bottom 10th percentile). That is, rank considers scarcity at the extremes.

Although rank is derived from grades, there is a subtle but important difference when using rank (and ultimately, RANK) rather than grades or relative grades (or some other technique that treats grades as being cardinal). A, B, and C are essentially ordinal. Receiving an A is better than receiving a B, but how much better is not specified. The standard 4-point system arbitrarily creates a cardinal system out of an ordinal system. Thus, getting an A (4 points) is worth twice as much as getting a C (2 points). But unlike grams and ounces, where 4 ounces is twice as large as 2 ounces, there is no universal agreement that an A is twice as good as C. Even the person teaching the class is unlikely to have a cardinal conceptual relationship between A and C. In contrast, rank and RANK are based on inherently cardinal relationships (the percentage of students with a grade below the student's grade).

Many Canadian and American high schools and colleges provide information on rank as well as GPA. But there is a crucial difference between what they do and what I do here. When a university provides the student's rank, it is based on GPA. For example, if the student has the 90th highest GPA out of 100 students graduating, then the student ranks in the 90th percentile. So, other things being equal, those students who choose courses where the grading is more generous will have higher GPA and a higher percentile based on GPA. But RANK is the average rank over all classes taken by the student and will not be directly influenced by the average grades in each class.

RANK as well as average grade and average relative grade ignore the possibility that some courses might attract higher performing students, while other courses might attract lower performing students. RANK would be a satisfactory measure of relative performance if course choice were random. But, as will be seen, the choice of course is not random. And if that is the case, one should consider the nature of the competition in coming up with an overall ranking of student performance. That is why in ranking tennis players, college basketball teams and chess players, not just the number of wins or the differential in scores between winners and losers are taken into account, but also the quality of the competition. For similar reasons, a student's average rank is not an adequate measure. So, one needs an estimation procedure that not only takes into account a student's performance (whether measured by grade, relative grade or rank) in a course, but also takes into account whether the other students in the course tend to be high performers in their other courses.

The intuition behind the procedure is as follows: Suppose that one looked at all students who took *both* course E and course S, and found that, on average, students who took course E ranked at the 50th percentile while these same

students on average ranked in the upper 65th percentile in course S. Then one could say that course S had academically weaker students. And if some students were only taking one of these two courses and one wanted to determine these students' ranking, it would be reasonable to subtract 0.15 from their rank in course S (or add 0.15 to their rank in course E, or some combination of the two) in order to account for the differences in composition for the two courses. In this way, one includes course fixed-effects by accounting for difference in the average student's rank between the two courses although the average rank or percentile performance *within* each course is 0.50 by construction.

While two-course comparisons provide the intuition, a more comprehensive method is needed. I now turn to a more sophisticated measure of student success—determining each student's rank fixed-effect (rkSFE). In so doing, each course's rank fixed-effect (rkCFE) is simultaneously determined. Speaking loosely, rkSFE not only considers the average rank the student is in all of the student's college courses, but also the average rank of their fellow course-mates in their other courses. rkSFE can then be used as an alternative to RANK. More formally, let C_j be a dummy (0, 1) variable for course j and s_i be a dummy variable for student i .² For now assume that each course is composed of just one class so that I can use the words course and class interchangeably. Let N_i be the number of classes that student i has taken for a grade, let I be the total number of students, and let J be the total number of courses. Then $\sum_{i=1}^I N_i = K = \sum_{j=1}^J n_j$ is the total number of observations. Assuming that the Gauss-Markov assumptions regarding the error term (ε_{ij}) hold, least-squares is the best linear unbiased estimator of the parameters in the following linear model:

$$\text{rank}_{ij} = \alpha + \sum_{i=2}^I \beta_i s_i + \sum_{j=2}^J \delta_j C_j + \varepsilon_{ij},$$

where $s_i = 1, s_{-i} = 0, C_j = 1, C_{-j} = 0$. (8)

All estimated parameters are relative to course 1 and student 1. Each observation is essentially about a particular student in a particular course; the other students and the other courses all have a zero value. The estimates of β_i are the student fixed-effects for rank (rkSFE) or student latent-values for rank, while the estimates of δ_j are the course fixed-effects for rank (rkCFE) or course latent-values for rank. This is a two-way ANOVA for students and classes. Before defining rkSFE, I used the phrase “adjusted student rank,” a less technical term for rkSFE. In the Appendix, I consider some technical issues that may arise when estimating unbalanced fixed-effects.

Substituting grade for rank in Equation (8) yields grCFE and grSFE. Caulkin et al. (1996), Arcidiacono et al. (2012), and Vanderbei et al. (2014) have provided student and course fixed-effects for grades. Arcidiacono, Aucejo, and Spenner went one step further and created percentiles from the set of adjusted GPAs. This transformation is analogous to the Canadian example provided earlier. Percentile would be different from the one that I develop as the method of least-squares is minimizing different objective functions. Note that relative grade, defined in Equation (3), can be seen as a different kind of grade adjustment, but subtracting the average grade in the

class from the individual grade is not based on minimizing the average squared-error.

I have suggested above that, compared to class grade and relative grade, class rank is a superior measure of class performance. If this is correct, then adjusted rank will likely be superior to adjusted grade, as well. The remainder of this article provides empirical corroboration for both statements. I first show that correlations across courses are greater for rank than for grade or relative grade. I then look at course average grades and show that either (1) academically weak students tend to take courses where the grading is easier or (2) faculty are easier graders when students are on average weaker performers. This tendency is so strong that courses filled with academically strong students actually tend to have lower average grades. Finally, I show that the explanatory power of SAT scores and HSGPA is much larger when student rank fixed-effect (rkSFE) and student average rank over all classes (RANK) are the dependent variables than when grSFE and GPA are the dependent variables.

Data

Here are the data in brief: I have grades for every student in every class taken from Fall 2009 through Summer 2013 at University of California Santa Cruz (UCSC). I also have SAT scores and high school GPA (HSGPA) for everyone who entered as a freshman during the fall term from 2006–2012 (SAT changed form in February 2005). In total, there are 22,314 students, 2,102 courses (note that one or more classes are in a course), and 414,156 observations.

The data used to determine course fixed-effect for rank (rkCFE) and student-rank fixed-effect (rkSFE) include *all* courses that have at least 150 students in total *over the 4-year period* 2009–2013 when only counting classes with at least 20 students. In total, this covered 22,106 students in 463 courses for a total of 298,659 observations. The 150–20 choice was a judgment call. Classes that were too small would add noise (in the limit, a single-student class would always rank the student at the 50th percentile). Limiting the sample to larger courses provides more accurate information about the individual courses (which is useful when discussing course fixed-effects, as I do later), but provide fewer courses for measuring student fixed-effects. So the 150–20 choice was a compromise between the competing desiderata. To make for a sensible comparison, measures for average grade (GPA), average relative grade (relGPA) and average rank (RANK) were also subject to the 150–20 criteria.

I have made a number of decisions about which aspects of the data that I actually employ. With one exception, these decisions were made in the absence of any knowledge about admission data such as SAT scores and HSGPA, I try to limit confusion by initially restricting the presentation to as few models and data sets as possible. Then, toward the end of the article, I consider a number of variations. I show that the results do not depend on the particular regression models or particular data sets employed earlier in the article. Details are provided at the appropriate times.

Empirical Strategies: Motivation and Results

In each of the following subsections, a different empirical strategy is undertaken to determine whether there are

Table 1***Average Grades in Each of the Five Divisions at UCSC 2009–2013***

| Division | Mean Grade in All Courses | Number of Observations | Mean Grade in First Year | Number of Observations |
|-----------------|---------------------------|------------------------|--------------------------|------------------------|
| Science | 2.81 | 135,275 | 2.75 | 42,085 |
| Engineering | 3.07 | 56,270 | 3.01 | 14,354 |
| Social science | 3.12 | 106,617 | 2.92 | 15,906 |
| Humanities | 3.23 | 63,516 | 3.17 | 16,808 |
| Performing arts | 3.38 | 52,478 | 3.28 | 12,761 |

Note. Divisions are administrative units but have no impact on student choice. That is, any student can choose any course in any division as long as the student satisfies the department's requirements. Average first-year grades in engineering are higher than average first-year grades in social science contrary to the usual characterization of STEM courses being more harshly graded.

Table 2***Average Grades in Selected UCSC Math Courses 2009–2013***

| Course # | Mean Grade | Name |
|----------|------------|--------------------------------|
| Math 3 | 2.30 | Precalculus |
| Math 11A | 2.31 | Calculus for nonmajors A |
| Math 11B | 2.51 | Calculus for nonmajors B |
| Math 19A | 2.42 | Calculus for majors A |
| Math 19B | 2.17 | Calculus for majors B |
| Math 20A | 2.47 | Honors calculus A |
| Math 20B | 2.85 | Honors calculus B |
| Math 100 | 2.31 | Required upper-division course |

problems with using grades and/or determine whether the other measures might have superior attributes.

Average Grades

It is useful to start with a summary of the average grades over the five academic divisions (see Table 1). These numbers immediately raise questions about using GPA as a measure of relative student performance. For example, does the higher mean grade in humanities relative to the mean grade in science indicate that students in humanities are stronger academically? STEM department grades tend to be the lowest elsewhere, as well (Aachen & Courant, 2009; Ost, 2010). Some might believe that academic performance is spread equally across divisions and therefore might recommend using relative grade instead, in which case all of the divisions would have the same mean relative grade.

For raising questions (but not necessarily providing answers), it is also useful to look at a particular department in detail. In this case, mathematics is the obvious choice. It is the only department that provides more than one sequence for the introductory courses, and, as can be seen in Table 2, there are three versions of calculus. Furthermore, the sequence is clear. A student does not take Math 11B first and then take Math 11A. In contrast, in many other divisions one can take the B course before the A course. All except Math 100 are likely to be taken during a student's first year. Does an average grade of 2.3, approximately a C+ on a 4-point scale, mean the same in precalculus as it does in calculus for nonmajors or junior-level mathematics for majors? Using relative grade does not seem like a good idea in this case. However, one would expect that the average level of academic performance is higher in honors calculus than it is in calculus for

math majors, and, in this case, average grades do reflect the expectations.

Correlations across Courses

Much of the argument in this article is ultimately based on there being a positive correlation in student performance across courses. That is, if individual B has a higher grade than individual C in course X, then individual B is likely to have a higher grade than individual C in course Z. Clearly, if there were zero correlation across courses, then information on how well a student did relative to another student in one course would provide no information on how well the student would likely do relative to the other student in another course and student fixed-effects (SFE) would be zero. To see whether positive correlations hold, I first look at pairwise correlations across each major's introductory course.³ It is logically possible that student X does better than student Y in course B, but student Y does better than student X in course C. After all, it is possible for an individual to be naturally gifted in economics but not so in English literature or to be motivated to do well in English literature, but not so motivated in economics when it just satisfies a requirement. Thus, we will have to look at the data to see whether there actually is a positive correlation.

A course may be composed of one or more classes. The data in this and the following section are presented in terms of courses rather than classes (even though the course grades relative grades, and ranks are based on performance within the relevant class within the course). There are both theoretical and practical reasons for focusing on courses. First, the differential selection among classes within a course is likely to be greatly muted in comparison to the differential selection

across courses and majors. This clearly holds when the course is only taught once or twice a year. Students are unlikely to wait a semester or year for a professor who is an easier grader, when who this might be is generally unknown. Even if there are more than two classes taught by different professors in a particular semester, many students may not be able to exercise their preference for one of the professors because of time conflicts with another course or because the department limits choice. In brief, the variation in average SFE across classes within a course is likely to be relatively moderate in comparison to the variation of average SFE across courses.

There is a practical problem when using class rather than course as the unit of analysis (when class performance is aggregated to the course level). The number of variables to be estimated is multiplied, and, at the same time, the average number of observations per variable is reduced by the same multiple. Hence, I group classes into courses when determining correlations.

Once there is more than one class in a course, a student's course rank is the student's class rank, the student's relative course grade is the relative class grade, and the student's course grade is the class grade. Essentially, within the same course, a grade of A in one class is viewed as being the same as a grade of A in another class, even though in one class 20% of the class received an A while in the other class only 10% received an A. Likewise, within the same course, being in the top 20% of the class is treated as the same as being in the top 20% in another class, even though the top 20% in the first class received an A while in the other class, the top 20% received a B. When estimating rkCFE and rkSFE, the same arguments for combining classes into course data hold and the same methodology as used here will be employed.

Results

To see whether positive correlations hold, I first look at pairwise correlations across each major's introductory course. There are 24 introduction to the major courses plus writing that is not a major but is required for nearly all students, and 300 pairwise correlations between these 25 courses. Of the 300 rank correlations, only 12 are negative (computer engineering and art each have 4 negative correlations out of the 12), and none of these negative correlations are significant at the 0.05 level. In contrast, there are 133 positive correlations with reported 0.0000 significance levels, and there are 91 positive correlations between 0.0001 and 0.05. Sixty-four positive correlations were not significant at the 0.05 level; 27 of these involved art or history of art and most of the rest had fewer than 10 observations. With regard to the earlier discussion, the correlation in rank between Economics 1 and Literature 1 for the set of students who take both courses is 0.43. Similar but weaker results hold for relative grade and grade. See Table 3, left half, for the average correlations. For more detailed data on correlations across courses, see the online appendix. Note again that the correlations are across courses, but relative grade and rank are based on the individual classes within the course. If each course were composed of only one class, then relative grade and grade would have identical correlations. It appears that that the average class grade within a course varies and therefore it is not surprising that the relative grades have a higher correlation across majors than plain grades. The expectation is that the variation in grading across

Table 3

Average Correlations of Grades, Relative Grades, and Rank across and within Majors

| Average of Correlations across Each Major's Required Introductory Course | | | Average of Correlations between Required Introductory Course and First Required Upper-Division Course within Each Major | | |
|--|----------------|------|---|----------------|------|
| Grade | Relative grade | Rank | Grade | Relative grade | Rank |
| 0.35 | 0.36 | 0.40 | 0.37 | 0.39 | 0.45 |

classes within a course is mainly due to idiosyncratic grading by the teacher rather than the differential selection of students.

It is insightful to look at the parallel data when finding correlations between the required introductory course to the major and the first required upper-division course in the same major (Table 3, right half). The average correlations are again highest for rank and lowest for grade. A more surprising result is that the average correlation between the introductory course grade and the required upper-division course grade, 0.37, is lower than the correlation between the introductory course rank and required upper-division grade, 0.39 (not reported in the table). Despite the fact that rank is a nonlinear transformation of grade, rank in the lower-division introductory course has, on average, a stronger linear relationship to grade in the upper-division required course than does grade in the lower-division course. In essence, rank in the lower-division introductory course is a better predictor than grade of the upper-division course grade. This set of results provide the first empirical evidence that rank might be a better performance measure than relative grade and that relative grade might be a better performance measure than grade. Grades are a noisier measure of performance than relative grade and rank. In turn, this suggests that adjusted rank (rkSFE) will be a superior measure to adjusted grade (grSFE) because the latter is based on noisier data.

Courses with Academically Stronger Students Tend to Have Lower Average Grades

Tables 1 and 2 suggest that grades across courses (as opposed to within classes and courses) might not reflect relative student performance very well. In this section, I consider the issue in a more systematic way and ask whether courses with academically stronger students tend to have lower average grades. But what evidence can be used to demonstrate that a course is composed of academically stronger students? There is a large literature showing a significant positive relationship between SAT scores, as well as HSGPA, with college GPA (my criticism in this article is not that this relationship does not exist but that there are stronger relationships when using other measures of college performance). So, one would expect that college courses composed of students with higher SAT scores would have higher average grades. Similarly, one would expect that college courses composed of students with higher HSGPAs would have higher average grades.

Table 4**Correlations of Course Average Grade with Course Mean Characteristics**

| Statistic | Course Mean | | | | | |
|--------------|-------------|-------|-------|-------|-------|-------|
| | SAT | HSGPA | rkSFE | rkCFE | grSFE | grCFE |
| Correlation | −0.15 | −0.10 | −0.16 | 0.16 | −0.17 | 0.8 |
| Significance | 0.00 | 0.035 | 0.00 | 0.00 | 0.00 | 0.0 |

Note: In each case, the number of observations is 463. Because the average rank in each class is 0.5, course fixed-effect for rank (rkCFE) and mean student rank fixed-effect (rkSFE) will be identical except for sign. This does not hold for course fixed-effect for grade (grCFE) and mean student grade fixed-effect (grSFE). During the time period under investigation, the SAT test was worth 2,400 points in total.

Table 5**Correlations between MEANgrCFE and Measures of Student Performance**

| Statistics | SAT | HSGPA | rkSFE | RANK |
|--------------|--------|--------|--------|--------|
| Correlation | −0.12 | −0.08 | −0.17 | −0.09 |
| Significance | 0.00 | 0.00 | 0.00 | 0.00 |
| Observations | 22,106 | 22,060 | 22,106 | 22,106 |

Note. MEANgrCFE is the weighted average of course fixed-effects for those courses that the student has taken. The greater MEANgrCFE is, the more generous the grading is. The results show that academically stronger students take courses that grade less generously. These data are not restricted to courses taken during the first year. Homeschooled students do not have HSGPAs. Therefore, the number of observations is lower for HSGPA.

Results

Looking at Table 4 (covering *all undergraduate* courses that meet the 150–20 criterion), the correlation between course average grade and course average SAT scores (column 1) and the correlation between course average grade and course average HSGPA (column 2) are both negative, contrary to what was just said in the previous paragraph.

As a third measure, rkSFE scores are a measure of academic strength based on the students' university performance. It is therefore to be expected that classes composed of students with high rkSFE scores would have academically strong students (almost by definition). So, one might expect a positive relationship between the mean rkSFE in a course and the course average grade. But Table 4, column 3, shows that the correlation between the average rkSFE in a course and the average grade in the course is again negative. One can also look at course fixed-effects for rank (rkCFE), which provides similar information to the course average rkSFE, but the larger rkCFE is, the less academically competitive the students in the course are. So, one would expect a negative correlation, but it is positive. Using grSFE and grCFE instead of rkSFE and rkCFE echos the results.

Courses with stronger students (whether measured by mean SAT scores, mean HSGPA, mean rkSFE, or rkCFE) have not just tougher grading standards, but so much tougher grading standards that in each case the sign of the slope is not just smaller in absolute terms than one might expect, but actually reversed from that expected. This result is worth reflecting upon. First, recall the earlier result that there is a high positive correlation of student grades across courses. That is, students who have higher grades than other students in one course will tend to have higher grades than these other students in another course. Now, suppose that every course is graded on a curve with B+ being the mean grade for each course. Then, all courses would have the same average grade and the correlation between the average grade for a course and the average SAT score (or average rkSFE) in the course

would be zero. Nevertheless, students in the more academically competitive course would tend to have higher grades when taking less competitive courses. But the results in this section say something more striking: the average grade is higher in courses with academically weaker students than the average grade in courses with academically stronger students.

It is useful to switch the focus from courses to students. Table 5 is based on each student's weighted grCFE (MEANgrCFE). The method of calculation is similar to GPA, but instead of a grade for each class, grCFE is used in its place. Hence, MEANgrCFE is the average weighted grCFE that the student has taken. The *greater* MEANgrCFE is, the *more* generous the average grading policy is in the courses that the student has taken. MEANgrCFE is regressed in turn against the student's SAT, HSGPA, rkSFE, and RANK (the student's average rank over all classes). All the correlations are negative.

The results show that GPA is a biased measure of relative performance because academically weak students tend to take courses where the grading is generous.⁴ However, most academically stronger students perform at a sufficiently higher level than the other students in academically weak and/or academically average courses. Therefore, their GPAs will be higher than average despite the bias, and a positive correlation between GPA and HSGPA, SAT, or rkSFE will still be observed.

As noted in the introduction, STEM courses have been shown to have harsher grading policies. Here I provide some compelling numbers. To do so, I make use of course fixed-effects for grade (grCFE). Again I focus on all undergraduate courses with at least 150 students in total over four years. Of the 50 courses with the most negative grCFE (meaning that the students received grades that were too low in these courses), all 50 were STEM courses. The downward bias ranged from −0.74 to −1.48 (where an F is worth 0 and an A+ is worth 4.33). I note that in these 50 courses, only three had an average SAT score below 1,660 and only two had a median SAT score less than 1,660. Next, I looked at the 50 courses

with the highest grCFE. Only 6 of the 50 were STEM courses and these 6 were mainly remedial or directed toward non-STEM students. In 38 of these courses, the average SAT was below 1,660 and in 29 of the courses, the median SAT score was below 1,660. The upward bias of these 50 ranged from 0.47 to 1.17. The differential between the course with the greatest bias for and the greatest bias against was $1.17 - (-1.48) = 2.65$. Focusing on just those courses that many freshmen take, the most extreme differential was where receiving 1 point (D) in one particular course was equal to receiving a grade of 3.25 (nearly a B+) in another course. So, the difference between the actual grade in a course and its adjusted grade need not be trivial. In a nutshell, this section suggests that GPA is a biased measure of relative performance because weak students tend to take course where grading is very generous.

Using SAT and HSGPA to Predict the Various Measures of College Success

So far, I have not considered, except indirectly, the relationship between high school GPA and SAT scores on the one hand and performance at the university on the other hand. But the previous results suggest that HSGPA and SAT scores will have a greater explained variation when predicting relGPA, RANK, and especially rkSFE than when predicting plain GPA.

Before proceeding, it is useful to consider some limits to the research design. Like all studies based on matriculated students, the range of HSGPA and SAT scores are limited because students with very low SAT *and* HSGPA values are not accepted and quite possibly did not apply in the first place, other variables that were relevant for admission are not available, and the values of the observed variables are endogenous to the admission process. To illustrate the last point, the positive correlation between HSGPA and SAT scores in the general population is reduced in the matriculated set of students because individuals with low SAT scores are likely to be admitted only if they have high HSGPAs, while those with low HSGPAs are likely to be admitted only if they have high SAT scores.

I do not have access to data for those who applied and did not matriculate and therefore range-adjustment, a technique to account for the reduced variability of HSGPA and SAT scores among the matriculates, is not possible. In any event, range-adjustment techniques are questionable (see Rothstein, 2004). Instead, I provide an alternative method in an online appendix to deal with this issue. Because I am comparing different measures of success under the same limited conditions, it is unlikely that the comparative results in this section would be undermined under less limited conditions.

I employ the following measures of a student's first-year college success: (1) GPA (average class grade); (2) relGPA (average class relative-grade); (3) RANK (average class rank); (4) rkSFE (student rank fixed-effect), and (5) grSFE (student grade fixed-effect). I then regressed each of these variables against HSGPA and SAT scores. I purposely chose this sparse regression equation so that the basic model is amenable to adding other variables depending on the interest of the researcher.

I consider two forms of the regression equations:

$$\text{Linear: } Y = B_0 + B_1\text{HSGPA} + B_2\text{SAT}, \quad (9)$$

$$\begin{aligned} \text{Double log: } \text{LOG}(Y) = & \text{LOG}(B_0) + B_1\text{LOG}(\text{HSGPA}) \\ & + B_2\text{LOG}(\text{SAT}). \end{aligned} \quad (10)$$

Y is one of the five measures of student success during the first year: fyGPA, fyrelGPA, fyRANK, fygrSFE, and fygrSFE. The double log form assumes that the effect of HSGPA and SAT is multiplicative rather than additive (the case for the linear form).

Results

As can be seen by looking at the linear regressions reported in Table 6, the explained variation (R^2) increases considerably when the dependent variable goes from GPA to relGPA to RANK to rkSFE. As before, I am only considering courses that have at least 150 students composed of classes of at least 20 students. During the years 2009–2013, 13,236 students took a total of 89,911 classes during their first year.

I also estimated course and student fixed-effects for grades (grSFE). I then regressed grSFE on SAT and HSGPA. This is the approach used by Caulkins et al. (1996), Vanderbei et al. (2014), and Koester et al. (2017). The results are shown in the last column. Not surprisingly, the explained variation for grSFE is higher than the explained variation of GPA, but as can be seen it is still lower than for RANK and rkSFE. Undertaking a similar exercise for relative grade (rgSFE) produces virtually identical results (not reported) to those when using grSFE as the dependent variable. The reason for their similarity is the adjustment for grade and the adjustment for relative grade shift these two measures up and down so that they become nearly the same measure.

Note that deriving the values of grCFE and grSFE involve a different objective function from that used in deriving the values of rkCFE and rkSFE. For grades, course fixed-effects (grCFE) identify those courses where the average grade is high (or low) given the average grades of the students in their other courses. In contrast, for rank, course fixed-effects (rkCFE) identify those courses where the average rank of students in their other courses is high (or low). Thus one would expect that ranking the student grade fixed-effect, that is, creating the function rank(grSFE) as Arcidiacono et al. (2012) have done and using it as the dependent variable will yield a lower R^2 than when using rkSFE as the dependent variable, which was indeed the case when I did it (results not reported).

To determine whether my results were sensitive to the model specification that I had made, I considered a number of variations.

1. Nearly identical results occurred when all classes taken by first-year students are included and CFEz is based on all classes (z is for zero restrictions on class and course size). A possible reason for the great similarity is that lower division courses in general and first-year courses in particular tend to be large so that they pass the criterion of 150 students over the 4-year period, and, in any event, small classes are weighted less in determining CFE and consequently SFE. But the presences of classes with few students means that the negative relationship between course average grade and course average rkSFE goes toward zero when classes as small as one student are included in the regression.

Table 6*Regressing Various Measures of First-Year Student Performance on SAT and HSGPA*

| Dependent Variable | SAT | HSGPA | Constant | Adjusted R^2 |
|--------------------|------------------|----------------|-----------------|----------------|
| fyGPA | 1.00 (0.02) | 0.65 (0.02) | -1.15 (0.08) | 20 |
| fyRELgpa | 1.00 (0.02) | 0.67 (0.02) | -4.06 (0.07) | 0.22 |
| fyRANK | 0.32 (0.00) | 0.21 (0.01) | -0.79 (0.02) | 0.26 |
| fyrkSFE | 0.37 (0.01) | 0.21 (0.01) | -0.88 (0.02) | 0.28 |
| fygrSFE | 1.16 (0.0.02) | 0.71 (0.02) | -1.41 (0.08) | 0.24 |

Note. Standard errors are in parentheses. In each case, there are 13,236 observations.

2. A similar ordering occurred when I employed a Cobb-Douglas (double-log) production function (Equation 10), but the difference between the R^2 when $\log(\text{rkSFE})$ was the dependent variable and the R^2 when $\log(\text{GPA})$ was the dependent variable was more dramatic.
3. When I used the method suggested by Brown and Van Niel (2012), $(\text{relative grade})/(\text{standard deviation})$, $R^2 = 0.2278$, which was an improvement over plain relative grade (relGPA), but still not close to RANK.
4. A similar ordering of R^2 to that found in Table 6 also occurred, but the explained variation was lower, when the sample only includes the introductory to the major courses, covered earlier. Note that students can take an introductory to the major course at any time in their career. However, these courses are almost always taken during the student's freshman and sophomore years.

To summarize, no matter how I divided the data or characterized the regression equation, GPA yielded the weakest results, while RANK and rkSFE yielded the strongest results.

Discussion

Student performance in each class can be characterized in a number of ways (grade, relative grade, rank, etc.). In turn, each characterization can be used as a basis for generating a measure of success over all classes. This article has evaluated the various measures. With regard to the unadjusted measures (grade and rank), in the abstract, one can argue in favor of either. Grades ignore percentile (e.g., how many students receive an A), while percentile (rank) ignores grades (if no one receives an A, then B+ will be in the top percentile). However, the evidence is clearly in favor of the percentile approach. For example, the correlation between the introductory course in the major and the intermediate required course in the major is higher for rank than it is for grade. Additionally, *rank* in the introductory course is a better predictor of *grade* in the intermediate required course than grade in the introductory course is a predictor of grade in the intermediate course. Furthermore, when using SAT and HSGPA as the explanatory variables of first-year college performance, the R^2 is greater when predicting average first-year rank (fyRANK) than when predicting average first-year grade (fyGPA).

Both grade and rank are weak measures of student success when compared to their adjusted counterparts (student grade fixed-effect, grSFE, and student rank fixed-effect, rkSFE, respectively). This is because academically stronger students, whether measured by their relative performance in their other college courses or by their SAT scores or by their HSGPA, tend to take courses with more stringent grading standards and therefore GPA and average RANK are severely biased measures of student success.

Given the previous two paragraphs, one should not be surprised that GPA is less predictable than adjusted rank (rkSFE). When regressed against HSGPA and SAT scores, the explained variation of GPA was 0.20 in contrast to the explained variation of adjusted rank, which was 0.28. In a nutshell, GPA has been shown to be the weakest measure. Of course, HSGPA is even more problematic than college GPA since the former typically includes variations both within high schools and across high schools (see Koretz & Langi, 2018).

In many universities outside of the United States, grading is based on percent of questions answered correctly (not to be confused with percentile or rank that refers to performance relative to other students). The difficulty of the questions asked can vary from course to course and thus getting 80% in one course may be easier than getting an 80% in another course. So this number has the same problem as using grades 0 to 4 or F to A. And the same holds for alternative grading systems using a different set of numbers unless the number stands for rank. And even then, as I have shown, that is not enough if the academic competitiveness of the class is not considered (that is, if student fixed-effects are not estimated). Students in European universities tend to take few courses outside of their major. This makes it harder, if not impossible, to determine student fixed-effects across majors. But that does not mean that academically weaker students are not drawn to majors where the grading is easier. And to the degree that there are elective courses in a major, one can again find student fixed-effects albeit limited to the major. Finally, there are weaker but still useful comparisons. One can see whether students with lower HSGPA flock to certain majors.

In tennis, college basketball, and chess, ranking depends not only on how well the individual or team did against the competitors, but also on how strong the competitors were. It is surprising that universities do not engage in a similar

exercise since the large numbers of observations allow for greater understanding.

This article has provided two new measures of competitiveness: student rank fixed-effect (rkSFE) and course fixed-effect for rank (rkCFE). These measures can be very useful for discovering issues and diagnosing problems. Is passing a course a mark of learning when a student can only fail by not doing the assignments? And, in accepting students, are universities doing the best that they can do in predicting who will be most successful? The results in this article suggest a number of policy refinements for higher education. Here are a few ideas for improvement.

1. Because class rank adds valuable information about a student's performance beyond what can be determined from a grade alone, some colleges might want to include class rank along with the grade in a student's transcript. This seems relatively easy to implement. Once grades have been entered on a spreadsheet, calculating rank only requires entering a few details into a relatively simple mathematical formula. This can be done centrally, rather than by individual faculty members. Such information can be valuable to students. Did an A– grade mean that the student's performance was superior to the average student in the class or did it mean the opposite because a majority of students received a solid A?
2. In some universities, a number of departments limit access to the major by requiring a certain grade in the introductory course in order for the student to be eligible for taking the intermediate-level course in the major. In comparison to grade, rank in the introductory course is a better predictor of performance in the required upper-division course. Therefore, a department might want to use rank rather than grade as a criterion for admission to the major. However, most students who would be admitted under one criterion would be admitted under both criteria. Thus, a department might want to raise the grade-point criterion somewhat but also allow the rank criterion to be invoked if the student did not satisfy the grade criterion. Students, themselves, may make more sophisticated major choices if they know their ranks in the introductory classes.
3. In some instances, students are required to have a certain grade-point average in order to maintain their scholarship. At times, this may encourage scholarship students to take courses where the grading is easier rather than where the knowledge gain is greater. This is an undesirable by-product of the grade requirement. It seems that a possible solution to this problem is for the scholarship committee to take into account that the student took courses where the grading was stricter if this were in fact the case.
4. In several universities, students whose GPAs are in the top 10% of the graduating class have “magna cum laude” written on their diploma. Some of those who majored in departments where the grading was stricter might not get the recognition they deserve. A solution might be that magna cum laude is awarded to those students who have a GPA within the top 8% of GPAs or who have a student-rank fixed-effect (rkSFE) within the top 8% of rkSFEs.
5. A number of universities use a formula for admission that to some degree relies on estimated parameters

when regressing GPA on independent variables such as HSGPA and SAT scores. Improvement in predictability is likely if, instead of GPA, admissions use student fixed-effects, as has been done here. Many applicants would be admitted under either criteria, but there could be improvement in predictability for those applicants who are close to the margin, one-way or the other. A test of this hypothesis would require the admission committee to track student performance where the two criteria were in disagreement.

I suggest that whenever the relevant data are available, future researchers should not only have grades in their equations, but also substitute rank for grades and student fixed-effects (SFE) for grades and rank. At times, these substitutions will confirm the results using only grades, and that is good because it makes the results more robust. At other times, contrary results may arise. The following are some topics that might gain from using a different measure of academic success from GPA: the big five personality traits as predictors of college GPA, the impact of financial aid on college GPA, and determining the characteristics of students whose academic performance improves over the 4 years of college. Even without asking new questions, there is room for new answers when using RANK and rkSFE.


One study can never be definitive. With regard to the question whether grading differences across departments are persistent and common across universities, there may be some characteristics of the students or the institution that are unique to University of California Santa Cruz, but the following suggests that this is probably not so. Grades in UCSC introductory courses to the major were regressed against University of Michigan lower division grades in the major (from Achen & Courant, 2009). For the 16 overlapping majors, the correlation was 0.77. More important is the tendency for academically weak students to avoid courses where the grading is strict, thereby undermining GPA as a measure of comparative performance across classes. This is not the first article to raise these issues, but it definitely puts more nails in the coffin.

The question as to whether rkSFE will continue to be superior to grSFE as a measure of student success will only be answered by more comparative studies. This article has provided an explanation for why this might be the case—grades are noisier than rank and this noise will tend to carry over in the adjustment process. The good news is that if performance by all students in all their classes is available, then both grSFE and rkSFE can be calculated, with the latter only requiring a simple transformation of the individual grades in each class. So in the future, whenever possible, research should follow what has been done here and provide results for the alternative measures of student success.

Acknowledgments

I would like to thank D. Bonett, G. Bulman, and S. Sinharay for very helpful suggestions.

ORCID

Donald Wittman  <https://orcid.org/0000-0002-1073-4345>

Notes

¹See Smits et al. (2002) for some other methods of dealing with missing values, and Young (1990) and Johnson (2003) for a discussion of Item Response Theory (IRT), a nonlinear estimation method for ordered values.

² s_i should not be confused with S_i .

³If there is more than one introductory course, I choose the most popular. Not all introductory courses are taken in the first year.

⁴This suggests that HSGPA might also be a biased measure of student success, and it very well could be, in general. However, University of California's admission policy does not treat all courses equally. Advanced placement courses receive an additional one point to the ordinary four-point scale, and other high school courses are not included in the calculations. So, it is unclear what the bias is, if any, and in what direction.

References

- Achen, A., & Courant, P. (2009). What are grades made of. *Journal of Economic Perspectives*, 23, 77–92.
- Arcidiacono, P., Aucejo, E. M., & Spenner, K. (2012). What happens after enrollment? An analysis of the time path of racial differences in GPA and major choice. *IZA Journal of Labor Economics*, 1, 5. <https://doi.org/10.1186/2193-8997-1-5>
- Beard, J., & Marini, J. (2013). *Validity of the SAT® for predicting first-year grades: 2013 SAT validity sample*, New York, College Board. Retrieved from <https://collegereadiness.collegeboard.org/pdf/national-sat-validity-study.pdf>
- Bowen, W. G., & Bok, D. (1998) *The shape of the river: Long-term consequences of considering race in college and university admissions*. Princeton NJ: Princeton University Press.
- Brown, P. H., & Van Niel, N. (2012). Alternative class ranks using Z-scores. *Assessment & Evaluation in Higher Education*, 37, 889–905.
- Carrell, S., & West, J. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy*, 118, 409–432.
- Caulkins, J., Larkey, P., & Wei, J. (1996). *Adjusting GPA to reflect course difficulty*. Carnegie Mellon University V1.1. Retrieved from <https://doi.org/10.1184/R1/6470981.v1>
- Felton, J., & Koper, P. T. (2005). Nominal GPA and real GPA: A simple adjustment that compensates for grade inflation. *Assessment & Evaluation in Higher Education*, 30, 561–569.
- Johnson, V. E. (2003). *Grade inflation: A crisis in college education*. New York, NY: Springer.
- Koretz, D., & Langi, M. (2018) Predicting freshman grade-point average from test scores: Effects of variation within and between high schools. *Educational Measurement: Issues and Practice*, 37, 9–19
- Kwon, S., (2021) Shrinkage estimation of fixed effects in linear panel data models. Retrieved from http://ses.wsu.edu/wp-content/uploads/2021/01/Soonwoo_Kwon_JMP.pdf
- León, M. B., Guest-Scott, A., Koke, A., Fiorini, S., & Rangazas, A. (2019). Claiming their education: The impact of a required course for academic probation students with a focus on purpose and motivation. *Journal of the Scholarship of Teaching and Learning*, 19, 43–57.
- Minaya, V. (2020). Do differential grading standards across fields matter for major choice? Evidence from a policy change in Florida. *Research in Higher Education*, 61, 943–965.
- National Association for College Admission Counseling (2016). *Use of predictive validity studies to inform admission practices*. Retrieved from <https://www.nacacnet.org/globalassets/documents/publications/research/testvalidity.pdf>
- National Association of Colleges and Employers (2018). *Job Outlook 2019*. Retrieved from <https://www.odu.edu/content/dam/odu/offices/cmc/docs/nace/2019-nace-job-outlook-survey.pdf>
- Ost, B. (2010). The role of peers and grades in determining major persistence in the sciences. *Economics of Education Review*, 29, 923–934.
- Rothstein, J. (2004). College performance predictions and the S78AT. *Journal of Econometrics*, 12, 297–317.

Searle, S. R., & Gruber, M. H. J. (2016). *Linear models*. Hoboken, NJ: John Wiley.

Smits, N., Mellenbergh, G. J., & Vorst, H. C. M. (2002). Alternative missing data techniques to grade point average: Imputing unavailable grades. *Journal of Educational Measurement*, 39, 187–206.

Strenta, C., & Elliott, R. (1987). Differential grading standards revisited. *Journal of Educational Measurement*, 24, 281–291.

Vanderbei, R. J., Scharf, G., & Marlow, D. (2014). Regression approach to fairer grading. *SIAM Review*, 56, 337–352.

Winter, J. d., & Dodou, D. (2011). Predicting academic performance in engineering using high school exam scores. *International Journal of Engineering Education*, 27, 1343–1351.

Young, J. W. (1990). Adjusting the cumulative GPA using item response theory. *Journal of Educational Measurement*, 27, 175–186.

Zwick, R. (2013). *Disentangling the role of high school grades, SAT® scores, and SES in Predicting College Achievement*, ETS Research Report, 13, 1–10. Retrieved from <https://www.ets.org/Media/Research/pdf/RR-13-09.pdf>

Zwick, R., & Himelfarb, I. (2011). The effect of high school socioeconomic status on the predictive validity of SAT scores and high school grade-point averages. *Journal of Educational Measurement*, 48, 101–121.

Appendix

Deriving fixed-effects

Fixed-effects cannot be determined for subgroups of students that are not connected to the larger group. For example, if 10 students only take courses 101 through 105, and no other students take these courses, fixed-effects for these students relative to the other students cannot be determined. Stata, the statistical package used for this article, automatically eliminates such students from consideration. This would be a problem when comparing student fixed-effects to GPA if this involved widely divergent data sets. However, this potential problem does not exist in the data. For example in Table 6, the number of students is the same for estimating SFE as it is for GPA. Table 6 restricts the data set to courses that have at least a total of 150 students over the relevant time period. If Table 6 had no restrictions on course and class size, the number of students would only be 3 more for GPA than it would be for SFE.

As one adds more students to the sample, the number of dummy variables increases even if the number of courses does not expand. Therefore, the typical formula used for estimating population variance is no longer appropriate, and the appropriate formula for estimating the population variance produces larger numbers than otherwise. In Stata, areg and xtreg produce the same estimates for the b -vector, but xtreg appropriately calculates larger variances.

If $X'X$ is not full rank, then the vector b of coefficients in the equation $b = [X'X]^{-1}X'Y$ cannot be determined because one is dividing by zero. Instead, a generalized inverse, G^{-1} , is found (see Searle & Gruber, 2016). There are many generalized inverses when $X'X$ is not full rank and thus the b vector depends on which G^{-1} is chosen, but, ignoring the intercept term, the difference in each element between a particular set of two b vectors is a constant. And because there are no other variables in the equation that determines fixed-effects, these differences are of little consequence when finding fixed-effects.

For unbalanced data, there is the concern that the estimates of β are unstable. A number of authors have undertaken various shrinkage methods (see, for example, Kwon,

2021). These methods reduce the mean squared error, but the coefficient estimates are biased. I have stayed with the unbiased least-squares estimates, which do not involve shrinkage (when the students are connected). The rationale for my decision is best understood when the shrinkage involves eliminating variables. Because the article is focused on comparing GPA to other measures, including fixed effects, it would make for a difficult comparison if GPA and SFE were covering different data or GPA were forced to measure only those observations that were optimized for SFE. Also, there is some

modest evidence that the estimates of SFE are stable. As noted in the main body of the article, SFE has also been estimated when there are zero restrictions on the size of the classes or courses (SFEz). When SFEz is estimated, the number of courses looked at increases by 1,638, while the number of students increases by a more modest 208. The correlation between rkSFEz and the rkSFE15020, which stands for course size of at least 150 students and class size of 20 or more and reported as rkSFE in the body of the article, is 0.9701.