An Improved Grade Point Average, With Applications to CS Undergraduate Education Analytics

JONATHAN H. TOMKIN, MATTHEW WEST, and GEOFFREY L. HERMAN,
University of Illinois at Urbana-Champaign

We present a methodological improvement for calculating Grade Point Averages (GPAs). Heterogeneity in grading between courses systematically biases observed GPAs for individual students: the GPA observed depends on course selection. We show how a logistic model can account for course selection by simulating how every student in a sample would perform if they took all available courses, giving a new "modeled GPA." We then use 10 years of grade data from a large university to demonstrate that this modeled GPA is a more accurate predictor of student performance in individual courses than the observed GPA. Using Computer Science (CS) as an example learning analytics application, it is found that required CS courses give significantly lower grades than average courses. This depresses the recorded GPAs of CS majors: modeled GPAs are 0.25 points higher than those that are observed. The modeled GPA also correlates much more closely with standardized test scores than the observed GPA: the correlation with Math ACT is 0.37 for the modeled GPA and is 0.20 for the observed GPA. This implies that standardized test scores are much better predictors of student performance than might otherwise be assumed.

CCS Concepts: • **Social and professional topics → Computer science education**; *CS1*; *Computer engineering education*; *Gender*;

Additional Key Words and Phrases: Learning analytics, GPA, gender disparity, women in computing

ACM Reference format:
Jonathan H. Tomkin, Matthew West, and Geoffrey L. Herman. 2018. An Improved Grade Point Average, With Applications to CS Undergraduate Education Analytics. *ACM Trans. Comput. Educ.* 18, 4, Article 17 (September 2018), 16 pages.
https://doi.org/10.1145/3157086

1 INTRODUCTION

Grades are a ubiquitous measure of academic performance in undergraduate education in the United States. Grades are used in most high-stakes decisions in academia, reflecting the judgment of the institution itself on students' academic success (Cohen 2000; Rosovsky and Hartley 2002). Minimum grades are needed to receive credit for coursework, minimum grade point averages (GPAs) are needed to enter a major or graduate, and grades are used to determine eligibility for scholarships and awards. A high undergraduate GPA is required to be considered for prestigious graduate programs; a low GPA can reduce employment prospects. Grades awarded to students

This work was supported by the National Science Foundation under grants DUE 1347722 and CMMI 1150490.

Authors' addresses: J. H. Tomkin, Department of Geology, 1301 W. Green St., Urbana, IL 61801; M. West, Department of Mechanical Science and Engineering, 1206 W Green St, Urbana, IL 61801; G. L. Herman, Department of Computer Science, 201 N. Goodwin Ave., Urbana, IL 61801.

https://doi.org/10.1145/3157086

have real impacts, so it is important that they are awarded fairly and without discrimination. Further, because GPAs are so tied to these high-stakes decisions, we must be sure that this metric provides as accurate an estimate of a students' ability as it can. If not, biased grading in colleges and universities could be discouraging students from staying in challenging programs—such as Computer Science (CS)—and related careers. Worse yet, could these biases be disproportionately affecting traditionally underrepresented populations, such as women in CS?

Because GPAs are treated as one of our best estimators of a student's ability level, it is essential that we ascertain whether the GPA is as accurate as possible given students' grade data. One core challenge to the usefulness of the traditional GPA is that not all faculty assign grades using the same practices, nor do they assign grades in similar distributions. For example, Science, Technology, Engineering, and Mathematics (STEM) faculty are twice as likely to use fixed curves than other disciplines, artificially deflating grades in these courses and the GPAs of students who take those courses (Hurtado et al. 2012). These differences in grading practices have contributed to persistent disparities in grades given across fields of study (Johnson 2003). A multi-institutional study (Koester et al. 2016) has further highlighted that introductory-level STEM and CS courses apply "GPA penalties" to students who take them (i.e., students perform worse in these courses than their GPAs would predict). These disparities reveal that the GPA is an inherently flawed measure, measuring the grading policies used in the courses that a student takes in addition to that student's ability in those courses. Consequently, academic departments are making high-stakes decisions about students' future careers based on a flawed number.

Students likewise make high-stakes decisions based on grades and GPA. Early grades in STEM courses, including CS, strongly predict students' persistence in their majors (Cromley et al. 2016). Critically, women are disproportionately more likely to leave computer science because of low grades in introductory computer science than their male counterparts (Katz et al. 2006), and the negative effect of low grades is particularly amplified when students perform worse in their STEM courses than they expect based on their non-STEM courses (Ost 2010; Rask 2010; Seymour and Hewitt 1997; Stinebrickner and Stinebrickner 2011). As the demand for competent computing professionals continues to rise, our efforts to increase the number of CS graduates is mired by low retention rates, sparked in part by low grades (Strenta et al. 1994; Katz et al. 2006; Stout et al. 2011). Consequently, we have a critical need to extract as much information as possible from students' grade data to determine the best courses of action for improving student persistence, improving academic advising, and giving our students every competitive advantage.

Despite its shortcomings, the GPA remains a staple of academic decision making because of the ubiquitous and easy access to grades and the readily digestible nature of a single summative number. We seek to maintain these affordances of the traditional GPA while improving its validity by accounting for the disparities in grade distributions from different fields of study. In this article, we use logistic modeling of students' grades in courses relative to the program of study to create a new modeled GPA that accounts for the average grade distributions of the courses that students take. We show that the modeled grades are more accurate than the observed grades in predicting student performance, and that the modeled GPAs more strongly correlate with standardized test scores than observed GPAs. The primary contribution of this article is a new method that any institution or student can use to better predict how students will perform in each course based on their performance in other courses. This new modeled GPA could provide a powerful new tool for enhancing future data and learning analytics that rely on students' GPA.

To demonstrate how this new model can be applied, we explore three research questions inspired by the work of Koester et al. (2016): (1) Does a logistic model of GPA improve predictions of student performance over observed GPA? (2) If we can improve estimation of students' ability level using a logistic model, what does this new estimator reveal about the GPA penalties that students experience by taking an introductory CS curriculum? (3) How is the GPA penalty against women affected by this modeled GPA as compared to the traditional GPA? We have chosen gender as a useful case study as it is of high interest to the community and we have sufficient data to address the question. While we specifically explore the impact of this new modeled GPA's effect on GPA penalties, it could also be applied by institutions in other contexts such as in admissions decisions, academic advising, and even in helping students find careers after graduation.

2 BACKGROUND

GPA is designed to measure academic potential, but it contains a major flaw: students increase their apparent GPA if they choose courses that award higher-than-average grades, or decrease it by choosing courses that award lower-than-average grades. But how can we tell if one course awards higher or lower grades than another?

One approach is to compare the grade students receive in a particular course with their GPA. Lower-grade courses reduce the GPA of students who take them, while higher-grade courses raise them: the relative difficulty of individual courses is expressed as the difference between the average grades in a course and the average GPA of students taking that course. This difference is called the "grade point penalty" (Koester et al. 2016). If, for example, "B" students (overall GPA = 3.0) have an average grade of B- (score equivalent to 2.67) in a given course, then that course has a grade point penalty of 0.33. We can then use the grade point penalty to describe if a course is lower graded or higher graded than the average course. We note that this method does not consider the content covered in a course, let alone how much students learn—it is purely defined by recorded grades. This definition does match a colloquial understanding of course difficulty, though: a higher-graded course is more likely to pass a given student, and more likely to award an A, than a lower-graded one.

We can also use the grade point penalty concept to measure if different groups of students face different penalties for taking a course, potentially providing evidence of discrimination. If we expect male and female students to do equally well, it would also be expected that they would share the same grade point penalty across different courses. If they don't have the same grade point penalty (if, for example, women do less well in STEM courses), then this can be expressed as a difference in the grade point penalty. As an example, if male students have an average grade point penalty of 0.2 while female students have an average grade point penalty of 0.6 in the same course, then the "female grade point penalty" for this course is 0.4. If this pattern is observed in CS courses, then this could be evidence of discrimination against female students.

As described, the grade point penalty method is flawed as it uses observed course grades to compute the GPA. This approach seems uncontroversial, and is in fact the standard method used to determine GPAs in transcripts. But the observed GPA is itself dependent on the cumulative grade point penalties of the courses taken by the student. So a student who takes a large number of high-penalty courses will have an artificially depressed GPA, which will reduce the size of any observed grade point penalty. Imagine a student who, if she took all courses at an institution, averaged a "B"; her GPA should be 3.0. But no student takes all courses at an institution. If a "B" student exclusively enrolled in grade-penalizing courses that have an average grade point penalty of 0.33, she would have an observed GPA of 2.67—not 3.0, as specified. This in turn would impact the observed grade point penalty of these lower-grade courses: this student records no grade point penalty for taking these courses. Her observed GPA is 2.67, her average grade in these courses is 2.67, and so the apparent grade point penalty is zero—even though the actual grade point penalty was defined as 0.33.

Furthermore, we would expect this issue with observed GPAs to be universal to all students, regardless of program. As different majors have different curricula, which have different average grade point penalties, a student's GPA is partly determined by his or her course of study. It is rare for students in even the same program to select the same set of courses, and since every course has its own grading penalty, every individual course also impacts a student's observed GPA. We need a more accurate measure of GPA than the observed GPA if a grade point penalty approach is to work.

If every student took every course, then the impact of curriculum and course choice would disappear, as the GPA would then accurately reflect the student's academic performance relative to all other students in the sample. We therefore have developed a "modeled GPA" that aims to remove course-choice effects. We use a two-parameter logistic model for student grades in courses and use this to define the modeled GPA as the GPA that would be calculated if a student took, and received a grade for, all courses. This corrects the systemic bias in the calculation of the GPA, which in turn enables a more realistic calculation of the grade point penalty. This is similar to the approach of Vanderbei et al. (2014) but differs in that they used linear models of course grades, which can give unrealistic course-level predictions (i.e., course grades above 4 or below 0).

It should be noted that this grade point penalty approach assumes that there is a single factor that explains a large portion of student success in any given course, regardless of discipline—history is treated the same as economics, which is treated the same as computer science. As we will show below, this assumption is surprisingly reasonable. Having said this, we do not expect that GPA is the only important determinant in individual course performance. One would expect CS majors to do better than nonmajors in CS courses, for example, as they should have greater motivation and interest.

3 STUDENT GRADE DATA

The dataset used throughout this article consists of 1,984,111 student grade records from the College of Engineering and the College of Liberal Arts and Sciences at the University of Illinois at Urbana-Champaign over a period of 10 years (2006 to 2015 inclusive). These two colleges were chosen because computer science majors reside in both of these colleges. This dataset included 64,860 students and 3,606 courses, so the students had grades from an average of 30.6 courses and courses had an average of 550.2 students per course over this time period. We did not distinguish between different semesters of a course or different instructors, and the dataset included only those courses with at least 30 students and students with at least 10 courses; 3,277 of the students in the data are CS majors—425 female (15%), 2,851 male (85%), and one unspecified.

The dataset consists of N enrollment records, where record (i_n, k_n, g_n) indicates that student i_n took course k_n and received grade g_n, for $n = 1, \ldots, N$. It is possible that the same student took a given course multiple times and received either the same or different grades each time. There are a total of $I = 64{,}860$ students and $K = 3{,}606$ courses. Grades are measured on a standard 4-point scale, with $g = 0.0$ being the lowest grade (F) and $g = 4.0$ being the highest grade (A or A+). We denote by K_i the number of course records for student i, so that the observed GPA is

$$\text{observed-GPA}_i = \frac{1}{K_i} \sum_{\substack{n=1,\ldots,N \\ \text{such that} \\ i_n=i}} g_n. \tag{1}$$

4 LOGISTIC GRADE MODELS

Logistic models are a special case of generalized additive models (Hastie et al. 2009). They are widely used in statistical learning and modeling, including in psychometrics and educational settings where they are known as item response or latent trait models (Nering and Ostini 2010).

Predictive models of student grades have been used previously by Vanderbei et al. (2014) to estimate true student aptitude and course grade inflation with two-parameter linear models.

Our two-parameter logistic model for the predicted grade $\hat{g}_{ik}$ of student i in course k is

$$\hat{g}_{ik} = \frac{4}{1 + \exp(-a_k(\theta_i - b_k))}, \tag{2}$$

where there is one student parameter and two course parameters given by

$$\theta_i = \text{“ability” of student } i \tag{3a}$$

$$b_k = \text{“difficulty” of course } k \tag{3b}$$

$$a_k = \text{“discrimination” of course } k. \tag{3c}$$

The difficulty of a course is the ability level θ at which the logistic model crosses 2.0 (i.e., a grade of C). A student with this ability level will have a 50% chance of getting a grade of C or better in the course. The discrimination of a course indicates how strongly a course distinguishes between students of different ability levels (higher-discrimination courses provide more information about a student's ability level). For a set of parameters (Equation (3)) the root mean square error (RMSE) of the predicted grades is

$$e = \sqrt{\sum_{n=1}^{N} (g_n - \hat{g}_{i_n,k_n})^2}. \tag{4}$$

The optimal parameters θ_i^*, b_k^*, and a_k^* are determined by minimizing the error e over all parameter values. This can be done, for example, by an iterative procedure that begins by initializing student abilities θ_i to observed GPA scores and then alternates between finding the optimal course parameters while the student abilities are held fixed and finding the optimal student abilities while fixing the course parameters. That is, we alternate between

$$(a_k^*, b_k^*) = \underset{a_k, b_k}{\operatorname{argmin}} \sum_{\substack{n=1,\ldots,N \\ \text{such that} \\ k_n=k}} (g_n - \hat{g}_{i_n,k})^2 \text{ for } k = 1, \ldots, K \tag{5a}$$

and

$$\theta_i^* = \underset{\theta_i}{\operatorname{argmin}} \sum_{\substack{n=1,\ldots,N \\ \text{such that} \\ i_n=i}} (g_n - \hat{g}_{i,k_n})^2 \text{ for } i = 1, \ldots, I. \tag{5b}$$

We fitted the logistic model (Equation (2)) to the student grade dataset (see Section 3) using the iterative procedure (Equation (5)) and we computed the RMSE of the predictions. Figure 1 shows the model RMSE plotted against the RMSE that results from using the observed GPA as a predictor of course grades for each student. This data is then projected onto the difference in RMSE in Figure 2, which shows that the modeled GPA is a better predictor of student course grades than observed GPA for 83.6% of students.

The total RMSE of observed GPA and our logistic model (Equation (2)) as student course grade predictors are shown in Table 1. From this we see that observed GPA has a 21% higher RMSE than the logistic model.

For comparison, we also fitted the two-parameter linear model of Vanderbei et al. (2014) and compared it to our two-parameter logistic model, with results as shown in Table 1. This linear model has the form $\tilde{g}_{ik} = \tilde{a}_k\tilde{\theta}_i + \tilde{b}_k$, where $\tilde{g}_{ik}$ is the predicted grade for student i in course k, $\tilde{\theta}_i$

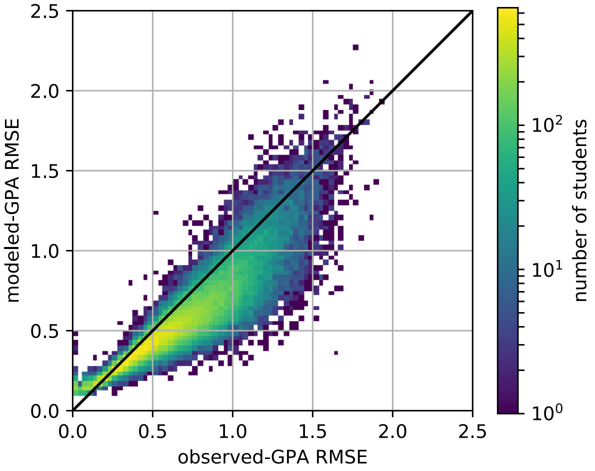

Fig. 1. The RMSE (root mean square error) of the logistic model (Equation (2)) as a predictor of student grades (vertical axis), plotted against the RMSE of the observed GPA as a predictor (horizontal axis). Points below the 45° line indicate that the model (Equation (2)) is a better predictor than observed GPA.

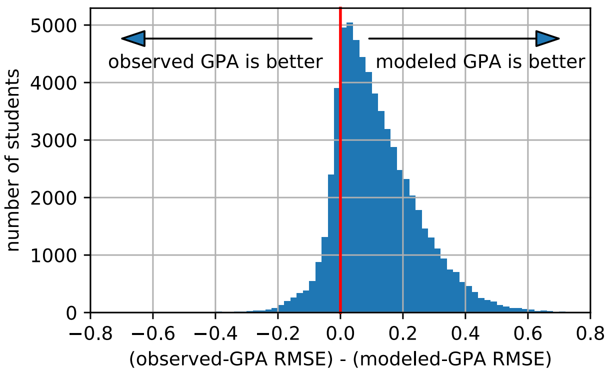

Fig. 2. The difference between the RMSE (root mean square error) of observed GPA and the logistic model (Equation (2)) as a predictor of student grades. Values to the left of center indicate that the model (Equation (2)) is a better predictor than observed GPA, while points to the right indicate the reverse. This plot is a projected view of the data in Figure 1 and shows that the model (Equation (2)) is a better overall predictor of student grades than observed GPA.

Table 1. RMSE (Root Mean Square Error) of All Student Course Grade Predictions from the Three Different Prediction Methods

Prediction Method	RMSE
Observed GPA	0.737
Logistic model (2)	0.609
Linear model of Vanderbei et al. (2014)	0.611

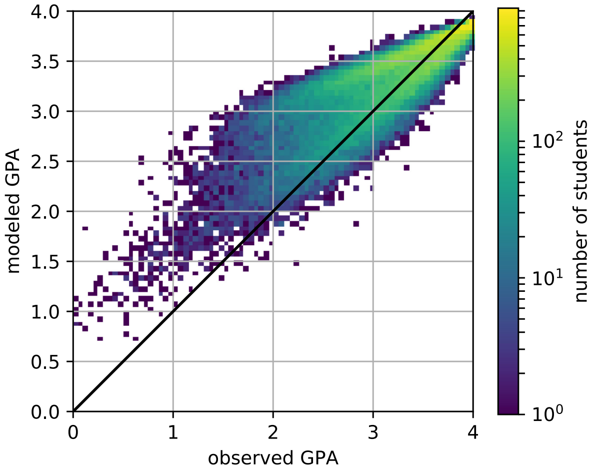

Fig. 3. Observed GPA plotted against modeled GPA (Equation (6)) for each student.

is the ability of student i, $\tilde{a}_k$ is the discrimination of course k, and $\tilde{b}_k$ is the difficulty of course k. While very similar to our logistic model on this dataset, the linear model had a 0.4% higher error than the logistic model. When used to predict course grades, our logistic model was better (lower RMSE) than the linear model for 53.2% of students. We interpret this as weak but positive evidence for the superiority of the logistic model for this dataset, given that both models have the same number of parameters and the logistic model has the additional advantage that it can only make plausible predictions (i.e., cannot predict below zero or above 4.0).

5 MODELED GPA AS AVERAGE PREDICTED GRADE OVER ALL COURSES

To remove the effect of course choice on GPA for student i, we first compute their predicted grade $\hat{g}_{ik}$ in all courses $k = 1, \ldots, K$. We then define their modeled GPA to be their GPA taken over all courses using these predicted grades:

$$\text{modeled-GPA}_i = \frac{1}{K} \sum_{k=1}^{K} \hat{g}_{ik}. \tag{6}$$

Because the modeled GPA is computed over all courses, all students are being compared on the basis of the same set of courses (i.e., all of them). That makes modeled GPA a fairer metric of student ability than observed GPA, which is heavily influenced by course choice.

Using the dataset described in Section 3, we computed the observed and modeled GPAs for all students and plotted them against each other as shown in Figure 3. We see that the modeled GPA tends to be somewhat higher than the observed GPA, especially for students with very low GPAs. This finding suggests that students with low GPAs also took lower-graded courses, partially explaining their low GPA. The model predicts that these students would do better if they took higher-graded courses. If CS students are generally taking lower-graded courses, this type of finding may inform academic advising to help students persist through what may be artificially low GPAs based on course selection effects.

6 EXAMPLE: UNDERGRADUATE STUDENTS IN CS

We have shown in the previous section that the modeled GPA developed more accurately predicts student grades than the traditional GPA, so it is a better metric for learning analytic approaches that require the use of student GPAs. We illustrate how these improved predictions significantly, and systemically, affect the observations we can make with GPA data in learning analytics. Specifically, we contrast findings generated from using observed GPAs with those generated from using modeled GPAs for three cases. As an additional benefit, these example analyses provide insight into the likelihood (or not) of gender bias in grading in CS courses at one institution. We choose gender as a case study in part because there are enough women in the sample to have confidence in the results. Other analyses are possible, including examinations of race, geographic origin, and social-economic status.

The following analyses highlight the importance of taking individual student course choices into consideration when using grade data by examining the grade point penalties of taking CS courses/curricula or being a female in those courses/curricula. We present evidence at different scales and timing. In this section, we begin by focusing on the performance of all students (not just CS majors) in the first-year courses required for CS majors. We then examine the performance of CS majors across all of their courses as well as in the CS curriculum specifically. Finally, we conclude by examining correlations between the different GPA measures and standardized test scores.

6.1 Penalty of Required First-Year CS Courses

As noted in the background section, grades in the first courses of a major deeply impact whether students perceive they can succeed in that major. If a particular group of students underperform expectations in introductory CS courses, then this may dissuade them from embarking on a CS degree, leading to underrepresentation. Consequently, we begin by examining the general student population, to see how introductory courses in the CS curriculum are graded. This dataset encompasses students in the College of Engineering and the College of Liberal Arts and Sciences, because CS majors come from both colleges at our institution. We then disaggregate this data by gender to examine whether students receive different grade point penalties based on gender.

The undergraduate CS program at Illinois requires a total of 128 credit hours to complete, with required courses in CS, mathematics, and physics, as well as a variety of general education and CS elective courses. We first examine how all students, from all majors, perform in the required STEM courses in the CS program, as the relative performance in these courses is important in determining student persistence (Ost 2010; Rask 2010). These courses consist of the calculus sequence (MATH 220/221, 231, 241), linear algebra (MATH 415), and physics (PHYS 211 and 212), as well as a computer science introductory core (CS 125 - CS1, 173 - discrete mathematics, and 225 - CS2). Students rarely take all 10 of these courses (MATH 220 and 221 both count as calculus 1, and students often have Advanced Placement credit for math and physics), but all CS and prospective CS students must have credit for these courses to graduate.

As can be seen in Figure 4, these courses are not easy. The average grade point penalty calculated when using the observed GPA (light blue bars) for these courses is +0.26; of these 10 courses, only CS 125 has a negative penalty (i.e., is "good" for a student's GPA). Furthermore, using the observed GPA actually underpredicts the difficulty of these courses. The observed GPA doesn't take into account that the STEM majors in general, and CS majors in particular, are taking these challenging courses and so have depressed GPAs as a result. As described previously, these depressed GPAs lower the observed penalties. The modeled GPA grade point penalties (dark blue bars) are more than twice as large, at +0.55, and none of the grade point penalties calculated in this way are negative. This implies that all of these courses lower, on average, the student's GPA, with an average penalty of over half a letter grade.

These results are very similar to STEM courses as a whole. If we examine all large (5,000 or more student records) STEM courses for majors, we find that using the observed GPA to calculate the

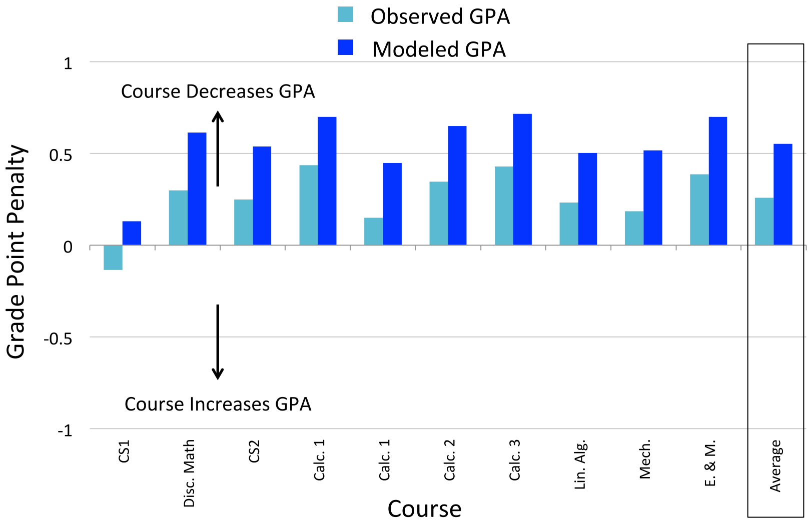

Fig. 4. Grade point penalties for all students who take the introductory courses in the CS curriculum. Positive values indicate that students get lower grades in these courses than they do in other courses. A grade point penalty of 1 would mean that a student would average a full letter grade lower in this course relative to his or her other courses. Without exception, using the observed GPA understates the GPA penalty. Note that CS1 has the rubric CS 125, Disc. Math is CS 175, CS2 is CS 225, Calc. 1 is MATH 220/221, Calc. 2 is MATH 231, Calc. 3 is MATH 241, Lin. Alg. is MATH 285, Mech. is PHYS 211, and E. & M. is PHYS 212. Error bars showing the Standard Error of the Mean (SEM) are omitted from the chart for clarity; in all cases they are 0.01 or less. The SEM of the average is 0.003, calculated using the weighted average of the standard deviations of the individual courses. These results and other popular STEM classes, with student numbers, are tabulated in the appendix.

grade point penalty yields an average of +0.24, and the modeled GPA grade point penalty average is +0.49. This broader result includes courses from biology, chemistry, and other engineering disciplines, as well as math, physics, and CS (see table in the appendix).

We also examined these courses for differences in GPA penalties according to gender (Figure 5): do women do relatively worse in these courses than men with equivalent GPAs? Gender-based differences in grade point penalties are relatively small compared to their absolute size (Figure 4) and can be positive or negative. In this case, the observed GPA (pink bars) yielded an average penalty difference between women and men of −0.15 (i.e., female students with the same GPA as male students did 0.15 grade points worse). In this case, the observed GPA systematically overestimates the penalty difference that women receive. The male/female grade point penalties calculated from the modeled GPA (red bars) are all less biased against women. Six of the 10 courses have positive penalty differences (i.e., men do relatively worse than women in these courses) and four are negative. The average penalty difference is −0.04 grade points. This is evidence that gender-based differences in course choice systematically bias the observed GPAs of men and women. Furthermore, if we examine the three CS courses in isolation, we find no gender penalty for women, who have a positive (and very small) grade point penalty difference of +0.01. Interestingly, the two physics courses in the sample are almost entirely responsible for the penalty differences between women and men, with the two courses having male/female grade point penalty differences of −0.26 and −0.22 (i.e., women have larger grade point penalties than men in these two courses). This is similar to the results found by Koester et al. (2016), who used observed GPA to calculate the grade penalty.

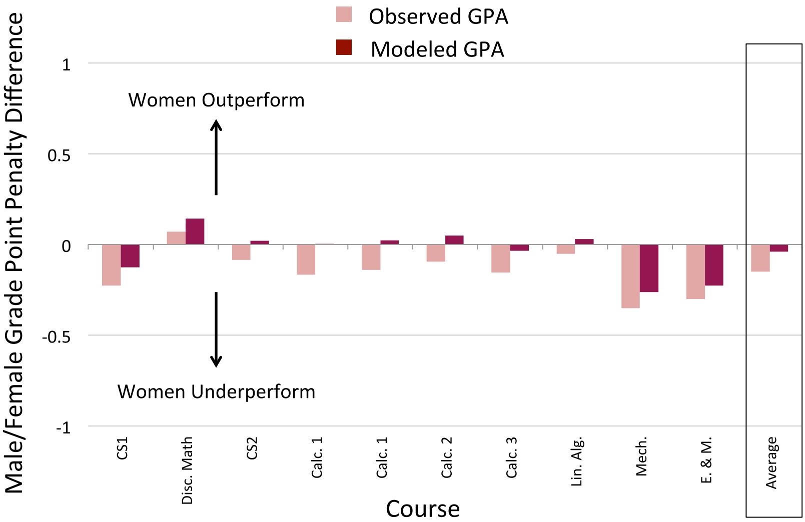

Fig. 5. Difference in grade point penalties for women and men in introductory courses in the CS curriculum. Positive values indicate that women relatively outperformed men. A grade point penalty difference of 1 would mean that a female student would average a full letter grade higher in this course relative to a male student with the same overall GPA. Using the observed GPA to calculate the grade point penalty consistently lowers the apparent female performance. Including course composition effects (red bars) reduces the average GPA difference in this sample of courses from −0.15 to −0.04. Note that CS1 has the rubric CS 125, Disc. Math is CS 175, CS2 is CS 225, Calc. 1 is MATH 220/221, Calc. 2 is MATH 231, Calc. 3 is MATH 241, Lin. Alg. is MATH 285, Mech. is PHYS 211, E. & M. is PHYS 212. Error bars showing the Standard Error of the Mean (SEM) are omitted from the chart for clarity; in all cases they are 0.01 or less. The SEM of the average is 0.003, calculated using the weighted average of the standard deviations of the individual courses. These results, with student numbers, are tabulated in the appendix.

6.2 Penalty of CS Courses

Alternatively, we may be simply concerned with how students perform in all introductory CS courses, as students may use their performance in either a major or a nonmajor CS course to inform their decision on whether to major in the field. If we examine the results for the five largest CS courses alone (CS 101 - CS1 for engineering majors, 105 - CS1 for nonengineering, non-CS majors, 125 - CS1 for CS majors, 173 - discrete mathematics, and 225 - CS2; data in the appendix), we find that the (adjusted) female/male GPA penalty is small in magnitude and mixed in sign, with three courses favoring women and two favoring men. The average male/female grade point penalty difference calculated from the modeled GPA for these five courses is zero (0.00).

This result is consistent with how CS-related majors perform in these courses. A previous survey of this dataset using observed GPAs (Tomkin et al. 2016) found that women did not do worse than men in CS and related majors across 12 introductory STEM courses. Female majors in Mathematics and Computer Science, and in Computer Engineering, have higher average scores than men with equivalent GPAs: the average difference in grade point penalty between men and women is 0.14 ± 0.03 and 0.06 ± 0.02 grade points, respectively. Female Computer Science majors have the same grade point penalty as men in these courses (−0.01 ± 0.01). The students in the sample who graduated with Computer Science degrees have an average observed GPA of 3.21 and an average modeled GPA of 3.46; on average, the courses that CS students take over the course of their degree lower their recorded GPA by a quarter of a grade point.

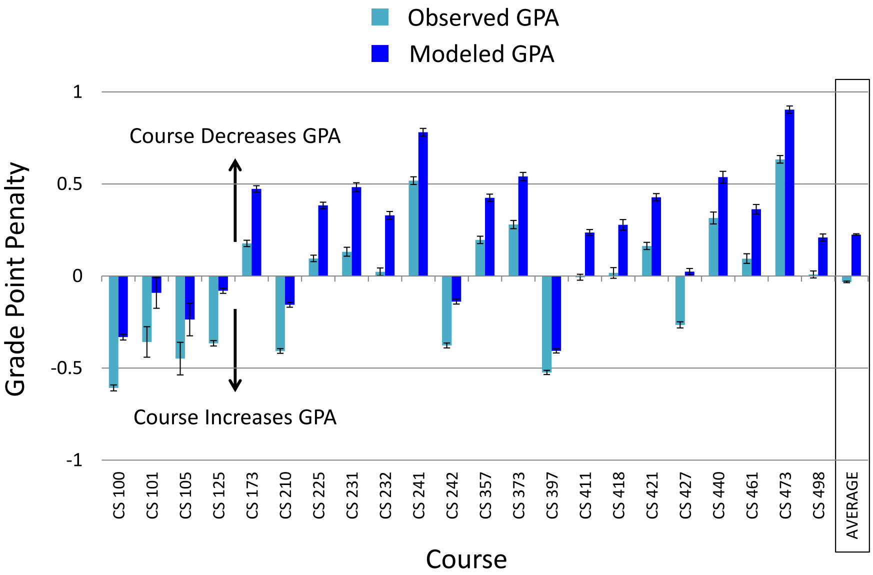

Fig. 6. Grade point penalties of CS majors for CS courses. Positive values indicate that students get lower grades in these courses than they do in other courses. A GPA penalty of 1 would mean that a student would average a full letter grade lower in this course relative to his or her other courses. Without exception, using the observed GPA understates the GPA penalty. The average grade point penalty (weighting all courses equally) is −0.03 for the observed data with a 95% CI [−0.02, −0.04], and +0.23 for the modeled data with a 95% CI [−0.02, −0.04]. The error bars show +/− 1 Standard Error of the Mean (SEM). The SEM of the average (0.004) is calculated using the weighted average of the standard deviations of the individual courses. These results, with course names and student numbers, are tabulated in the appendix.

This CS grade point penalty is reflected in the grades awarded in all CS courses, as shown in Figure 6. The average grade point penalty determined from the observed GPA is small and negative (−0.03), while the average grade point penalty determined from the modeled GPA is large and positive, at 0.23. The gender difference in the grade point penalty, shown in Figure 7, is usually negative (i.e., women have higher grade point penalties than men). Again, using the observed GPA overestimates the difference: the average across all courses shrinks from −0.09 to −0.05 when the modeled GPA is used to make the calculation. Of the 22 courses, women do better than expected, relative to men, in nine, and do worse than expected, relative to men, in 13.

6.3 GPA and Standardized Test Scores

The grades of CS majors are somewhat predicted by standardized test scores: we find that the modeled GPA of CS majors and their ACT composite score has a correlation coefficient of 0.34. ACT math scores are slightly more predictive of GPA, with a correlation coefficient of 0.37. Note the the ACT math scores have a weaker correlation with the observed GPA (0.20)—higher-scoring students are more likely to enroll in STEM courses, which penalize GPAs. The standardized test scores of CS students do not differ greatly by gender: average composite ACT scores, ACT math scores, and SAT totals for female CS majors are 31.22, 32.00, and 1, 362; male averages are 31.35, 32.61, and 1, 386.

This gender difference in the CS majors' GPA penalty appears to be partly related to characteristics of students captured by these standardized test scores. If we redo the GPA penalty calculation shown in Figure 7 but only include students with ACT math scores of 34 and higher (the 99th

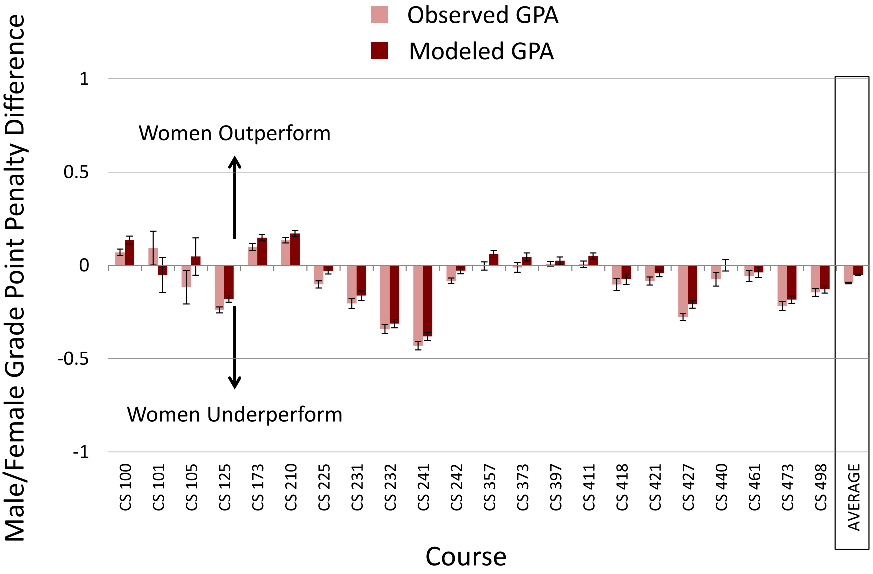

Fig. 7. Difference in grade point penalties of CS majors for women and men in CS courses. Positive values indicate that women relatively outperformed men. A grade point penalty difference of 1 would mean that a female student would average a full letter grade higher in this course relative to a male student with the same overall GPA. The average grade point penalty difference (weighting all courses equally) is −0.09 for the observed data with a 95% CI [−0.09, −.1], and −0.05 for the modeled data with a 95% CI [−0.04, −0.06]. The error bars show +/− 1 Standard Error of the Mean (SEM). The SEM of the average (0.004) is calculated using the weighted average of the standard deviations of the individual courses. These results, with course names and student numbers, are tabulated in the appendix.

percentile of all students who take the math ACT), we find that the GPA penalty difference shrinks and changes sign (to +0.02).

There is a large gender imbalance in the number of students completing CS bachelor degrees in the dataset, which is produced by a large gender imbalance in the gender of incoming students. Of the 3,277 students whose last declared major is CS, 2,851 were male and 425 were female: 13% of finishing CS majors were female at this institution. It is possible that the proportion of women in courses impacts their performance in those courses. Unfortunately, it's not clear from this data if this is the case, as there is insufficient spread in the female/male fractions. Almost all CS courses matched, to within a few percent points, the overall proportion of female CS majors (13%). The one course that is close to gender parity was CS 105, a nonmajors course that has 41% female enrollment, and in which women outperformed men by 0.05 grade points when using modeled GPA in the calculation (Table 3).

7 DISCUSSION AND CONCLUSIONS

The GPA is an attractive way to measure student college success; it's a single-number metric, the majority of institutions use the same system (so results can be compared widely), and grades reflect a judgment by the institution itself as to the academic aptitude of students. Furthermore, all institutions have access to grade data, and so can readily use individual course grades and GPAs to analyze student performance across the curriculum. This analysis must be done with care, however: although one might expect that grades will differ between institutions, it is clear from the results here that grades systematically vary within institutions, by course and by program. The overall GPA of a student is dependent on each individual student's course of study.

Different courses of study have different GPA penalties, so the observed overall GPA is flawed as a yardstick of student academic potential: a student will increase his or her apparent GPA if he or she chooses a course of study made up of many low (or positive) grade point penalty courses. Insidiously, if we use observed GPA to try to detect this problem (by examining the difference between average GPA and average course grade), we greatly underestimate the size of the penalty, as students in lower-graded courses cluster (CS students, for example, are all required to take a large number of high-penalty courses).

The logistic model described here accounts for course heterogeneity, producing a "modeled GPA" that much more accurately predicts student performance. The modeled GPA consistently better predicts how well students perform in the courses taken than the observed GPA, with a significantly lower root mean square error between predicted and actual grades. The logistic model also performs better (has a lower error) than a linear model. The logistic model does so by making use of all of the available data (all courses and students are included) and does not require any subjective selection of what to include. It also enables us to get a better idea of the size of grading variation across courses, as we can now use the modeled GPA to determine course grading penalties.

We find that the average grade point penalties of the introductory CS courses and introductory STEM courses sampled in this study are 0.55 and 0.49, respectively. This implies that students lower their GPA by about half a letter grade when they take only introductory courses in CS and STEM—possibly dissuading some students from continuing in the CS program. Note that calculations that use the observed GPA to determine the grade point penalty significantly underestimate this effect (Figure 4). The problem of using the observed GPA to calculate penalties is underlined when we survey CS classes as a whole: it would suggest that taking CS courses has no effect on your GPA (Figure 7), whereas in fact they reduce your GPA by about a quarter of a grade point.

As a consequence, CS students suffer from observed GPAs that are lower than their actual academic performance warrants. We find that the overall grade point penalty for CS majors across all of their courses is 0.25—a quarter of a letter grade. Enrolling in a CS major is bad for your GPA.

Interestingly, gender differences in course penalties change when using the modeled GPA rather than the observed GPA. In the courses examined here, using the observed GPA to calculate the gender penalty systematically overestimated the impact of gender on expected course performance. This suggests that male and female students choose different patterns of courses, even when they share the same major (we see the same pattern both for all students and for CS majors only). The reasons for this difference in course selection are worthy of further study. Although the (modeled GPA) grade point penalty often showed that women underperformed expectations relative to men, changing the specifications altered this: restricting the sample to the highest-scoring students reversed the sign of the effect. Furthermore, women did better than expected in individual courses about as often as doing worse than expected. If gender is a factor in CS academic achievement in this dataset, it appears that it is relatively small, and is of uncertain sign. There was no treatment of the data that resulted in a programmatic GPA gender difference larger than 0.05 of a point.

We avoid making strong statements about the presence or absence of discrimination in this article. Readers should be wary of "researcher degrees of freedom" (Simmons et al. 2011) when assessing claims about the statistical significance of studies such as this one—there are many ways in which the method of partitioning the data (by deciding what courses to include in the analysis, for example) has the potential to change the significance of the findings. It is clear, however, that any study that does not control for heterogeneity in course grades may produce significantly biased results.

The modeled GPA generated by the model is a surprisingly good predictor of course performance across all courses examined, which includes humanities and social sciences as well as STEM courses. There are, for example, no obvious bimodal features in the error between observed and predicted grades, which might be evidence of discipline-specific aptitudes. This suggests that a single dimension predicts much of the variation in student success in college courses. This single dimension is well characterized by standardized tests for this dataset; the correlation coefficient between the ACT math scores and modeled GPA is 0.37, for example. Note also that standardized tests are better predictors of student academic potential than studies that use observed GPA would suggest: the correlation coefficient between math ACT and observed GPA is just 0.20 for the same data. This difference in correlations is the result of the course selection bias discussed in this article: students with higher test scores are more likely to be enrolled in courses with greater grade point penalties, which depresses their observed GPA and weakens the correlation between standardized tests and observed GPA. The result found here suggests that studies that do not take into account course grading heterogeneity will significantly underpredict the correlation between standardized tests and GPA.

Grades and GPA are ubiquitous in assessing student performance, both formally and in education research. This article describes the problem that course grading heterogeneity presents in making use of observed grades and GPA in several learning-analytic applications. It also shows how a logistic model can produce a modeled GPA that can be used to overcome the problems associated with course grading heterogeneity and enable researchers and educators to better assess student performance.

APPENDIX

The observed GPA data for the large STEM courses and all CS courses discussed in the article is shown in Tables 2 and 3.

Table 2. Data for Large STEM Courses Used in Campus Analysis Section and Figures 4 and 5

Course Rubric	Number of Grade Records	Observed GPA Penalty	Modeled GPA Oenalty	Observed GPA Male-Female Penalty	Modeled GPA Male-Female Penalty
CHEM 102	22,356	0.545	0.797	−0.360	−0.203
CHEM 103	23,665	−0.389	−0.156	0.009	0.174
CHEM 104	13,815	0.493	0.707	−0.270	−0.144
CHEM 105	15,107	−0.102	0.087	−0.032	0.098
CHEM 232	9,177	0.510	0.682	−0.276	−0.202
CHEM 233	7,880	0.187	0.321	−0.096	−0.029
CS 101	9,481	0.053	0.320	−0.098	0.001
CS 105	6,448	0.115	0.197	−0.207	−0.058
CS 125	5,373	−0.135	0.130	−0.227	−0.126
CS 173	5,301	0.300	0.614	0.071	0.143
CS 225	6,464	0.249	0.539	−0.085	0.021
ECE 110	6,635	0.047	0.420	−0.120	−0.042
ECE 205	5,104	0.223	0.472	−0.040	−0.014
IB 150	8,037	0.007	0.161	−0.167	−0.073

(Continued)

Table 2. Continued

Course Rubric	Number of Grade Records	Observed GPA Penalty	Modeled GPA Oenalty	Observed GPA Male-Female Penalty	Modeled GPA Male-Female Penalty
MATH 220	8,464	0.435	0.699	−0.168	0.000
MATH 221	8,137	0.151	0.448	−0.141	0.022
MATH 225	6,336	0.157	0.460	0.020	0.119
MATH 231	17,682	0.347	0.648	−0.095	0.050
MATH 241	20,172	0.429	0.716	−0.154	−0.035
MATH 285	9,469	0.194	0.451	0.078	0.138
MATH 415	11,331	0.234	0.503	−0.052	0.030
MCB 150	10,201	0.366	0.537	−0.247	−0.156
MCB 250	5,582	0.687	0.854	−0.224	−0.166
MCB 251	5,582	−0.033	0.116	−0.125	−0.068
PHYS 211	17,329	0.186	0.516	−0.351	−0.262
PHYS 212	16,157	0.387	0.700	−0.301	−0.227
PHYS 213	10,962	0.481	0.771	−0.203	−0.129
PHYS 214	11,995	0.466	0.755	−0.244	−0.180
STAT 400	5,115	0.199	0.445	0.049	0.163
TAM 212	6,251	0.482	0.744	−0.133	−0.099

Table 3. Data for CS Courses as Displayed in Figure 7

Course Rubric	Course Name	Number of Grade Records	Observed GPA Penalty	Modeled GPA Penalty	Observed GPA Male-Female Penalty	Modeled GPA Male-Female Penalty
CS 100	Freshman Orientation	1,600	−0.61	−0.33	0.07	0.14
CS 101	Intro Computing: Engrg & Sci	73	−0.36	−0.09	0.09	−0.05
CS 105	Intro Computing: Non-Tech	53	−0.45	−0.24	−0.12	0.05
CS 125	Intro to Computer Science	2,162	−0.37	−0.08	−0.24	−0.18
CS 173	Discrete Structures	2,278	0.18	0.47	0.10	0.15
CS 210	Ethical & Professional Issues	2,040	−0.41	−0.16	0.13	0.17
CS 225	Data Structures	2,390	0.10	0.38	−0.10	−0.03
CS 231	Computer Architecture I	1,151	0.13	0.48	−0.20	−0.16
CS 232	Computer Architecture II	1,218	0.02	0.33	−0.34	−0.31
CS 241	System Programming	2,462	0.52	0.78	−0.43	−0.38
CS 242	Programming Studio	1,911	−0.38	−0.14	−0.08	−0.03
CS 357	Numerical Methods I	1,622	0.20	0.42	0.00	0.06
CS 373	Theory of Computation	1,323	0.28	0.54	−0.01	0.04
CS 397	Individual Study	786	– 0.52	– 0.41	0.01	0.03
CS 411	Database Systems	1,494	– 0.01	0.24	0.01	0.05
CS 418	Interactive Computer Graphics	786	0.02	0.28	−0.10	−0.07
CS 421	Progrmg Languages & Compilers	1,909	−0.27	0.43	−0.08	−0.04
CS 427	Software Engineering I	974	−0.27	0.02	−0.28	−0.21
CS 440	Artificial Intelligence	703	0.32	0.54	−0.07	0.00
CS 461	Computer Security I	751	0.09	0.36	−0.06	−0.04
CS 473	Algorithms	1,683	0.63	0.90	−0.22	−0.18
CS 498	Special Topics	1,731	0.01	0.21	−0.14	−0.13

ACKNOWLEDGMENTS

Thanks to Lin Fan and Debbie Dilman for their invaluable help in gathering and formatting the data for this analysis. Thanks to the reviewers for insightful comments that greatly improved the article.

REFERENCES

W. D. Cohen. 2000. The grade point average (GPA): An exercise in academic absurdity. *National Teaching & Learning Forum* 9, 5 (2000), 1–4.

J. G. Cromley, T. Perez, and A. Kaplan. 2016. Undergraduate STEM achievement and retention: Cognitive, motivational, and institutional factors and solutions. *Policy Insights from the Behavioral and Brain Sciences* 3, 1 (2016), 4–11.

T. Hastie, R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning*. Springer-Verlag. DOI: http://dx.doi.org/10.1007/978-0-387-84858-7

S. Hurtado, M. K., Eagan, J. H. Pryor, H. Whang, and S. Tran. 2012. *Undergraduate Teaching Faculty: The 2010–2011 HERI Faculty Survey.* Higher Education Research Institute, UCLA, Los Angeles, CA.

V. E. Johnson. 2003. *Grade Inflation: A Crisis in College Education.* Springer-Verlag, New York.

S. Katz, D. Allbritton, J. Aronis, C. Wilson, and M. L. Soffa. 2006. Gender, achievement, and persistence in an undergraduate computer science program. *SIGMIS Database* 37, 4 (2006), 42–57. DOI: http://dx.doi.org/10.1145/1185335.1185344

B. P. Koester, B. G. Galina, and T. A. McKay. 2016. Patterns of gendered performance difference in introductory STEM courses. *Arxiv Preprint.* DOI: http://dx.doi.org/arXiv:1608.07565

M. L. Nering and R. Ostini (Eds.). 2010. *Handbook of Polytomous Item Response Theory Models.* Routledge.

B. Ost. 2010. The role of peers and grades in determining major persistence in sciences. *Economics of Education Review* 29, 6 (2010), 923–934.

K. Rask. 2010. Attrition in STEM fields at a liberal arts college: The importance of grades and pre-collegiate preferences. *Economics of Education Review* 29, 6 (2010), 892–900.

H. Rosovsky and M. Hartley. 2002. *Evaluation and the Academy: Are We Doing the Right Thing? Grade Inflation and Letters of Recommendation.* American Academy of Arts & Sciences, Cambridge, MA.

E. Seymour and N. M. Hewitt. 1997. *Talking About Leaving: Why Undergraduates Leave the Sciences.* Westview Press, Boulder, CO.

J. P. Simmons, L. D. Nelson, and U. Simonsohn. 2011. False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22, 11 (2011), 1359–1366.

T. R. Stinebrickner and R. Stinebrickner. 2011. *Math or Science? Using Longitudinal Expectations Data to Examine the Process of Choosing a College Major.* National Bureau of Economic Research, Cambridge, MA.

J. G. Stout, N. Dasgupta, M. Hunsinger, and M. A McManus. 2011. STEMing the tide: Using ingroup experts to inoculate women's self-concept in science, technology, engineering, and mathematics (STEM). *Journal of Personality and Social Psychology* 100, 2 (2011), 255–270. DOI: http://dx.doi.org/10.1037/a0021385

A. C. Strenta, R. Elliott, R. Adair, M. Matier, and J. Scott. 1994. Choosing and leaving science in highly selective institutions. *Research in Higher Education* 35, 5 (1994), 513–547.

J. Tomkin, M. West, and G. L. Herman. 2016. A methodological refinement for studying the STEM grade-point penalty. In *46th Annual Frontiers IEEE Frontiers in Education Conference (FIE'16).*

R. J. Vanderbei, G. Scharf, and D. Marlow. 2014. A regression approach to fairer grading. *SIAM Review* 56, 2 (2014), 337–352. DOI: http://dx.doi.org/10.1137/12088625X

Received September 2016; revised April 2017; accepted October 2017