A Methodological Refinement for Studying the STEM Grade-Point Penalty

Jonathan Tomkin
University of Illinois at Urbana-Champaign
Champaign, IL, USA
tomkin@illinois.edu

Matthew West, Geoffrey L. Herman,
University of Illinois at Urbana-Champaign
Urbana, IL, USA
mwest, glherman@illinois.edu

***Abstract*— We present a study that explores the grade-point average (GPA) penalties that students face when taking introductory STEM courses. Previous work has found that there is a large and significant grade point penalty for women and minorities in most STEM classes (that is, these students perform worse in these classes than their overall GPA would suggest). We recreated this work using a new, large data set (63,012 students over 10 years) of student performance, and found that the initial results held when using the original approach. We argue that there are methodological shortcomings to the original approach, however, as there is no attempt to control for individual student program difficulty (STEM majors and non-STEM majors share some classes, but have very different overall suites of courses that determine their overall GPA). As the female/male and racial ratios vary across majors it is therefore likely that a division by gender is not comparing equivalent sample populations. By controlling for student test scores or major most of the penalty is removed. The initial findings of large GPA penalties in STEM courses appears to be an example of "Simpson's Paradox".**

Keywords— grade point penalty; engineering; gender; race

I. INTRODUCTION

Women, African-Americans, and Hispanic-Americans are underrepresented in math, engineering, and physical science undergraduate programs in higher-education [1-3]. Differential treatment of these underrepresented populations may further dissuade these students from persisting in these fields, necessitating that we explore various factors in the undergraduate environment such as the effect of instructors, fellow students, or the curricula [1, 3].

Students' grades in their first STEM courses strongly predict students' persistence in STEM majors, signaling to students about whether they belong [4]. Some research suggests that these introductory courses are disproportionately pushing women and minorities out of STEM careers [4, 5]. Consequently, we need measures that can identify bias in STEM teaching practices. In this paper, we present a new study that explores the grade-point average (GPA) penalties that students face when taking STEM courses.

If introductory STEM coursework is a factor in the relative lack of diversity in the field, it should be possible to detect the impact of the undergraduate environment by examining large data sets of student performance. Do underrepresented groups become more underrepresented over the course of the program, (i.e. is there a differential in the rate of retention in STEM)? Do underrepresented groups underperform in STEM courses, relative to their non-STEM courses? If found, such a result would be highly suggestive: although correlations between race or gender and achievement are not sufficient to indicate some type of discrimination, they are necessary. We need to establish under what circumstances these patterns exist.

In this study, we examine a large data set (63,012 individual students) for evidence of differential student outcomes. The data includes all students majoring in the Colleges of Engineering and the College of Liberal Arts and Sciences at the University of Illinois in Urbana-Champaign between 2005 and 2015. This study was motivated by a Committee on Institutional Cooperation (CIC) Learning and Research Analytics Meeting workshop held at the University of Michigan on November 13, 2014 that encouraged CIC members to perform data-based analysis of their student's performance in STEM fields. Initial findings at the University of Michigan found that women underperformed in physics courses, relative to expectations [5]. By underperform, we mean that students perform worse in these classes than their overall GPA would suggest. The GPA penalty is calculated by comparing the average difference between a student's overall GPA and their letter grade in a course. If, for example, a "B" student (overall GPA=3.0) scores a B- (GPA equivalent of 2.67) on average in a given course, then that course has a "grade point penalty" of 0.33. The female grade point penalty works similarly. If we compare male and female students that have the same overall GPA we would expect the same performance in STEM courses if there is no bias. This may not be the case [5]. If, for example, male students have an average grade point penalty of 0.2 in a class while female students have an average penalty of 0.6, then the "female grade point penalty" for this class is 0.4.

This work was supported by the College of Engineering and the National Science Foundation grant DUE 1347722.

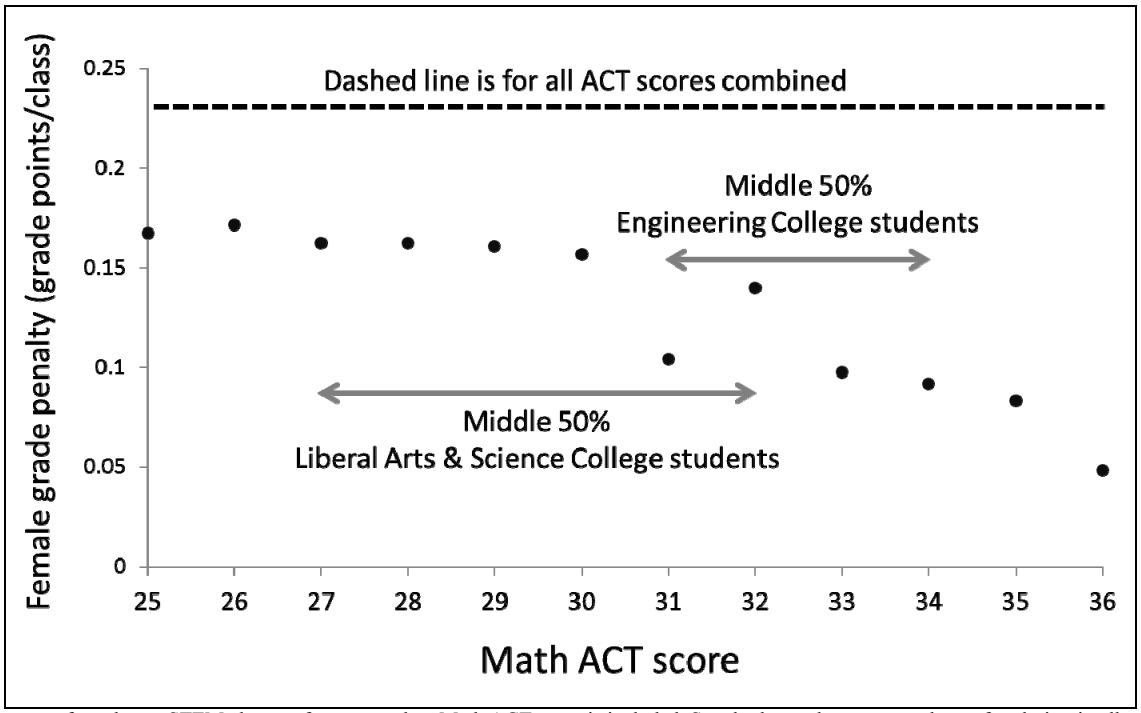

Fig 1. Impact of gender on STEM class performance when Math ACT score is included. Standard error bars are not shown for clarity; in all cases they are smaller than 0.01 of a grade point. The Engineering College is, on average, more selective than the Liberal Arts & Sciences College, as evidenced by the difference in median incoming Math SAT scores. Note how the female grade penalties are lower for all the individual scores than the average. This surprising result is evidence for Simpsons paradox, and arises from compositional differences in the different Math ACT grades.

We replicated the earlier study and found broad agreement: female students, and underrepresented minority students, underperformed in STEM classes relative to their GPA. However, we also observed that this underperformance was mitigated when controlling for a single relevant observable (e.g., major or math ACT score). This observation suggests that the original finding is at least partly, and perhaps wholly, due to "Simpson's Paradox" (as observed in other higher education settings, [6]): measured differences in gender or racial outcomes are largely the result of differences in the composition of different college majors, and not the result of any discrimination in the courses themselves.

II. Methods and Results

A. Data Set and core STEM classes

The data set contained academic performance data for all 63,012 undergraduate students enrolled in the College of Liberal Arts and Sciences and the College of Engineering at the University of Illinois 2005-2015. The data was anonymized and all students were given a randomized unique identifier. The data set includes students' grades in all courses taken at the University, incoming test scores, and demographics.

We chose to examine students' performance in 12 introductory STEM courses (the major sequences in Calculus, Biology, University Physics, and Chemistry). All physical science and engineering students are required to take most of these courses, regardless of major, and these courses form the introduction to these courses of study. We used this data to answer two research questions: 1) is there a gender performance gap in core STEM courses, expressed as a grade point penalty and 2) does this grade point penalty persist when controlling for composition effects?

These courses are challenging – students get lower grades in these courses than they do in other courses in the curriculum. We found that, for all students in the data set, a student typically underperforms (that is, does worse than their average GPA) by between 0.2 and 0.5 of a grade point in the core lecture chemistry, biology, calculus, and physics courses.

B. Female/Male student performance gap and Math ACT

By averaging the grade point penalty over all students over the 12 introductory STEM courses, we found an average grade point penalty of 0.23, as shown by the dashed line in Fig 1. This data set is heterogeneous, however, containing students from two different colleges, representing over 80 different majors. These programs have differing levels of student selectivity and attract students with differing interests, backgrounds, and goals. To control for this heterogeneity, we performed the same analysis, controlling for Math ACT test score. As can be seen in Fig. 1, this unilaterally reduced the observed gender difference in STEM underperformance.

This finding is indicative of "Simpson's Paradox," suggesting that the magnitude of the penalty is due in part to the compositional heterogeneity between programs: different proportions of male/female students enroll in different majors. Biology has more females than males, for example, and is in the College of Liberal Arts and Sciences, while Mechanical Engineering students are in the Engineering College, which is more selective and also skews male. As different degree programs have different entrance requirements, we are not comparing similar students in each group. Note that for the

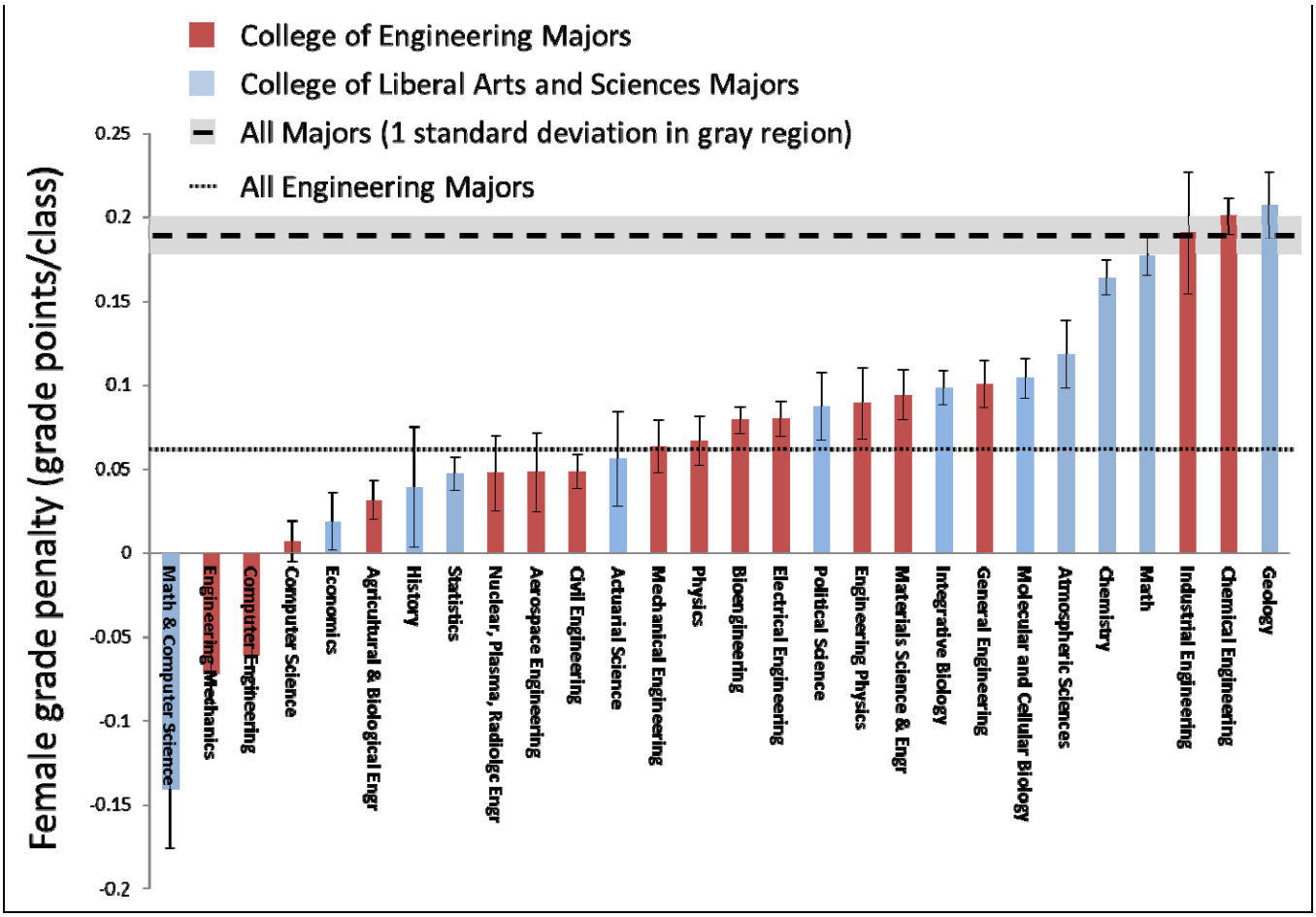

Fig 2. The average gender gap in core STEM course performance: male (GPA – grade in course) – female (GPA – grade in course) for 12 introductory STEM classes (calculus, chemistry, biology and physics) by major. Error bars are one S.D.. Includes selected Liberal Arts and Science majors (including all large natural science degree programs) and the 16 largest Engineering Majors (all have at least 150 students in the sample). Note that Computer Science and Physics are both part of the Engineering College at the University. Of the 16 large engineering degree programs shown, 2 have the same gender grade penalty

very highest Math ACT scores we have removed almost all of the composition effects – and almost all of the apparent gender bias (there is a 0.05 grade point penalty for being female if you have a Math ACT score of 35).

This composition effect works in two ways. Firstly, students in some majors will be less prepared or interested in specific courses, and so will do less well in them (Biology and Mechanical Engineering majors both take calculus and physics, for example).

Secondly, some majors require a much more demanding curriculum than others. Introductory STEM classes lower student GPA, on average. The apparent GPA of students in engineering and science programs is therefore depressed in comparison with students who take only a few science courses to satisfy general education requirements and may otherwise have a less stringently graded curriculum, giving them a higher (but largely non-STEM) GPA.

Math ACT and overall GPA are correlated, but only with an r^2 of 0.20 – math ACT is only weakly predictive of overall GPA, but including it as a single control removes around half of the apparent gender penalty in STEM courses.

C. *Female/Male student performance gap and major*

Another way in which we can control for student characteristics is to bin students by major. This is done in Fig. 2. Note that in this case the measured gender performance gap is slightly smaller (0.19) than in Fig 1. as this group includes students without ACT scores.

The addition of majors recreates the finding in Fig. 1; paradoxically, all the majors have equal or lower gender gaps than the overall average, and the average gender gap for the Engineering College majors shrinks to 0.06 of a grade point; this is less than a third of the difference seen in the whole sample, and indicates the importance of composition effects between Colleges.

D. *Racial performance gap and majoring in engineering*

This data also includes demographic information, enabling us to determine the relative under-performance/over-performance of different racial groups in core STEM classes. We focus here on performance in large Engineering majors (Fig. 3). Here, we compute racial grade point penalties (white students (GPA minus grade-in-course) minus minority students (GPA minus grade-in-course)) for the average of 12 introductory STEM courses (calculus, chemistry, biology and

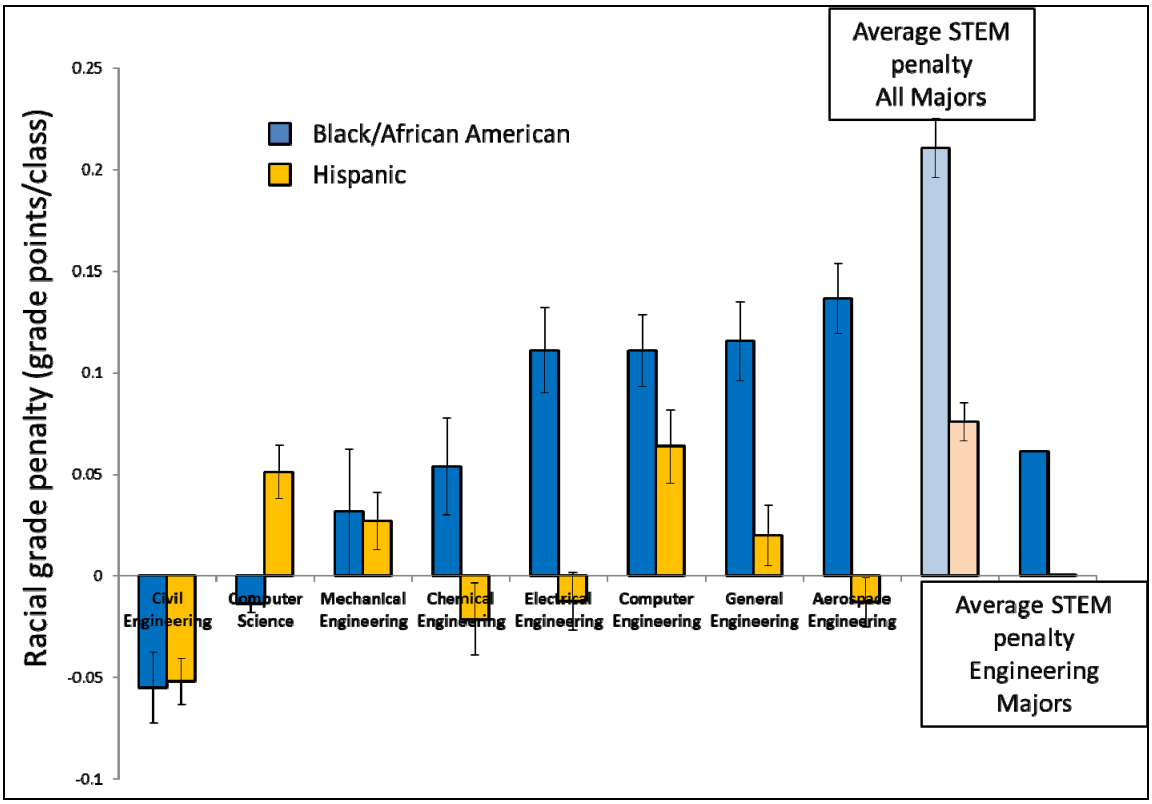

Fig 3. Impact of race on STEM class performance is reduced when course of study is included. Error bars are one standard deviation. The light colored bars are for all students in the sample, while the rightmost bars are for engineering students only. Note that evidence of Hispanic underperformance (relative to white students) entirely disappears, while the difference between white and Black/African-American students shrinks by over two thirds.

physics) by engineering major. The grade point penalty in STEM courses appears large when all students are examined: 0.21 for Black/African American students and 0.07 for Hispanic students. When we examine the 8 largest engineering majors we find that most of this STEM penalty disappears, suggesting that the observed penalty is largely the result of composition effects. The average penalty scores for students in these 8 engineering majors is 0.06 for Black/African American students (less than a third as big as for the whole data set), and 0.01 for Hispanic students (and indistinguishable from zero).

It should be noted that underrepresented minorities are very much underrepresented in our data set. All 8 programs listed have at least 1000 graduates in the time period examined, and it was necessary to have this data set be this large so as to capture sufficient numbers of minority students to run the analysis.

III. Discussion and Future Work

This study underlines the importance of controlling for student characteristics when examining large data sets. At the very least, obvious observables should be included, and a careful consideration of unobserved factors should be encouraged. By adding a single control variable we are able to eliminate the majority of the apparent difference in gender and racial performance in the STEM courses we examined. We do not claim that there is no grade point penalty, but that data must be more critically examined before being used to inform future research or policy decisions.

A previous work with a smaller, but more highly specified, sample [7] found a similar result: when SAT math scores are used to control for STEM course performance, gender differences almost disappeared. After binning students by SAT math quartile, they found that female students did 0.05 of a standard deviation better in STEM classes when the instructor was also female. Tellingly, there is good reason to think that even this small difference is largely due to the coarseness of their control of observables: SAT math scores are only divided into quarters. Most of their result is driven by difference in the top quartile of SAT math scores, in which there is significant gender heterogeneity in their sample (Figure 1 of [7]). We suspect that a finer resolution matching of controls (by using identical scores, for example, as we are able to do here) would have reduced their observed disparity even more.

Our findings are from a single institution, limiting the generalizability of the findings. The primary contribution of this paper is the evidence it provides for future methodological refinements when examining grade-point penalties. Future studies must incorporate data from other institutions before strong claims can be made about grade-point penalties at large.

Another important future work is to correct for differences in course challenge by constructing a “synthetic GPA” that accounts for the degree of difficulty of different courses. We will explore treating each course as a polytomously scored exam item using Item Response Theory to construct this synthetic GPA. It could be that the most significant driver of the GPA underperformance gap is that raw GPA is a biased measure of student quality.

ACKNOWLEDGMENT

We thank Lin Fan and Debbie Dillman for their invaluable help in gathering and anonymizing the student grade data.

REFERENCES

[1] G. Lichtenstein, H. L. Chen, K. A. Smith, and T. A. Moldonado, "Retention and persistence of women and minorities along the engineering pathway in the United States," in Cambridge Handbook of Engineering Education Research, A. J. a. B. Olds, Ed., ed Cambridge: Cambridge University Press, 2014, pp. 311-334.

[2] B. M. Holloway, T. Reed, P. K. Imbrie, and K. Ried, "Research-informed policy change: A retrospective on engineering admissions," Journal of Engineering Education, vol. 103, pp. 274-301, 2014.

[3] M. J. Graham, J. Frederick, A. Byars-Winston, A.-B. Hunter, and J. Handelsman, "Increasing persistence of college students in STEM," Science, vol. 341, pp. 1455-1456, 2013.

[4] J. G. Cromley, T. Perez, and A. Kaplan, "Undergraduate STEM achievement and retention: Cognitive, motivational, and institutional factors and solutions," Policy Insights from the Behavioral and Brain Sciences, pp. 1-8, 2016.

[5] M. Huberth, P. Chen, J. Tritz, and T. A. McKay, "Computer-tailored student support in introductory physics," PloS one, vol. 10, p. e0137001, 2015.

[6] P. J. Bickel, E. A. Hammel, and J. W. O'Connell, "Sex Bias in Graduate Admissions: Data From Berkeley," Science, vol. 187, pp. 398-404, 1975.

[7] S. E. Carrell, M. E. Page, and J. E. West, "Sex and science: How professor gender perpetuates the gender gap," The Quarterly Journal of Economics, vol. 125, pp. 1101-1144, 2010.