I. SUPPLEMENTARY TEXT

This supplementary sections contains a detailed description of the data set, supporting analysis, and additional results derived from this study. Supplementary figures are contained in a compressed file ("gpen.zip"), and tabulated results can be found in Table1.xlsx.

DATA

Our analysis is based completely on administrative data made available by the University of Michigan Registrar. The data used in this study include all undergraduate students admitted between Fall 2008 and Winter 2015 who took one of the 116 large courses defined in the main text, a total of 65492 individuals. 86% of these students were admitted as first years, 14 % were transfer students. In general we find that our results are unaffected by the inclusion/exclusion of transfers. We start by considering only students taking a course for credit, and who have received a letter grade, A-E; course grades and credit are available at 100% over the range of these data, which likewise yields a nearly 100% rate of defined GPAO.

GPAO is calculated as the credit-hour weighted average grade over all other courses completed by the end of the term in which the grade of interest is received. This allows us to consider the relative performance of first semester students - comparing their grade in a class of interest to their grades in all the other classes they have completed. GPAO is undefined for the less than 1% of students who take a single course in their first semester.

ACT/SAT and high school GPA are available for freshman admissions at the 94% level. The rate drops to $\simeq 70\%$ for transfers. We further note that HSGPA is not self-reported. It is recalculated and recorded by University of Michigan admissions from high school transcripts. 94% of students are recorded as having taken either the SAT or the ACT. In total, 75% (48750) took the ACT and 39% (25497) the SATI. When only SATI scores are available, we estimate ACT scores using multiple imputation [1]. In total, 4019 (6%) had neither score, only 504 of which are NOT transfers. In regression and matching analyses, individuals that are NA in these covariates are ignored. In addition to high school GPA, these test scores carry the most pre-college explanatory power (see below) of the metrics listed in the table. The extent to which they form the basis of our study, or serve merely as covariates will be explored in due course.

A. University of Michigan Course Catalog

Undergraduate courses are designated by subject with numbering between 100-499. 100-level courses are typically introductory and 300 and 400-level are usually specific to majors. Courses in the catalog are designated as LAB (laboratory), LEC (lecture), REC (recitation), or IND (independent study). These serve as proxies to the course format. They do not fully describe course structure; for instance, Math 215 (Multivariable Calculus), labelled LEC, involves both a lecture and lab component in the grading and course format. This is in contrast to, for example, CHEM 210 and CHEM 211 (Organic Chemistry I and the accompanying lab) which are considered separate courses. For the courses considered here, we retain the most recent subject/catalog number designation. In this supplement, we ask additional questions about the effect of course format on our observations. Where possible, we access syllabi to confirm the detailed format of each course. In addition to analysis results, Table 1 includes links to the University of Michigan course catalog that also contain descriptions of each course.

II. STUDENTS TAKING CLASSES VARY, AS DO MALE AND FEMALE STUDENTS IN EACH COURSE

Course content and structure affect grades, but the students in each course may differ from the student body at large as well. At the University of Michigan, well-maintained administrative data allow us to view this variation firsthand and to explore other predictors of student performance. Figure 1 shows the distributions of several different variables for students in Physics 140, comparing them to the overall student population. We can also compare male and female students with the course.

For each variate, we include a histogram the D-statistic from a Kolmogrov-Smirnov test, as well as the associated p-value. The differences between students in this course and the University as a whole are apparent, as are differences in these variates between men and women. The latter is especially important in the assessment of any apparent gender effects present in grades. Comparable imbalances are present in most of the courses we examine in this study. We address these through a combination of regression and matching methods (details below).

III. BOOTSTRAP ESTIMATION OF AVERAGE GRADE ANOMALIES AND GRADES VS. GPAO

To estimate average grade anomalies and their standard errors for each course we use bootstrap resampling. We create $N = 100$ bootstrap samples of all students X_i and compute the mean, $\bar{X} = \sum_i^N X_i/100$ and standard error, $\sigma = Var(X_i)$. Similarly, for the aggregate female or male grade penalty, we create boostrap estimates of the grade penalty by resampling only males or only females in the course.

In Figures 1 and 2 in the main text, the mean grades (and standard errors) in different GPAO bins are estimated by bootstrap resampling, where $N = 100$, and resampling is restricted to individuals within in the same GPAO bin.

IV. GPAO AND OTHER PREDICTORS OF COURSE GRADES

In the big picture, grades have real consequences for students: University honors, admission to undergraduate and graduate programs, and employment (e.g. clerkships,[2–4]) are often sensitive to grades. The fact that Universities meticulously record and preserve grades makes them an ideal tool for defining an expectation. This, in addition to the strong scaling of grade with GPAO demands that we include GPAO in our analyses.

Figure 2 shows that grade-point average in other classes (GPAO) is highly correlated with mean grade in a course (for courses with $N \geq 40$); as expected the correlations decrease with increasing mean grade because of saturation effects. While these are average trends, there is considerable scatter, such that grades in a few individual courses (especially those taken in the first semester) may have comparable ACT and GPAO components.

In general, the GPAO correlation grows for courses with more senior students, while the predictive power of precollege measures shrinks (Figure 3). Again, GPAO includes cumulative information about a student over her college career, so this should not come as a surprise.

A. Systematic Selection of Covariates

In the course of this study, we investigate differential grading patterns between genders, it will be natural to ask if the observed patterns are due to something other than gender, at least among information that is present in GPAO and administrative data, e.g. ACT scores and high school GPA. The predictive power of standardized test scores and HSGPA have been routinely dissected and employed in Higher Education [5–7], Psychology [8, 9], and Economics [10] literature. More recently, it was shown [11] that ACT Math and English scores far exceed the remaining ACT component scores in their ability to predict college success. A larger effort has been made to assess the impact of various predictors on college success [8, 12**?** –14]. For instance, [8] compares SAT, high school GPA and 32 measures of personality for their abilitity to predict grades. These studies typically employ some form of hierarchical or forward regression to single out predictors of grades.

Other demographic factors present in our data no doubt play a role in grades, hence the subject of this work. Remaining additional administrative data (e.g. total credits) have little support in the literature as orthogonal predictors of grade, and others (e.g. estimate gross income) are strongly correlated with other covariates (e.g. ACT scores) that are themselves relatively strong predictors.

Following previous work, we select the covariates to include in our models computationally. Linear regression-based techniques are commonly used to identify best subsets of variates. In short, one adds (forward) or removes (backward) variables to the model one-at-a-time, assessing adjusted R^2, AIC, or something similar for both the overall value and the relative change with each new addition.

We have investigated the use of the LASSO (least absolute shrinkage and selection operator) method which is well-known in the machine-learning community and explained in great detail in The Elements of Statistical Learning ([15]). This is akin to stepwise regression: in short, variables are added one at a time to a model, and the order is chosen to maximize the correlation with the residuals of each previous model. The "best" subset of covariates is then chosen according to the approach of the over model mean-squared error to an asymptote. The "lars" package in R provides a convenient interface to LASSO and generates informative diagnostic plots.

Once again using our Physics 140 example, since Fall 2008, we have 6500 students with 90% complete measurements in the following covariates: estimated gross income (self-reported), sex, ethnic group, ACT MATH, ACT ENGL, total credits, college, HSGPA, and GPAO. The LASSO evaluation of each covariate is shown in Table IV A. GPAO, ACT MATH, HSGPA, and SEX appear first (determined by the line at which each becomes non-zero). A soft cut in these variates can be set on the basis of the mean standard error, which is smoothly reduced in LASSO with the addition of new variates (Figure 4). Cutting at a penalty paramter of ~ 0.6 suggests that we set GPAO, ACT MATH, SEX, HSGPA as the most important covariates for grade in Physics 140.

	EGI	SEX	EGR	ACT_MATH	ACT_ENGL	CREDITS	COL	HSGPA	GPAO	TERM
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.79	0.00
3	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.93	0.00
4	0.00	0.02	0.00	0.03	0.00	0.00	0.00	0.00	0.94	0.00
5	0.00	0.12	0.00	0.04	0.00	0.00	-0.03	0.00	1.03	0.00
6	0.00	0.16	0.00	0.04	0.00	0.00	-0.03	0.07	1.05	0.00
7	0.00	0.16	0.00	0.04	0.00	0.00	-0.03	0.07	1.05	0.00
8	0.00	0.19	0.00	0.04	0.00	-0.00	-0.04	0.13	1.07	0.00
9	-0.00	0.20	0.00	0.04	0.00	-0.00	-0.04	0.14	1.07	0.00
10	-0.00	0.20	0.00	0.04	-0.00	-0.00	-0.04	0.16	1.08	0.00
11	-0.00	0.20	-0.00	0.04	-0.00	-0.00	-0.04	0.17	1.08	0.00

TABLE I. Stepwise addition of covariates by LASSO, where the dependent variable is Physics 140 grade. Each column contains regression coefficients in units of grade points. EGI=estimate gross income, COL=college of admission, EGR = ethnic group.

A similar exercise can be run for any course, and we are limited only by the scope and completeness of the data on hand. The subset of columns selected can vary. In PSYCH 111 (Introductory Psychology), GPAO and HSGPA stand out far above other predictors, and SEX is rejected altogether by LASSO. If fact, our experience thus far has been that in courses with a previously known minimal gender bias, LASSO rejects SEX as having predictive power.

Following [16], we can diagrammatically represent our efforts to model the association of gender with course grade (Figure 6). Even though gender and grade are marginally associated, we model them as conditionally independent given GPAO, ACT/SAT/ scores and HSGPA. Conditioning on these covariates blocks the flow association through the paths connecting gender with grade, and any residual association is attributed to a direct association between gender and course grade.

The remaining variables (Ethnicity, Estimated Gross Income, College, and Credits) are now taken to be associated with course grade through the same sturctural diagram as that for gender, such that conditioning on them does not affect the association between gender and course grade.

V. MODELING AND MEASURING AN EFFECT

In our efforts to investigate the connection between course grades and gender, we have used several measures to identify covariates which might explain our observed GPDs. In this section we describe two approaches to testing the robustness of our observed GPDs.

A. Linear Regression

Linear regression assumes a grade, g_i^j, for student i in course j can be predicted by a linear combination of nominally independent variables including a gender effect for the course (α_j), and covariates, β_i^k, which enter with coefficients χ_k:

$$g_i^j \sim \alpha_j \times gender_i + \sum_k (\chi_k \times \beta_i^k) + \epsilon_i \tag{1}$$

where the sum j is taken over the covariates (β_i^j) which include GPAO, ACT scores, and HSPGA. Finally (ϵ_i) is assumed to be a normally distributed error term. We then estimate α_j, the course dependent gender effect using OLS linear regression.

B. Matching

The shortcomings of linear regression are well-known: the error term is often non-gaussian and the covariates are most certainly correlated with one other. Matching [17, 18] is an attractive alternative. A number of methods, each

well-deployed in R packages, facilitate the exploration of matching schemes. We opted to use "optmatch", which performs optimal matching of control and treatment samples in a kind of quasi-experiment. We found that a range of more sophisicated schemes including optimal and greedy matching [19, 20] yield similar results for effect sizes, and that these are all consisent with the regression analysis. After matching, the effect sizes are measured using the ETT esimator suggested by [19]:

$$e_i = \frac{\sum_i w_i d_i}{\sum_i w_i} \tag{2}$$

The weights, w_i, are the number of controls used, which depends on the matching scheme. Each d_i is the difference between the case grade, and the mean of the matched control grades. Matching analysis generally produces noiser, but less biased estimates in comparison with linear regression, which returns potentially biased estimates with much higher precision. The mean difference between the two methods is 0.009 grade points with a SD = 0.026, showing broad consistency.

On the whole, matching on average reduces the estimated gender gap by 0.015 grade points over all the classes considered in this study and 0.022 over the intro STEM lecture courses.

VI. ADDITIONAL RESULTS

For each of the 116 courses described in the text, we include grade penalty plots and LASSO results as supplementary material ("gpen.zip"). The main results of the analyses for each of theses courses is also provide in this supplementary material as an Excel table ("Table1.xlsx").

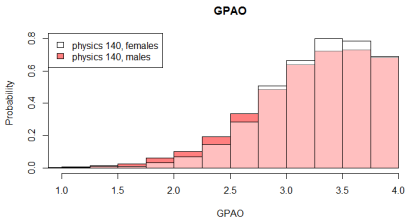
A. The Gender Gap After Matching

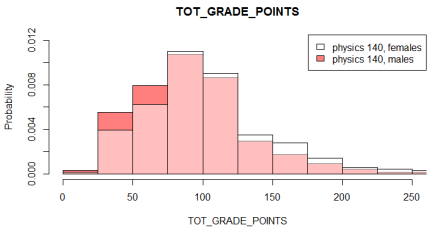
The main body of the paper relies upon raw measures of grade penalty to assess the gender gap. It is noted that the gender gap persists even after more detailed analysis that controls for sources of covariation. In Figure 7, the scaling of the matched gap with the raw gap is clear, recreating Figure 4 in the main text with the matched results for the gender gap tells the same story.

[1] S. Van Buuren and C. Oudshoorn, MICE V1. 0 user's manual. Leiden: TNO Preventie en Gezondheid (2000).
[2] P. L. Roth and R. L. Clarke, J. Vocat. Behav. **53**, 386 (1998).
[3] T. Strenze, Intelligence **35**, 401 (2007).
[4] A. B. Marks and S. A. Moss, A Longitudinal Study Correlating Law Student Applicant Data and Law School Outcomes (July 6, 2015) (2015).
[5] J. P. Noble and R. L. Sawyer, College and University **79**, 17 (2004).
[6] E. J. Shaw, J. L. Kobrin, B. F. Patterson, and K. D. Mattern, New York, NY: College Board (2012).
[7] K. D. Mattern, B. F. Patterson, and J. L. Kobrin, College Board (2012).
[8] R. N. Wolfe and S. D. Johnson, Educ. Psychol. Meas. **55**, 177 (1995).
[9] R. Zwick, *Rethinking the SAT: The future of standardized testing in university admissions* (Psychology Press, 2004).
[10] E. Cohn, S. Cohn, D. C. Balch, and J. Bradley, Jr., Econ. Educ. Rev. **23**, 577 (2004).
[11] E. P. Bettinger, B. J. Evans, and D. G. Pope, American Economic Journal: Economic Policy **5**, 26 (2013).
[12] J. W. Lounsbury, E. Sundstrom, J. M. Loveland, and L. W. Gibson, Pers. Individ. Dif. **35**, 1231 (2003).
[13] J. L. Zheng, K. P. Saunders, M. C. Shelley, II, and D. F. Whalen, J. Coll. Stud. Dev. (2002).
[14] M. Richardson, C. Abraham, and R. Bond, Psychol. Bull. **138**, 353 (2012).
[15] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, Math. Intelligencer **27**, 83 (2005).
[16] M. A. Hernan and J. M. Robins, *Causal inference* (CRC, 2010).
[17] E. A. Stuart, Stat. Sci. **25**, 1 (2010).
[18] F. J. Thoemmes and E. S. Kim, Multivariate Behav. Res. **46**, 90 (2011).
[19] B. B. Hansen, J. Am. Stat. Assoc. **99**, 609 (2004).
[20] B. B. Hansen, R News **7**, 18 (2007).

FIG. 1. Variation between Physics 140 students and the University population, and between men and women within Physics 140. Where applicable, we test the hypotheses (two-sample KS tests) that the distribution of a variate is different between Physics 140 and the larger Michigan population, or between men and women. In this course, as in most others, small but statistically significant differences exist between students who take the course and the overall student body, and between male and female students who take the course. Because of these imbalances, we examine the impact of covariates on performance prediction in every course both by regression modeling and using matching methods.

Balance of covariates for Physics 140 for terms > FA 2008

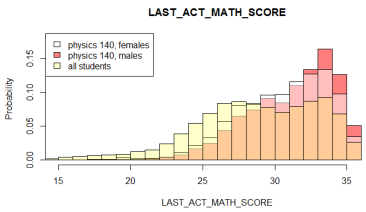


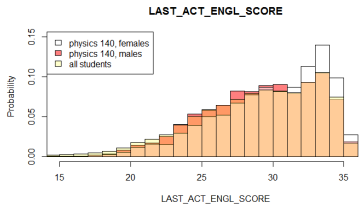


Two-sample Kolmogorov-Smirnov test data:

D = 0.052, p-value = 3e-04 for females/males in PHY140

D = 0.098, p-value = 3e-14 for females/males in PHY140

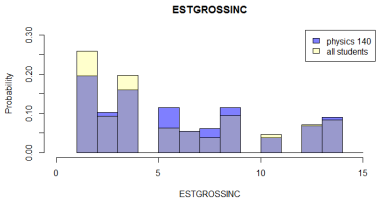


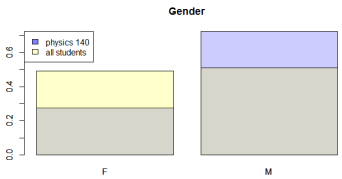


D = 0.082, p-value = 1e-09 for females/males in PHY140
D = 0.24, p-value <2e-16 for PHY140/all

D = 0.096, p-value = 4e-13 for females/males in PHY140
D = 0.043, p-value = 2e-11 for PHY140/all

Balance of covariates for Physics 140 for terms > FA 2008





Two-sample Kolmogorov-Smirnov test data:

D = 0.092, p-value <2e-16 for PHY140/all

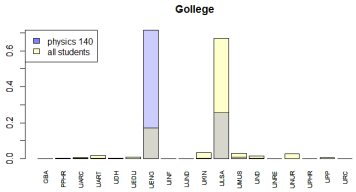


FIG. 2. Information content: Correlation of grades and various predictors. For courses with ≥ 40 students, we fit course-specific univariate models, $GRADE_i \sim \alpha + \beta x_i$ (one each for x_i = GPAO, ACT_MATH, ACT_ENGL, HSGPA) to pre-Fall 2014 student data. We then use each to predict grades for Fall 2014, and measure the correlation between the expected and observed grade as function of mean course grade. Results are shown both binned and smoothed.

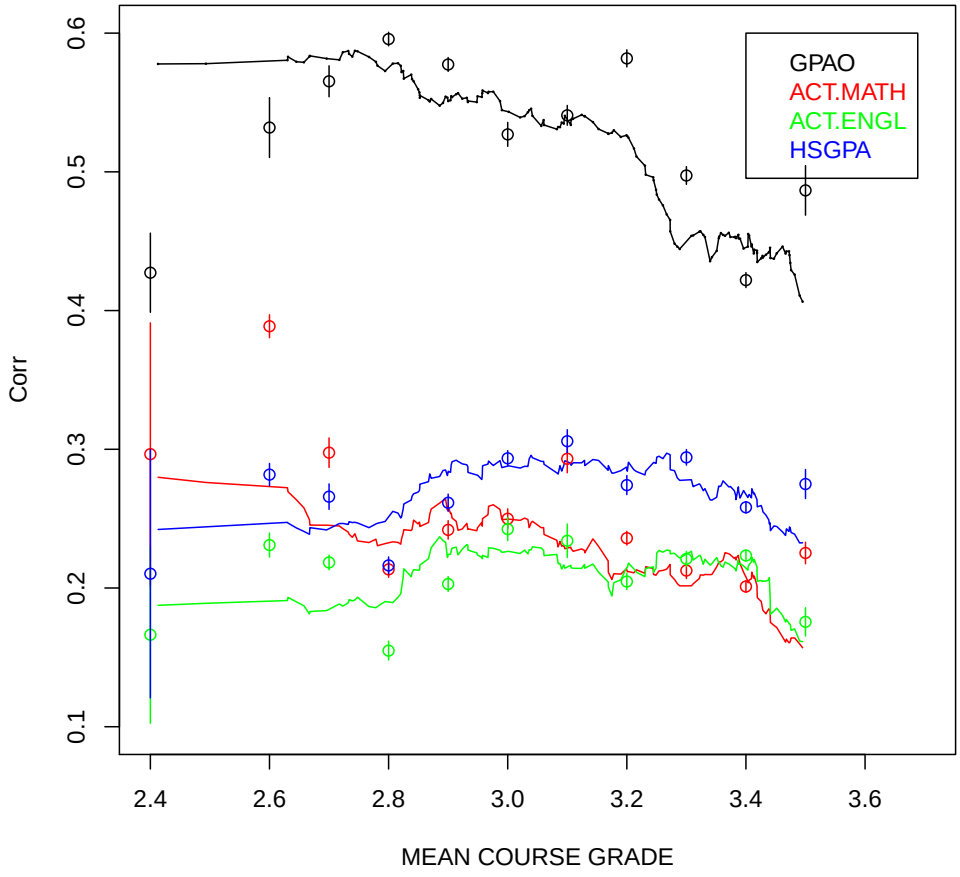


FIG. 3. Information content: Correlation of credits and various predictors. For courses with ≥ 40 students, we fit course-specific univariate models, $GRADE_i \sim \alpha + \beta x_i$ (one each for x_i = GPAO, ACT_MATH, ACT_ENGL, HSGPA) to pre-Fall 2014 student data. We then use each to predict grades for Fall 2014, and measure the correlation between the expected and observed grade as function of total credits. Results are shown both binned and smoothed.

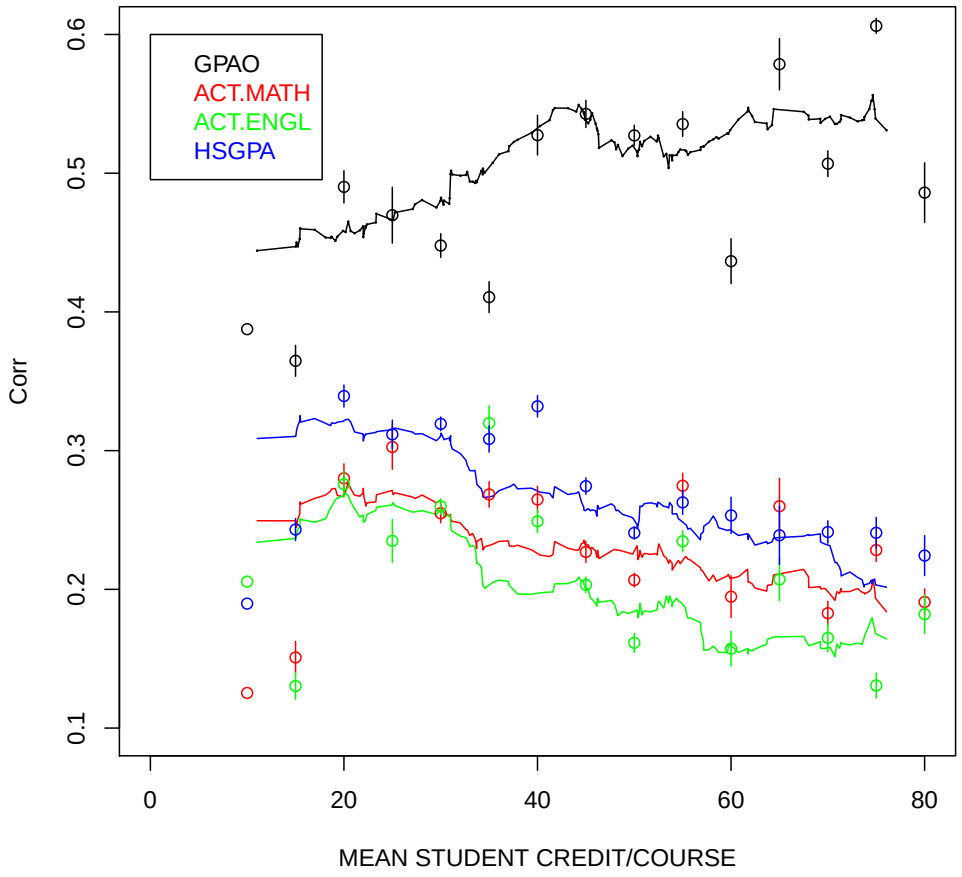


FIG. 4. LASSO MSE: Reduction of the mean-squared error vs. penalty parameter. The asymptote of the MSE occurs at ~ 0.6, which suggests an appropriate cutoff in the LASSO at about the fifth variate

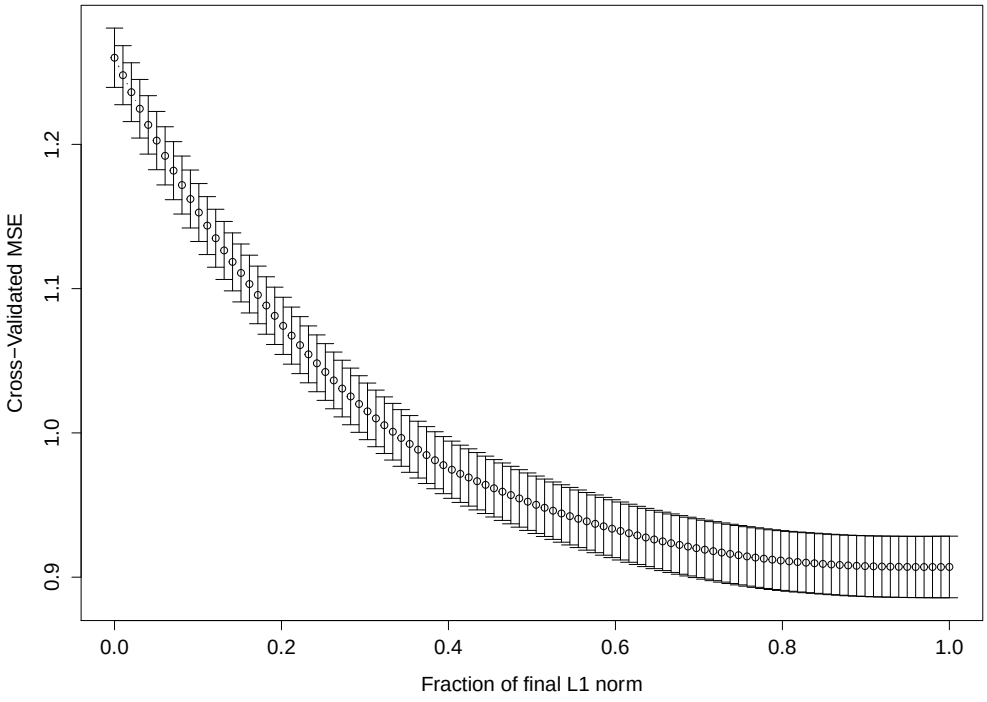


FIG. 5. LASSO iteration of coefficients, with grade in Physics 140 as the dependent variable. The horizontal axis gives the value of a penalization parameter that is central to LASSO, and coefficients are standardized for visualization purposes. Each trajectory is numbered according to the column number in Table IV A (i.e. column 9 is GPAO), and each shows the growth of a coefficient as the model is iterated.

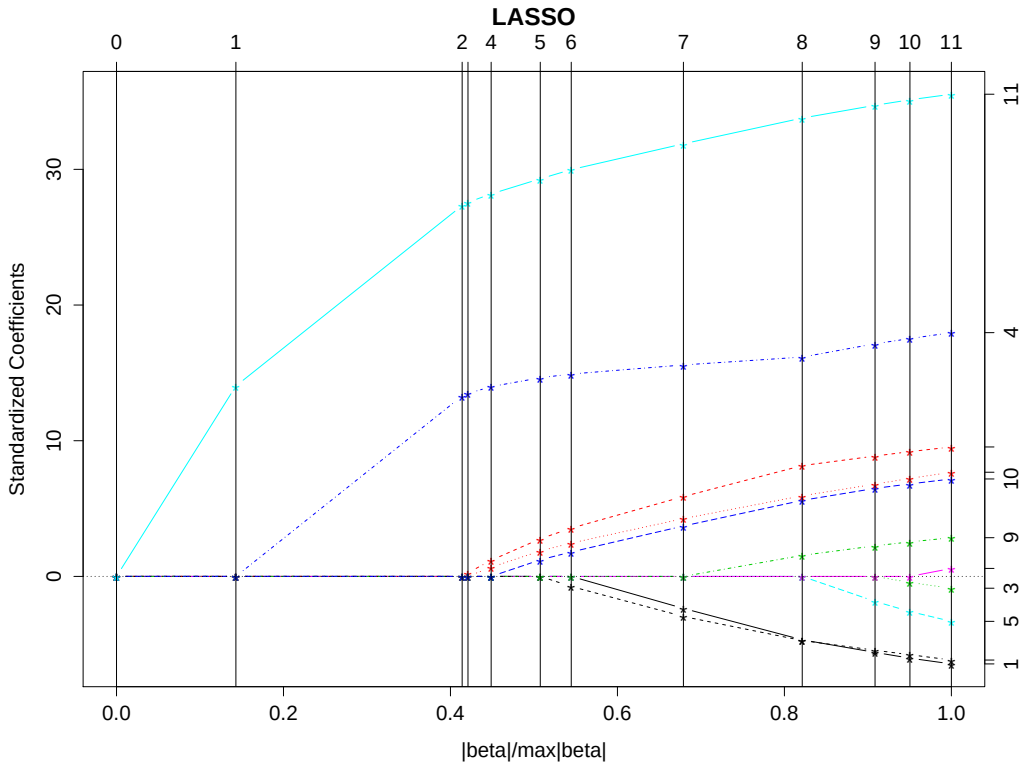
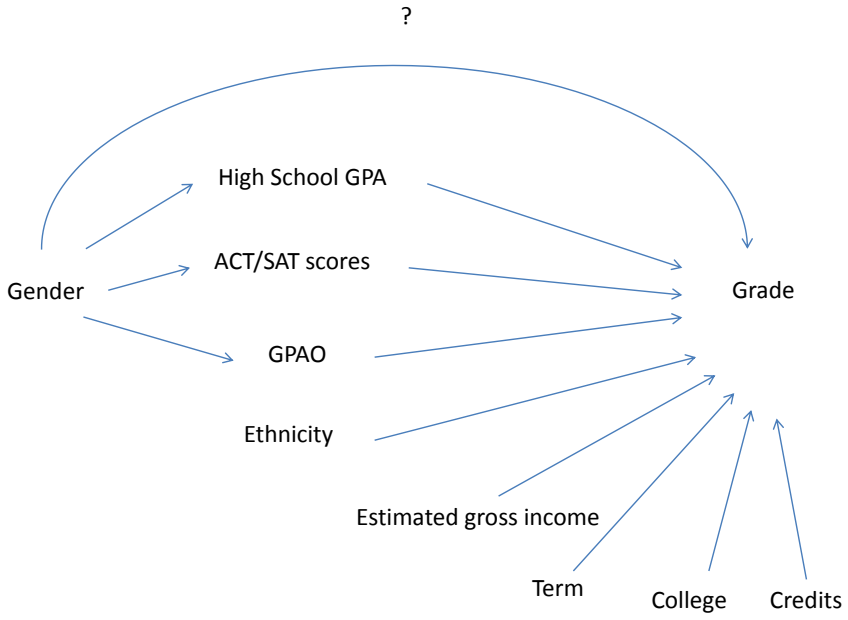


FIG. 6. Directed acyclic graph (DAG) graph that represents the relationships between variables leading to the observed associations between them. HSGPA, ACT/SAT, and GPAO are mediators and are used as covariates to remove the indirect association between gender and grade.



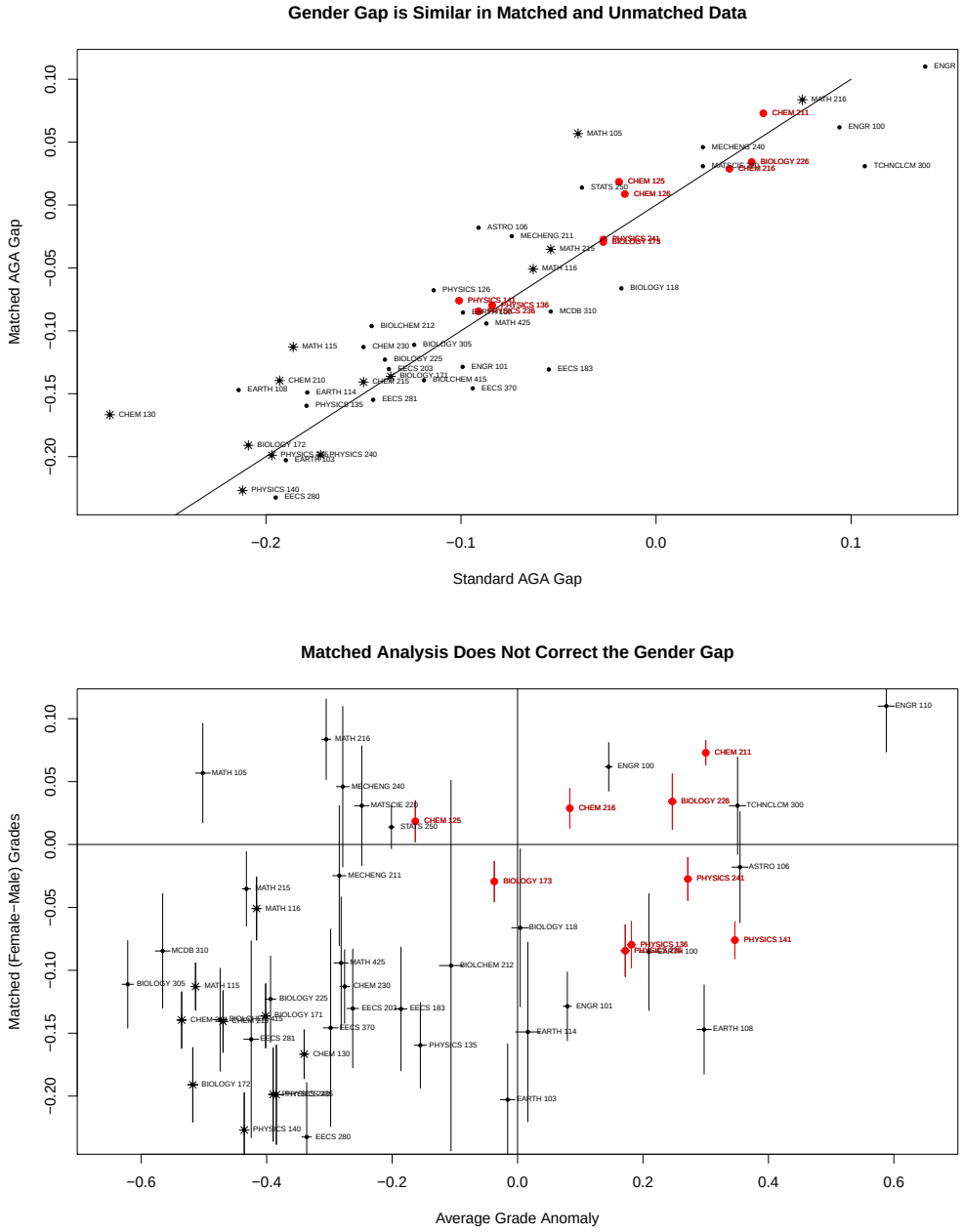


FIG. 7. The top panel shows the scaling of the raw and matched gender gaps in the grade penalty. Matching on LASSO-derived covariates does not explain the observed difference. The lower panel is a re-creation of Figure 4 in the main text, confirming that the pattern observed in the raw grade penalty plots is not diminished.