How Did You Get that A? Selectivity's Role in Rising Undergraduate Grades at a Large Public University

August E. Evrard
Departments of Physics & Astronomy
University of Michigan
Ann Arbor, Michigan, USA
evrard@umich.edu

Kyle Schulz
Center for Academic Innovation
University of Michigan
Ann Arbor, Michigan, USA
kwschulz@umich.edu

Caitlin Hayward
Center for Academic Innovation
University of Michigan
Ann Arbor, Michigan, USA
cholma@umich.edu

ABSTRACT

For nearly a century, pre-college standardized test scores and undergraduate letter grades have been de facto industry standard measures of achievement in US higher education. We examine a sample of millions of grades and a half million pre-college test scores earned by undergraduates between 2006 and 2019 at a large public research university that became increasingly selective, in terms of test scores of matriculated students, over that time. A persistent, moderate correlation between test score and grades within the period motivates us to employ a simple importance sampling model to address the question, "How much is increased selectivity driving up campus grades?". Of the overall 0.213 rise in mean undergraduate grade points over the thirteen-year period, we find that nearly half, 0.098 ± 0.004, can be ascribed to increased selectivity. The fraction is higher, nearly 70%, in engineering, business and natural science subjects. Removing selectivity's influence to surface curricular-related grade inflation within academic domains, we find a factor four range, from a low of ~0.05 in business and engineering to a high of 0.18 in the humanities, over the thirteen year period.

CCS CONCEPTS

• **General and reference** → **Estimation**; **Empirical studies**.

KEYWORDS

Undergraduate Education; Standardized Tests, Learning Analytics, Student Grades

1 INTRODUCTION

Across American institutions of higher education, letter grades and the associated four-point numeric grading scale are the canonical means to summarize student academic performance within the

curriculum [15]. During the first half of the twentieth century, the full spectrum of letter grades was employed [14]. Historical analysis of 170 four-year colleges shows that the modal grade received by a student before 1970 lay between letter grades of C or B (2.0-3.0 on the numeric scale), but in recent decades the rate at which students receive the highest level of A, the coveted, "excellent" level of 4.0 grade points, has risen sharply [13], a phenomenon popularly referred to as *grade inflation*.

Perspectives on the desirability, utility, and root causes of rising grades vary considerably. Consider an enlightenment point of view, in which the ideal of higher education is to produce rational, responsible citizens capable of supporting human progress through empiricism and open discourse [3]. To the extent that letter grades reflect the *quality* of a student's education, then higher overall grades on a campus may be a desirable outcome, so long as the higher grades truly indicate enhanced quality and more enlightened alumni. Novel instructional practices (*e.g.*, flipped classrooms, gameful learning, engaged learning techniques) and education technology services (*e.g.*, online learning, learning modules) aim to improve educational quality, and so some portion of rising grades may be ascribed to increased adoption of such tools in campus classes. Another confounding factor is the growing importance of college preparation at the high school level. If the first-year student body, especially at selective universities, is simply better prepared for college-level study, shouldn't campus grades naturally go up?

Weighing against this optimism is the perspective that rising grades reflect an erosion of educational standards by faculty responding to competitive market forces [9]. College business models create incentives for departments to maximize their course enrollments, and grades are one of many factors affecting course demand [1]. Higher tuition costs and the increasing importance of tuition in college budgets, particularly at public universities in the United States (US), fuel impressions that faculty are under pressure to "satisfy the customer." Students at highly selective institutions, almost by definition, are accustomed to being rated as excellent. Anything less than a grade of A can be a source of severe dissatisfaction to students at such campuses [10]. The professional academic market also puts pressure on undergraduate grades, since students with high grade-point averages (GPAs) are more likely to be rewarded with admission to post-graduate or professional programs than their lower-GPA counterparts.

In 2013, the Dean of the Faculty of Arts and Sciences confirmed that the median grade in Harvard College "is an A-", and that "the

most frequently awarded grade is actually a straight A".[1] Administrators at Harvard, Princeton, and other highly selective US institutions have wrestled with faculty to implement deflationary policies, with little success. Princeton revisited its 2001 grade rationing policy, designed specifically to deflate student GPAs, after considering negative consequences for their students in the postgraduate market.[2] Simply put, if only the very best students are selected for your college classes, then why would you expect any of them to earn less than an A? This is an especially pressing question if the faculty primarily consider grades as a motivational tool.

Our work is rooted in the explicit assumption that student body quality is a key contributing factor to grades earned on a campus. Prior research at Cornell has implicated this factor [1] but did not attempt to model or estimate the magnitude of its effect. In this work, we use pre-college standardized test scores as a proxy for student quality. Combining test scores with the grades earned by these students, we develop an importance sampling method [7] that allows us to estimate the mean grades expected in student populations with varying degrees of quality.

We employ the model to estimate mean grades over time based on the known evolution in campus student quality. We do this by forward-projecting from the beginning of the study (2006), and comparing grade projection to the observed grade outcomes of the student body to reveal the influence of test-score based selectivity on mean grades. Removing this influence from the actual grades exposes a fairer measure of grade inflation to be ascribed to elements within faculty control (curricular change, teaching practices, more lenient assessment, etc).

1.1 Pre-college standardized test scores

A college's admission process is an important factor affecting student grade performance. In the United States (US), the *Scholastic Aptitude Test* (SAT) and the *American College Testing* (ACT) standardized tests have been employed for decades as an important—and often required—component of an applicant's case file. A recent study of 72 institutions by the College Board found that those that accept less than 50% of applicants had student populations with GPAs that were $0.4 - 0.5$ higher than those earned by students at institutions with acceptance rates of $> 75\%$ and above [11]. That student GPA positively correlates with pre-college test scores is often perceived as a measure of validity for such testing.

Social inequalities in the US mean that the quality of a student's primary and secondary education can vary significantly, depending on geography, parental income and education level, and other factors [17]. Longitudinal studies find racial performance differences on standardized tests for math and reading in early childhood, and black students lose ground relative to other races in the first two years of school [6]. To the degree that such systematic differences, by class or race or other characteristics, persist throughout students' education, standardized test scores at the college level will inherit these accumulated social biases.

The existence of bias, or gaps, in mean test scores among different student cohorts poses important questions regarding the fairness

of their use in the admissions process. Some institutions, like the University of California (UC) state system, have decided to drop standardized tests as a required element for admission. UC will phase out the usage of the exams over five years, and has taken this step despite their faculty senate recommending to maintain the exams, citing uncertainties in how applicants who do not report scores would be evaluated against those who do [5].

Bar, et al, [2] note that grade outcomes and standardized test scores are highly correlated, and that students' own academic history then relates to their future course selection. In a study on the relationship between a large-scale effort to share median course grades with students and the impact on grades earned at Cornell, they found that the majority of students were likely to choose courses that were shown to be more leniently graded, but that higher achieving students (as identified by higher percentile scores on the SAT) were more likely to select courses with lower median grades.

A 2013 study examined outcomes in 1,683 courses from 28 departments at a large public research university over two decades. The authors found that grades increased by .188 between 1984 and 2005, and that this was less prominent in departments that awarded doctoral degrees than those that did not. [16]

At the University of Michigan, Achen and Courant conducted an extensive analysis of grade inflation within the College of Literature, Sciences and Arts. They observed that over a 16-year period, grades rose at the rate of approximate .011 increase each year on a 4.0 scale, and hypothesize that this may be because awarding higher grades "is costless and it makes students happy" [pg 78].

1.2 Study Subject and Goals

Against this complex backdrop, we seek to address a simple question: *How much does increased selectivity explain rising grades at a large, public American university?* We answer this question using a statistical model applied to over one hundred thousand students enrolled in undergraduate courses over a thirteen-year period, 2006-2019. The university's wide range of undergraduate programs allows us to address domain differences in grade inflation. In the process we uncover differences in mean grading behavior across the classical divisions within liberal arts, as well as the domains of engineering and business [8].

2 STUDY SCOPE AND METHODS

We examine undergraduate student data from a large public university between 2006 and 2019. While the university houses nineteen schools and colleges in total, the large majority (~90%) of undergraduate enrollment lies in three units: liberal arts and science (LAS), engineering (ENG) and business (BUS). These three units are the focus of our study, along with divisions within LAS.

Campus grades have risen over the study period. Figure 1 unpacks the percentages of letter grades earned by students in different credit-hour categories[3] at the start and end years. All credit hour categories display similar shifts, with a roughly 15% increase in A-level grade frequency that comes at the expense of a 5% drop in B's and a 10% drop in C-level and lower letter grades. Lower credit

---

[1]This Harvard Crimson article quotes a faculty member describing an "embarrassed silence" provoked by this public pronouncement.
[2]Sourced from this 2013 Boston Globe article.

[3]Five and higher-credit hour courses are not included in our analysis but account for only ~1% of all grades earned.
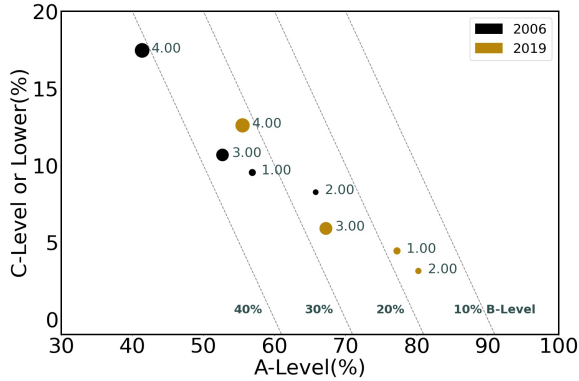
Figure 1: Grading behavior of undergraduate courses within credit-hour categories (point labels) at the beginning (dark) and end (light) of our study period. The horizontal axis gives the percentage of A-level grades while the vertical gives the percentage of grades of C or lower levels. The fraction of B-level grades can be inferred from the dotted lines. Symbol size is proportional to the fraction of undergraduate student grades in each category. Mean grades have systematically risen in all categories by similar amounts, justifying our uniform, rather than traditional credit-hour, weighting of courses when computing a mean grade point in equation (1).

.

hour classes issue higher grades than higher credit hour classes consistently across the study period.

The similar shifts in grades and consistent fractions of grades issued in classes at each credit hour level motivate us to use a simple measure of "mean grade" in undergraduate classes. For a cohort of undergraduate students, S, who take courses within academic domain, D, we define the mean grade point earned, $\langle \text{GPE} \rangle(t)$, in calendar year, t, via an equal-weight sum over courses taken in the domain

$$\langle \text{GPE} \rangle(t) \; = \; \frac{1}{N_{\text{gra}}(t)} \sum_{i \in S} g_{i,j}(t) \; ; \; j \in D. \tag{1}$$

Here $g_{i,j}(t)$ is the grade earned in the calendar year, t, by a student, i, in a course, j, that lies in domain, D, and $N_{\text{gra}}(t)$ is the total number of such grades.

We chose this measure for its simplicity. Using credit hour weighting would shift $\langle \text{GPE} \rangle$ values somewhat lower, but the fact that three- and four-credit hour courses account for the large majority of undergraduate grades means that the effect is small. The similar shifts seen in Figure 1 imply that the *fraction* of grade rise ascribed to selectivity will not depend strongly on the credit hour weighting.

2.1 Test Score and Grade Data

In Table 1, we list characteristics of the complete student sample by calendar year. Between 2006 and 2019, this measure of total undergraduate enrollment, N_{stu}, at the university rose steadily, from roughly 31,000 to nearly 38,000.

Over the study period, the university required prospective undergraduate students to submit either SAT or ACT scores as part of their application. Student records therefore contain a mix of SAT and ACT scores, with a minority of students having both, and a smaller portion, mostly international and transfer students, having neither. During the study period, two relevant changes occurred regarding the tests themselves. The SAT was redesigned in 2016, resulting in changes to both the assessment and the scoring [4]. In addition, the university in question is located in a state that switched from requiring all high school students to take the ACT to the SAT exam in 2016 [12]. The fraction of students with SAT scores declines from a high of 55% in 2006 to a low of 28% in 2016 before rebounding back to 43% in 2019. The ACT student fraction is steadier, lying in the range $(68 \pm 5)\%$ over the fourteen-year span.

Table 1 details the numbers of students with SAT scores, N_{SAT}, in each year, along with that population's mean score, $\langle \text{SAT} \rangle$, and the same for the ACT. From 2006 to 2019, the mean test scores of admitted students have risen steadily. In none of the fourteen years does the average score in either test decline. The overall increases of 100 points in SAT and nearly 3 in ACT test scores illustrate how the university became increasingly selective over time.

The final two columns of Table 1 show the total numbers of grades, N_{gra}, issued to undergraduates in all three colleges, as well as the mean grade, $\langle \text{GPE} \rangle$. The increase in number of grades roughly tracks that of enrolled students, at a rate of roughly six and two-thirds grades per person per year. Over the fourteen-year period, undergraduate grades rose by 0.2, from a letter grade of B+ (3.3) to a value of 3.5, midway between B+ and A−.

The increase in $\langle \text{GPE} \rangle$ is not as steady over time as that of test scores. Grades rise slowly at first, with the mean reaching 3.39 in 2014, a rate of 0.011 per year, before trending more strongly upward to 3.51 in 2019, a rate of 0.024 per year. If the latter trend continues it will take only twenty more years — just one and half times the length of our study, and less than the length of tenure of a typical permanent faculty member — for average grades to hit the ceiling of 4.0.

2.2 Grade Estimation Method

Our method takes as input the joint likelihood of test scores and grades earned by a population of students in undergraduate courses in some (base) year. The measures are persistently correlated — see the correlation coefficients, ρ, listed in Table 1 — which can be interpreted as affirming the validity of test score as a measure of student academic quality (premised on the implicit assumption that grades are also a measure of such). There is also significant variance in grades earned by students with particular test scores, a feature accounted for in our importance sampling method.

We view the conditional likelihood[4] — the probability of grades earned by a population of students having a given test score — as essentially a snapshot of the curriculum in the base year, its structure reflecting a rich stew with multiple ingredients including academic content, instructional and grading practices of faculty, support services for students, administrative dicta on grading policy, and other factors of the day.

---

[4]We are unaware of a common term to describe this particular statistic. A physicist might refer to this measure as a "grade susceptibility" to test scores.

Year	N_{stu}	N_{SAT}	$\langle SAT \rangle$	ρ_{SAT-GP}	N_{ACT}	$\langle ACT \rangle$	ρ_{ACT-GP}	N_{gra}	$\langle GPE \rangle$
2006	31,049	18,264	1287.4	0.198	22,147	27.9	0.219	218,786	3.301
2007	31,843	18,302	1292.2	0.181	22,945	28.1	0.207	223,029	3.306
2008	32,111	17,283	1297.2	0.195	23,533	28.3	0.216	226,256	3.318
2009	32,421	15,994	1300.3	0.182	24,221	28.5	0.213	228,502	3.331
2010	33,077	14,709	1304.4	0.173	25,102	28.7	0.214	232,799	3.331
2011	33,689	13,489	1310.9	0.172	25,914	28.9	0.211	238,638	3.336
2012	34,035	12,388	1318.7	0.165	26,241	29.0	0.200	244,714	3.356
2013	34,585	12,025	1327.1	0.160	26,515	29.3	0.195	248,183	3.365
2014	35,038	11,649	1335.9	0.158	26,825	29.5	0.188	249,282	3.382
2015	34,782	10,842	1343.7	0.146	26,868	29.8	0.183	251,429	3.417
2016	35,373	10,467	1352.7	0.145	27,500	30.1	0.178	252,017	3.444
2017	36,508	12,014	1362.8	0.162	28,015	30.3	0.172	259,995	3.464
2018	37,255	14,558	1376.6	0.167	27,050	30.6	0.159	265,481	3.474
2019	37,990	17,583	1389.3	0.168	25,514	30.8	0.170	271,268	3.505

Table 1: Undergraduate student population characteristics in all three colleges by calendar year. The rightmost column shows the mean grade point earned, equation (1). See text for details.

We combine the conditional likelihood of student grades given ACT/SAT score in the base year with weights of standardized test scores in other (target) years to predict how grades *under a fixed curriculum* would vary as student quality is varied. Monte Carlo sampling the target multiple times allows us to estimate a mean and population-related uncertainty for each target year. Our analysis uses 2006 as the base, but we confirm similar results using 2019 as the base and running the analysis backward in time rather than forward. The large statistical sample, comprised of roughly 3.5 million grades and 0.5 million test scores, allows us to estimate the magnitude of selectivity's influence on grades with high numerical precision.

Consider again the set of grades, $\{g_{i,j}\}$[5], received by undergraduate students enrolled in a domain of courses, $j \in D$, in some year, t, along with the ACT/SAT scores, $\{s_i\}$, of the same set of students. The joint probability distribution of grades and pre-college scores at some time is expressed in terms of the conditional probability, $Pr(g_{i,j}|s_i, t)$, via the Bayesian definition

$$Pr(g_{i,j}, s_i|t) = Pr(g_{i,j}|s_i, t) Pr(s_i|t), \qquad (2)$$

where $Pr(s_i|t)$ is the probability density of pre-college test scores for students in year t. The explicit use of time in the equation (2) emphasizes that this relationship is, by construction, *exact* when grade and test scores are measured for a coeval population. When the applied target is the base year, our Monte Carlo method provides mean grade points that match the actual values to a precision of ~0.001 (as shown in Figure 2 below).

The distribution of pre-college scores, $Pr(s_i|t)$, is the key measure of an institution's selectivity that we employ in our model. A university that becomes increasingly selective over time will see its student population test scores shift to higher values.

Recall that our model for determining the effect of selectivity on grades is based on the *ansatz* that the conditional probability, $Pr(g_{i,j}|s_i, t)$ reflects institutional academic rigor that we assume to be independent of student selection. That is, for a student having a

certain ACT or SAT score, their likelihood to receive a particular grade in an introductory calculus course (for example) reflects the academic standard of that particular department (mathematics, in this case), not the admissions standards of the enrollment management office. In other words, *the conditional probability reflects the structure of the curriculum and, thereby, faculty values*. While entwined by institutional priorities, the faculty's academic values are, most professors at least would argue, largely independent of the admissions office values that drive acceptance decisions.

We therefore build a predictive model that employs the conditional likelihood measured at some fixed reference time, t_{ref}, to realize expectations of the joint probability, $P_{mod}(g_{i,j}, s_i|t)$, of grades and test scores in academic domain D at some time, t, via

$$P_{mod}(g_{i,j}, s_i|t) = Pr(g_{i,j}|s_i, t_{ref}) Pr(s_i|t) ; \ j \in D. \qquad (3)$$

We can interpret the above equation as representing a hypothetical scenario in which the curriculum and faculty assessment standards are frozen in time while the student population varies year to year.

In this study, we focus on the model average grade point,

$$\langle GPE \rangle_{mod}(t) \equiv \langle g_{i,j} \rangle(t) = \frac{1}{N_{gra}(t)} \sum_{i \in S} g_{i,j} P_{mod}(g_{i,j}, s_i|t) ; \ j \in D, \qquad (4)$$

earned in different domains, D. We detail the specific implementation in the appendix.

3 RESULTS

Applying the method to all three undergraduate academic domains yields the result shown in Figure 2 and the first row of Table 2. The solid line gives the actual mean campus grade while the model prediction is shown by dotted line. The 95% bootstrap confidence range is shown by the thin band around the dotted line. Statistical uncertainties are typically $O(10^{-3})$ in magnitude.

Overall, we find that nearly half of the 0.213 rise in grades can be associated with selectivity, leaving a residual 0.115 ± 0.004 that we ascribe to the curriculum.

Both grading behaviors and selectivity's level of influence over the study period are domain dependent. Figure 3 breaks down grade

---

[5]For simplicity we drop the explicit time-dependence of the grade earned here and below, replacing it with an implicit condition.
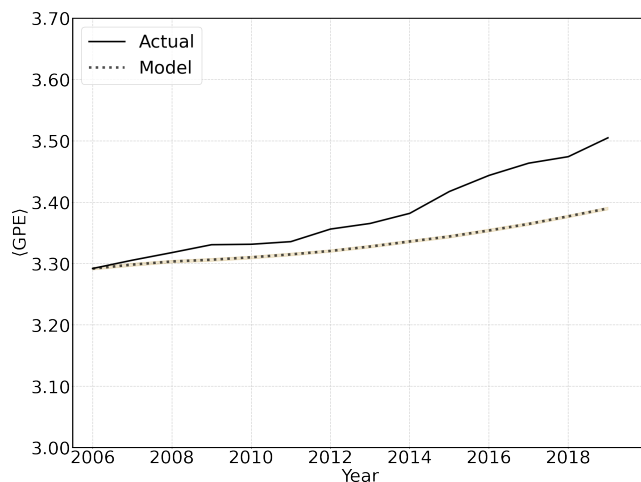
Figure 2: The trend in average grade point in all undergraduate courses in the three-college sample (solid black, equation (1), Table 1) compared with selectivity-induced model expectations (dotted, equation (4), Table 2). The shaded region around the dotted line is the model's 95% confidence bounds from bootstrap re-sampling.

behaviors and model expectations in the three colleges. In 2006, BUS awarded the highest average grade, 3.34, followed by LAS (3.26) and ENG (3.21), but by 2019 LAS had risen to the top, 3.48, closely followed by BUS (3.47), then ENG (3.39).

Deflationary efforts by the Dean of BUS in the 2010's appear to have been effective. That college had the lowest overall increase over the study period, with the average grade point remaining nearly steady since 2013. In contrast, both LAS and ENG display an upturn in average grades since 2013. We note that all domains display a slight pause or depression after 2008, but we are unaware of university-wide deflationary actions at that time. It is tempting to speculate a causal link to the 2008 recession, but we have no evidence for this.

The last column of Table 2 displays the curricular-related grade inflation values for each college. Both ENG and BUS have low values, ∼0.05, corresponding to a rate of 0.004 per year. LAS displays a value more than twice as large, 0.12. In addition, Table 2 breaks down the LAS behavior into its classical academic divisions with LAS. Perhaps not surprisingly, the natural science courses (LAS-NAT) behave more like engineering, with curricular grade inflation of 0.11. The fact that this number is twice as large as that of ENG is due to the higher growth in natural science mean grades, which rose nearly 0.3 over the study period compared to 0.18 in engineering. The curricular amount of grade inflation is intermediate in social science (LAS-SOC) and highest in humanities courses (LAS-HUM) with a value of 0.18.

The market forces associated with filling classes and producing majors are likely to be a contributing factor to these differences [1]. At the university, the number of graduates with a BA in English has declined by nearly a factor of three since 2010, and History BA's declined by a factor of two. Over the same period, the number of
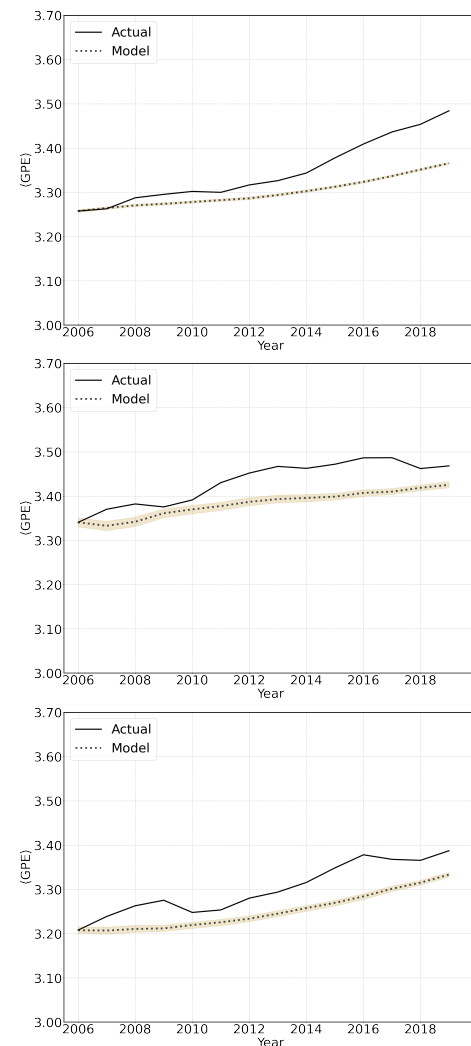

Figure 3: Mean grade points (solid black) and model expectations (dotted brown) for LAS, BUS and ENG (top to bottom) domains. Shaded regions around the dotted lines show the model's 95% statistical confidence bounds from bootstrap re-sampling.

students leaving with computer science degrees has increased by nearly a factor of five.

While the rise to prominence of the information technology sector in the global economy during the last two decades is likely to be a driving force, there are surely other factors at work. The degree of codification, or consensus on an objective set of learning goals, varies across subjects, and a recent study of grades within subjects at a large, public US university [8] finds evidence that subjects with more subjective grading practices tend to issue higher grades. We leave examination of the factors that may be driving differential rates of grade inflation across domains to future work.

Domain	$\langle\text{GPE}\rangle_{2006}$	$\langle\text{GPE}\rangle_{2019}$	Model $\langle\text{GPE}\rangle_{2019}$	f_{sel}	$\Delta\text{GPE}_{\text{cur}}$
ALL	3.292	3.505	(3.388, 3.392)	0.460 ± 0.009	0.115 ± 0.004
ENG	3.208	3.387	(3.328, 3.338)	0.698 ± 0.056	0.054 ± 0.010
BUS	3.341	3.468	(3.419, 3.433)	0.664 ± 0.073	0.043 ± 0.014
LAS	3.258	3.484	(3.363, 3.368)	0.477 ± 0.011	0.118 ± 0.005
LAS-HUM	3.374	3.644	(3.458, 3.469)	0.332 ± 0.014	0.180 ± 0.011
LAS-SOC	3.268	3.514	(3.360, 3.372)	0.399 ± 0.019	0.148 ± 0.012
LAS-NAT	3.045	3.337	(3.219, 3.229)	0.614 ± 0.021	0.113 ± 0.010

Table 2: Breakdown of model results in multiple domains, giving mean grade points, selectivity model 95% confidence expectations (derived from 200 bootstrap samples) for 2019 grades, the implied selectivity-induced fraction, f_{sel}, and derived magnitudes of curricular-ascribed grade inflation, $\Delta\text{GPE}_{\text{cur}}$. Quoted errors reflect the 95% confidence bootstrap range.

4 SUMMARY AND DISCUSSION

Grade inflation is an unappealing and largely dormant topic of discussion on most American college campuses, and we appreciate and recognize that there are plenty of good reasons to move beyond letter grade summaries of student achievement in the curriculum. Still the practice continues. Faculty issue letter grades each term and registrar offices collate them to produce transcripts that distill a student's academic history down to a single number, the GPA. That same student's entry to the university was facilitated, to a degree, by another number, a standardized test score.

We employ these historical measures in a new way, taking the conditional likelihood of letter grades as a function of test score in a particular year as a snapshot of the curricular structure in some set of academic domains. Freezing this aspect of the university, we then apply an importance sampling approach to predict grades expected to be achieved within that curriculum by student bodies with different distributions of test scores. Since the university in our study became increasingly selective, as reflected by rising test scores, over time, we find that grades on campus would also rise. Overall, we find that roughly half of the actual rise in mean grade point can be ascribed to enhanced selectivity.

Selectivity's role is not uniform across academic disciplines; we find it to be higher in engineering, business, and natural science domains. A corollary finding is that the component of grade inflation that can be ascribed to curricular drift is smallest in these areas and largest in the social sciences and humanities.

Of course, caution is required when interpreting results from a single university study. The successful deflationary tactics in the business school may or may not have played out on other campuses during the past few decades, for example. We encourage educational data scientists to address selectivity's influence, and the complementary influences of curriculum and administration, on rising grades using a larger body of evidence incorporating multiple campus histories.

5 ACKNOWLEDGEMENTS

6 APPENDIX

To perform the importance sampling we take data from the initial year of 2006 and group students into buckets based on the standardized test scores. There are four types of groupings, and then many buckets within three of them:

- The first group contains a single bucket with all 2006 grade outcomes for students have no SAT or ACT score on record. In 2006 there were roughly 1600 students in this category, or 5% of the total, a fraction that remains fairly persistent in time.
- The second group contains all 2006 grades from students who have both an ACT and an SAT score on record, with a bucket generated for every unique combination of ACT/SAT score in 2006 (i.e, if a student has a 1600/35, there will be a bucket for 1600/35 that houses all course outcomes for these students). In 2006 there are just under 11,000 students in this bucket (33%) a fraction which declines by a factor of two by 2019.
- The third group contains all 2006 grades from students who have only an ACT score on record, with a bucket generated to house grades associated with each ACT score.
- The fourth group contains all 2006 grades from students who have only an SAT score on record, with buckets generated to house grades associated with SAT score.

To estimate model GPE we do the following. For each year, beginning in 2006, loop through the $\{g_{i,j}, s_i\}$ collection of grades and scores for the sample of students taking courses in the chosen domain in that particular year. For each grade, identify the 2006 bucket, based on s_i, the student is associated with and randomly sample a grade event from that bucket. Note that s_i here may be a pair of scores or no score at all, as described above. Record the randomly sampled grade event, $\{g'_{i,j}, s_i\}$. At the end of the list, average the randomly sampled set grade events. This is our projected GPE. Repeating this process 200 times allows us to estimate an uncertainty, which we quote using the 2.5% to 97.5% range of rank-ordered results.

REFERENCES

[1] Alexandra C. Achen and Paul N. Courant. 2009. What Are Grades Made Of? *Journal of Economic Perspectives* 23, 3 (2009), 77–92.
[2] Talia Bar, Vrinda Kadiyali, and Asaf Zussman. 2009. Grade Information and Grade Inflation: The Cornell Experiment. *Journal of Economic Perspectives* 23, 3 (2009), 93–108. https://doi.org/10.1257/jep.23.3.93

[3] Gert J J Biesta. 2006. *Beyond learning: democratic education for a human future.* Paradigm Publishers, New York. https://doi.org/10.4324/9781315635811

[4] The College Board. 2016(?). *Compare New Vs. Current SAT Specifications.* https://collegereadiness.collegeboard.org/sat/inside-the-test/compare-old-new-specifications

[5] Eddie Comeaux and Henry Sánchez. 2020. Systemwide Review of the Report of the Academic Council's Standardized Testing Task Force (STTF).

[6] Roland G. Fryer and Steven D. Levitt. 2004. Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics* 86, 2 (2004), 447–464. https://doi.org/10.1162/003465304323031049

[7] J. M. Hammersley, D. C. Handscomb, and George Weiss. 1965. Monte Carlo Methods. *Physics Today* 18, 2 (1965), 55–56. https://doi.org/10.1063/1.3047186 Publisher: AIP Publishing.

[8] Joseph C. Hermanowicz and David W. Woodring. 2019. The Distribution of College Grades across Fields in the Contemporary University. *Innovative Higher Education* 44, 6 (2019), 497–510. https://doi.org/10.1007/s10755-019-09474-w

[9] Valen E. Johnson. 2003. *Grade Inflation: A Crisis in College Education.* Springer, New York. https://doi.org/10.1086/421860

[10] Craig Evan Klafter. 2019. Good Grieve! America's Grade Inflation Culture. *Academic Questions* 32, 3 (2019), 328–333. https://doi.org/10.1007/s12129-019-09810-8

[11] Jessica Marini, Emily Shaw, Linda Young, and Maureen Ewing. 2018. Getting to Know Your Criterion: Examining College Course Grades and GPAs over Time. , 22 pages. Publication Title: College Board.

[12] Jake Neher. 2015. Michigan switches to SAT over ACT starting in 2016. https://www.michiganradio.org/post/michigan-switches-sat-over-act-starting-2016

[13] Stuart Rojstaczer and Healy Christopher. 2010. Grading in American colleges and universities. *Teachers College Record* (2010). http://www.tcrecord.org/PrintContent.asp?ContentID=15928

[14] Stuart Rojstaczer and Christopher Healy. 2012. Where A is Ordinary: The Evolution of American College and University Grading, 1940-2009. *Teachers College Record* (2012).

[15] Jack Schneider and Ethan Hutt. 2014. Making the grade: a history of the A-F marking scheme. *Journal of Curriculum Studies* 46, 2 (2014), 201–224. https://doi.org/10.1080/00220272.2013.790480 ISBN: 1075501292161.

[16] R. Todd Jewell, Michael A. McPherson, and Margie A. Tieslau. 2013. Whose fault is it? Assigning blame for grade inflation in higher education. *Applied Economics* 45, 9 (2013), 1185–1200. https://doi.org/10.1080/00036846.2011.621884

[17] Rebecca Zwick. 2004. *Rethinking the SAT: The future of standardized testing in university admissions.* Ringgold, Inc, Portland. https://doi.org/10.4324/9780203463932 Publication Title: Rethinking the SAT: The Future of Standardized Testing in University Admissions.