Preliminary and Incomplete
Please do not cite

The Great Deflation:
A Quasi-Experimental Analysis of the Impact of an Anti-Grade Inflation Policy on Students and Instructors

Kristin F. Butcher (Wellesley College and NBER)
Patrick McEwan (Wellesley College)
Akila Weerapana (Wellesley College)

September 2013

Acknowledgements: We appreciate the support and feedback from administrative staff, faculty, and students at Anonymous College. We are grateful to the administration and institutional research staff for making the data available. We thank Ashley Longseth for helpful research assistance. All errors are our own.

Abstract

Average grades in colleges and universities are markedly higher now than they were in the 1960s: the average GPA was about 2.4 in 1960 and about 3 in 2006. Many critics express concern that grade inflation (and compression at the top of the distribution) erodes incentives for students to learn, gives students, employers, and graduate schools poor information on students' absolute and relative abilities; they argue there is an implicit quid pro quo between grades and student evaluations of their professors, which puts upward pressure on grades, and that grade inflation ends up eroding the value of higher education. This paper presents a quasi-experimental evaluation of an anti-grade inflation policy where a cap of a B+ was placed on the course average grade. This cap was binding for high grading departments (in the Humanities and Social Sciences) and was not binding for low-grading departments (in Sciences and Economics), allowing for a differences-in-differences analysis of the effect of the policy on grades, receipt of honors, student sorting, and students' evaluations of their professors. We find that professors complied with the policy; that they complied by reducing compression at the top of the grade distribution; that this had little effect on receipt of top honors, but affected receipt of magna cum laude designations. We also found that the effect of the policy was to expand racial gaps in GPA. There is evidence that enrollments shrank in the capped departments, but no evidence of a shift in major. Finally, students in capped courses were less likely to "strongly recommend" or "recommend" their professors and were more likely to report that they were "neutral" or did "not recommend" their professors.

I. Introduction

Colleges and Universities in the U.S. are grappling with grade inflation. Myriad reports in outlets ranging from the popular press to peer reviewed journals note that average grades are much higher now than in the past (see, for example, Rojstaczer and Healy 2010). Many express concern that grade inflation erodes the incentives for students to learn, gives students, employers, and graduate schools poor information on students' absolute and relative abilities, and ends up eroding the value of higher education. Despite these concerns, attempts to address grade inflation are generally controversial and have met with varied success.

This paper assesses the impact of a particular anti-grade inflation policy implemented at Anonymous College (AC) in 2004. The grading policy adopted a cap on average grades for introductory-level and intermediate-level courses with at least 10 students. As there were persistent differences in grade levels across departments, this policy was binding for high-grading departments and was not binding for low-grading departments. We use a differences-in-differences framework to examine the effect of the grade cap on students' outcomes and choices, on professors' grading behavior, and on students' evaluations of their professors.

We find that the cap had an immediate impact, bringing average grades down in the affected departments. Faculty complied with the policy by reducing compression at the top of the grade distribution, but there is little evidence that they increased the use of very low grades. The reduction in grades at the top of the distribution reduced the probability of students graduating with Latin honors from

historically high-grading departments. There is little evidence of a change in skill sorting of students across capped and uncapped courses. There is also little evidence that students' major choices were affected, however enrollments in courses in capped departments were reduced. Finally, it appears that students reduced their evaluations of their professors' performance in response to the change in the grading policy.

Additionally, we examine the impact of the change in grading policy for different subgroups. For African American students and students with low initial test scores, the GPA gap between them and other students increased in the departments where grades were reduced. It appears that grade compression in the high-grading departments was masking differences in "latent" grading outcomes that were revealed when grades were lowered. There is little evidence of a corresponding increase in gaps in students' evaluations of their teachers by faculty subgroup.

The paper proceeds as follows: Section II below provides background on grade inflation in U.S. Colleges and Universities. Section III discusses the potential consequences of grade inflation. Section IV describes a number of anti-grade inflation policies that have been adopted by various institutions, and describes Anonymous College's policy in some detail. Section V presents the data and methodology used in the empirical analysis. Section VI discusses the empirical results. Section VII concludes with discussion.

II. Background

Average grades in colleges and universities in the United States are markedly higher now than they were four decades ago. Figure 1, from Rojstaczer and Healy (2010), shows average grades from 1930 to 2006 for public and private colleges and universities (for which they have data), and shows these averages by type of institution. Grades rose rapidly in the 1960s, were relatively flat in the late 1970s and early 1980s, and have risen steadily through the 1990s and 2000s. In 1960, the average GPA for all institutions was about 2.4. In 2006, this number was about 3, or roughly a "B." From 1930 to the late 1950s, grade levels across public and private institutions were very similar. After this period, grade levels began to diverge, with private institutions giving out higher grades. By 2006, the average grade at private institutions was about 3.3, or a "B+."

There are many potential reasons for this increase in grades. If students are better prepared than in the past, higher grades may indicate that faculty apply a consistent evaluation method to a better performing pool of students, resulting in higher grades. This type of increase would reflect true increases in performance and would not deserve the name "grade inflation." However, SAT scores, an arguably more objective measure of student ability, have not seen the average increases that grades have, leading most observers to reject this explanation for the increasing grades, and to argue that the rise is "inflationary" (Rosovsky and Hartley 2002).

The steep rise in grades in the 1960s is often attributed to the social upheaval in that period. Maintaining a certain grade point average was necessary to

maintain military draft deferment, and faculty may have responded to this pressure with higher grades, whether or not they were merited by performance (Rosovsky and Hartely 2002).

A number of studies implicate the increased use of student evaluations of faculty as a main source of grade inflationary pressure since the late 1970s. Studies suggest that faculty members can improve students' evaluations of their performance by grading more leniently (see for example, Zangnenehzadeh 1988, Nelson and Lynch 1984, and Krautman and Sanders 1999).[1] Student evaluations are heavily used to evaluate teaching quality in promotion and tenure decisions, and thus faculty may increase grades to try to improve perceptions of their teaching quality.

A stylized fact about the grade inflation story is that there are persistent differences in average grades across departments. To a first approximation, grades tend to be highest in the humanities and lowest in the sciences and math, with the social sciences somewhere in the middle (Rojstaczer and Healy 2010). A number of researchers have used economic models to investigate the reasons for these persistent differences in grades across disciplines (see for example, Achen and Courant 2009, and Franz 2010). The basic story is as follows: in some disciplines the typical assessment methods make it more costly to evaluate student

[1] Note that studies find that students' perceptions of instructors' clarity (c.f. Nelson and Lynch 1984), for example, tend to dominate students' expected grades in determining evaluations of the instructor. However, as is discussed by researchers in this literature, estimating these relationships is not straightforward: good teaching may actually improve student performance, leading to higher grades; students expecting a high grade may infer that they are doing well and attribute that to instructor quality.

performance, and, importantly, to defend that evaluation when students complain. For example, it is both easier to see that a student's answer to a math question is incorrect and for a student to accept that it is incorrect, than to detect and explain a student's failure to support an argument adequately in a paper. In addition, when small enrollments may lead to a program being reduced or eliminated by an institution's administration, faculty are under pressure to maintain enrollment levels. If students view the difficulty of achieving a desired grade as the "price" of a particular course, then departments may improve their enrollments by reducing the "price" of taking courses in their department. This sort of arms race extends across institutions in addition to across departments within an institution. In the short run, if employers and post-graduate programs can only imperfectly observe the true quality of an applicant, institutions can (likely temporarily) improve their graduates' outcomes relative to other institutions' by arming them with a seemingly better academic outcome: a higher GPA.

III. Consequences of Grade Inflation

If grade inflation were simply a matter of a steady increase in average GPA, and the process were well-understood by all parties who use GPAs to make decisions, then it should not cause many problems. This is not, however, the typical view of the current higher education landscape. Grades are used as an indication of student ability and achievement by the students themselves, in making educational and career choices, and by employers and post-graduate programs in choosing whom to accept. Chan, Hao, and Suen (2007) for example, present a model in which

it is difficult for employers to tell whether a student is a good student, or merely goes to a school that gives easy grades. If it is difficult for employers to extract true information about a student's abilities, then schools have an incentive to help mediocre students by inflating their grades. Ultimately, this hurts the institutions of higher learning and reduces the efficiency of the job market.

Sabot and Wakeman-Linn (1991), using data from Williams College, show that grade inflation may affect students' choice of field of study. If students incorrectly interpret a high grade in a high-grading department as an indication of their superior talent in that field of study, rather than an indication that they are average and got an "A" because everyone did, then they may make the wrong choices about where to apply their talent and effort. Policy-makers and educators are concerned with the fact that relatively few students in the United States choose to major in math and science fields. It is possible that giving students mis-information about their comparative advantage in educational pursuits may exacerbate the dearth of students trained in scientific and mathematical fields, and as Sabot and Wakeman-Linn suggest, correcting this may improve our ability to produce students trained in these critical fields.

Grade inflation may also exert an adverse effect on students' efforts to learn. If one does not have to try very hard to achieve a nominally good grade, then perhaps students will not put in the study time necessary to actually master a subject. Babcock (2009) finds evidence for just such a link as higher expected course grades appear to reduce reported study time.

If post-graduate programs and employers are aware that the information contained in grades may be misleading, then they may, over time, seek alternative ways to evaluate applicants. The implications of this for students and undergraduate programs are unclear. Wongsurawat (2009) examines law school outcomes from 1995 to 2007, and finds that during the 1995-2000 period, grade inflation seems to have pushed law schools to place less emphasis on grades and more on standardized tests (LSATs) in the admissions process. One might argue that as long as graduate programs have an accurate way of determining whom they should admit, it matters little whether those are grades or standardized tests. However, standardized tests are, of course, controversial. For example, some groups of minority students tend to perform more poorly on LSATs than other groups, even after controlling for other measures of ability. Wongsurawat finds evidence that minority students' admission to law school experienced a relative decline in the period that grades were deemphasized in favor of LSAT scores.

One may or may not care about law school outcomes per se, but Wongsurawat's example is emblematic of potential problems that grade inflation poses. If employers and post-graduate programs find that grades are a poor indication of how well an individual will perform, then they will likely seek alternative ways to evaluate candidates, like standardized tests. If persistent differences in performance on standardized tests, by race, ethnicity, and gender, are due to differences in individual characteristics that have little to do with how one will perform in the legal or other professions, then there are important societal

consequences of employers' and post-graduate admissions' committees switching toward these as selection methods.

In sum, observers worry that grade inflation, and the accompanying differences in grading levels across departments, results in grave problems for higher education in the United States, and ultimately for the economy overall as resources are misallocated. Students may have little incentive to study hard if there is little reward for doing so. Further, students receive poor information about their true talents and may distort their field choices, leading them away from low-grading sciences and mathematics – fields in which employers report a lack of qualified applicants. Employers and post-graduate programs receive poor information about students' abilities and achievement, potentially hiring and admitting the "wrong" candidates, and resulting in inefficiencies from these bad matches. In an attempt to address this problem with poor information from grades, employers and post-graduate programs may seek alternate forms of information on candidates, and these choices may have their own problems, perhaps disadvantaging some candidates over others.

Despite the adverse consequences of grade inflation, there are powerful forces that contribute to it: in a world of imperfect information, institutions may be able to advantage their students over those of others by signaling that they are higher achievers; faculty have an incentive to protect enrollments and improve evaluations of their own performance; particularly in departments where assessments of student work may be more controversial, faculty have an incentive to inflate grades in order to spend more time on activities that are more productive

(like research and teaching) than defending their assessment decisions. The fact that these powerful forces all push toward grade inflation suggest that the trend is unlikely to correct itself without overt policies to address it. The next section discusses some of these potential policies.

IV. Anti-Grade Inflation Policies

A. Examples of Policies

Despite the near universal agreement in institutions of higher education that grade inflation represents a problem, there is little consensus about how to address it, and when such policies are proposed and implemented, they are strikingly controversial and do not always have the intended outcome.

The types of policies proposed tend to fall into two categories. The first is to try to inform the users of grades -- students, post-graduate programs, and employers -- about the meaning of those grades. The second is to try to set grade targets so that faculty do not inflate grades.[2] There are examples of each type of policy. Dartmouth College's policy falls into the first category. Since 1994, the

[2] This is, of course, a simplification, and each policy has its advocates. For example, Pressman (2007) suggests including average, not median, course grades on transcripts. Love and Kotchen (2010) develop a theoretical model that shows how grade inflation distorts behavior for both faculty and students, and how both more emphasis on research productivity AND on student course evaluations in promotion decisions, can reduce teaching effort. They suggest that grade targeting, at a level consistent with grades in low-grading departments, can help align behavior with institutional goals. Other recommendations focus on changing the ways faculty are evaluated. Pressman (2007) suggests eliminating student evaluations of faculty altogether in order to eliminate the inflationary pressure stemming from the implicit quid pro quo in student and faculty evaluations of one another. Zangenehzadeh (1988) suggests a less radical change in this direction: regression adjusting evaluations of faculty for the grade-level given out by the professor.

median grade in the course is listed on students' transcripts next to their own grade in the course. Presumably, this informs both the student and the other users of grade information about the content of that grade. Princeton University's policy, adopted in 2004, falls into the second group: no more than 35% of grades in a course are meant to be in the A category. Pressman (2007) reports that little has changed in terms of grade levels at these two institutions. Nonetheless, one does not have to look far to see evidence of the controversy.[3] Students clearly perceive that unilateral disarmament by their institution is going to disadvantage them in an increasingly competitive post-graduate world.

In 1996, Cornell University adopted a policy that falls into the first category. The policy was not explicitly about curbing grade inflation, but about trying to provide information about what an individual's grade means in the context of a particular course. The policy was meant to be implemented in two parts: first, median grades for each course would be made available on the web, which happened in 1998. Second, median grades in each course would be published on students' transcripts. However, the second of these steps has not taken place, and Bar, Kadiyali, and Zussman (2009) offer a fascinating analysis of the impact of the first step: grade inflation accelerated at Cornell. Once students were able to know the median grade in each course, enrollments grew in more leniently graded courses and the fact of this compositional shift – more students taking the courses

[3] At one point, The Princeton Tiger had an on-line game where one gets to play a "whack-a-mole" inspired game in which Dean Malkiel bashed students receiving A's down with a mallet to enforce the policy. As of this writing, http://gradedeflation.com/ allows students to enter their Princeton GPA and then calculates what (higher) GPA they would have received at Harvard.

with higher grades – can account for about 45% of the increase in grades in the post policy period.[4]

Clearly, constructing the "right" policy is difficult. Anti-grade inflation policies are generally controversial with students and faculty a like. There is potential for unintended consequences. There has been little evaluation of the impact of the adopted policies on the outcomes they were intended to affect. Below, we describe the policy adopted by Anonymous College (AC), and then turn to evaluating it.

B. The policy at AC

AC is a small, private, elite liberal arts college, and the ultimate goal of the grading policy adopted by AC is to insure that it persists in its commitment to excellence. The faculty and administration viewed that grade inflation and compression obscured vital distinctions in performance among students, gave students misleading information about their abilities, adversely affected students' learning by giving fewer incentives to work hard, penalized students in lower-grading departments, and undermined the institution's credibility and reputation. To address this, the College proposed the following policy in Fall 2003 and implemented it in the Fall of 2004: average grades in courses at the introductory (100) level and intermediate (200) level with at least 10 students, should not exceed a 3.33, or a B+. There was some latitude in this in that if a professor felt that her students in a given section were particularly meritorious, she could write a letter to

[4] Bar et. al. (2009) also finds evidence of a compositional shift with higher quality students being relatively less likely to take the higher grading courses.

the administration explaining the reasons for the average grade exceeding the cap. Grades by course are regularly reported to the administration, so that persistent violators of the policy can be identified. Average grades by department are regularly reported during "Academic Governance" meetings, the main governing body at the institution, such that peers can see if some departments are regularly violating the agreement. Explicit penalties, however, were not built into the policy. Penalties were left to the discretion of the administration (which, at least in the early years of the policy, may have made the threat of penalty more ominous than if it had been explicitly detailed).

Grades across departments at AC, prior to the policy, follow the stylized facts about grade inflation and grade compression reported in the literature. At AC, as at other institutions, it is largely the "sciences" that are low-grading and other departments that are high-grading. Figure 2 shows how department (and program) average grades deviated from the policy target in the years prior to the adoption of the 3.33 cap. These data are for 100 and 200 level courses (with more than 10 students) taught from Fall 1998 to Spring 2003. The departments and programs where average grades were below the cap are: Astronomy, Chemistry, Biological Sciences, Quantitative Reasoning, Economics, Geological Sciences, Math, and Physics. All other departments had grades above the cap, but there is still quite a lot of variation in grades: with several departments quite close to a 3.33, but with at least three departments exceeding the cap by about a third of a letter grade.

As Figure 2 indicates, some departments had grades below the cap, and thus should not have been directly affected by the policy when it was adopted in Fall

2004. On the other hand, there were many departments with average grades above the cap that needed to make a change in grading practices in order to come in line with the new policy. As described in more detail in the section below, our analysis splits departments into those "treated" by the policy, and those that are "untreated" by the policy, where the untreated are the eight departments listed above. Figure 3 shows what happened to average grades in the treated and untreated departments during our study period from Fall 1998 to Spring 2008. Grades dropped a bit in the soon to be treated departments in the Fall of 2003 when the policy was proposed, and data from this interim period is dropped from the analysis. After the adoption of the policy, treated departments lowered their grades such that they were in compliance with the policy. There is some evidence of an upward trend in grades in the treated departments such that by Spring 2006 average grades in those departments slightly exceeded the cap. Nonetheless, by and large, the policy was effective in its most basic goal of lowering average grades in high-grading departments to the agreed upon target.

In the next section, we present the data and methodology we use to assess the impact of the policy on student and faculty behavior, and student outcomes.

V. Data and Methodology

Our main data set is transcript level data on student grades and courses from Fall 1998 to Spring 2008.[5] Summary statistics pre- and post-policy adoption are

[5] As Appendix Table 1b indicates, AC has a diverse and high quality student body. Over the whole time period, 45% of the students identified as White, 22.8% identifying as Asian, and 5% identifying as African American/Black, and 5%

shown in Appendix Table 1a. There are 149,171 student- course- semester observations covering outcomes for over 8,000 students. There are over 8,000 students in the data. The average numeric value for grades before the policy was about 3.5. About 8% of students elected to take a given course credit/non, which means that a student had to receive a "C" or better in order to pass the course, but her GPA is not affected by receiving a "credit." In addition, 1.2% of students withdrew from classes; if a student withdraws after the "drop" period, but before the end of classes, she receives a "withdrawal" on her transcript, but, again, this does not affect her GPA.

Roughly a third of the courses were taken in the Humanities, 40 percent in the Social Sciences, and 23 percent are in the Sciences, with the remainder in other programs. Thus, the majority of courses taken by students were in "treated" departments where the grade cap is binding. AC has substantial distribution requirements such that all students must take courses in all the different divisions in the College, meaning that all students will have experiences in both the treated and untreated departments.

In order to assess the impact of the policy change on outcomes, we use a standard differences-in-differences methodology, illustrated by Figure A, below.

---

identifying as Latino. About 10% of students are first generation college students and about 10% are legacy students, related to an alum of the institution. Average SAT scores are high, consistent with its status as an elite private institution.

Figure A:

Difference-in-Differences Framework

	Pre Fall 2003	Post Fall 2004	Difference
Treated Group	a	b	(b-a)
Control Group	c	d	(d-c)

The difference (b-a) will pick up the effect of the grading policy and other general changes affecting grades, such as the quality of a given cohort of students. The change (d-c) will pick up general changes affecting grades in the untreated departments. Under the assumption that the general changes affecting grades are the same in the treated and untreated departments, the difference-in-differences [(b-a)-(d-c)] will estimate the impact of the grading policy on a given outcome. An example of a violation if this assumption would be the following: suppose that during the time when the policy changed, untreated departments were in a period where they were able to recruit students of a relatively higher caliber, perhaps because a particularly dynamic faculty member had a well-received book published during that period, changing the admissions yield for students with different types of interests, then the reduction in relative grades might be overstated in the simple differences-in-differences scheme, because not only are grades going down in the treated departments because of grade cap, but they are rising in the untreated departments because of the change in the quality of students.

In order to allow us to control for other factors, like observable dimensions of student quality, that might affect grades we estimate regressions of the following form:

$$Y_{idt}=\beta_0 + \beta_1 PostPolicy_t + \beta_2 Treated_d + \beta_3 Post_t * Treated_d + X_{idt}\Gamma + \varepsilon_{idt}$$

where i indexes the individual, d the department, and t the year. Y is the outcome of interest; PostPolicy is a dummy for Fall 2004 and beyond. Treated is a dummy variable equal to 1 for those departments with average grades above the cap prior to the policy. The coefficient β_3 is the coefficient of interest as it indicates whether the outcomes for students in courses in the treated departments changed differentially before and after the policy was implemented.[6] X_{itd} is a vector of fixed and time varying individual or course/department characteristics. In some specifications, described below, we use individual, department, and semester fixed effects to control for both observed and unobservable differences across people, departments, and years. Standard errors are clustered at the department level.

VI. Results

A. Grades

i. Overall

Table 1 presents the results for the numeric value of the letter grade received by each student. Each observation is at the student-course level.

The first column reports the unadjusted differences-in-differences estimate (corresponding to the results in Figure 3). The second column includes a vector of student and course characteristics.[7] The final column replaces the "post" dummy

[6] Figure 2 indicates that some departments had more change to make in order for their grades to be compliant with the new policy. We experimented with an "intensity of treatment" variable rather than the simple dichotomous "treated" variable described above, and found that the results are qualitatively similar.

[7] These include math and verbal SAT scores, dummies for Black, Latino, Foreign, Asian, Non-traditional aged student, first generation college student, and legacy student. Missing data for SAT scores are entered as the mean and a dummy variable

with a vector of semester dummies, replaces students' fixed characteristics with individual fixed effects, and replaces the "treated" dummy with a vector of department fixed effects. The coefficient on "treated*post" shows the estimated impact of the anti-grade deflation policy: grades in treated departments dropped by about a sixth of a letter grade after the adoption of the policy. The estimate is very stable across the three columns, suggesting that the policy change was orthogonal to other factors that might also affect changes in grades across departments.

The other coefficients are also of interest. First, note that this was a period when grades fell in the untreated departments as well. Whether this was itself due to the policy – for example, because the public discussion of grade inflation induced all faculty to adopt more stringent grading standards – or whether it is due to other factors (such as changes in student performance), is not something that can be identified in this framework. Second, note that the other coefficients in column 2 are consistent with those found in other papers studying the relationship between student characteristics and grades (McEwan and Soderberg 2006, Zimmerman 2003, Elzinga and Melaugh 2009).[8] Higher SAT scores are associated with higher grades, and African American, Latino, and Legacy students get lower grades, conditional on other characteristics.

---

equal to 1 when missing data were replaced is included. The only course characteristic is class size.

[8] McEwan and Soderberg (2006) and Zimmerman (2003) report very similar ethnic and racial gaps to those reported in column 2 of Table 2A. Elzinga and Melaugh (2009) find evidence that legacy students perform less well than other students with similar characteristics. Espenshade et. al. (2004) finds that legacy students are about three times more likely to be admitted to selective colleges and universities than non-legacy students, conditional on characteristics.

The evidence in Table 1 shows that when the policy was adopted that average grades should not be above a B+, faculty complied with the policy and reduced grades in courses in departments that were above the cap. We now turn to examining *how* faculty changed grades to come into compliance with the new rule. Table 2 shows the results of a series of linear probability models, where the outcomes are dummy variables equal to 1 if the student got a particular grade and zero otherwise. So, in the first column, the outcome is 1 if the student got a straight A, and zero otherwise. In the last column, the outcome is equal to 1 if the student got a C- or below, and zero otherwise. Note that in this column, students who took the course Credit/No-Credit, but ended up failing the course are included as "1," so students electing credit/non are included in this regression and there are more observations. The specification includes student fixed effects and class size and a post-policy dummy variable. The results are robust to the other specifications as well.

The results indicate that after the policy change, students were about 14 percentage points less likely to get a straight A in the treated departments. On average in the pre-period, about 29 percent of the grades were straight A's, so a 14 percentage point drop is substantial. Column 2 of Table 2 indicates that students were about 18 percentage points less likely to get an A or A-. Correspondingly, the probability of receiving a B+ increased by about 7 percentage points and any type of B increased by about 17 percentage points. There was no increase in the incidence of very low grades, C- or below (including no-credit for those electing the credit/non option). This evidence then, suggests that the policy did alleviate the compression

of grades at the very top of the distribution. There is no evidence that faculty began using very low grades for some students in order to "preserve" A's for other students.

Of course, grade point averages may also change if students' choices about electing to take courses using the credit/non-credit options changed due to the policy, or if their decision to withdraw in the fact of a low-grade changed. We analyzed both of these outcomes as well. First, there is no evidence of a differential change in withdrawal probabilities in treated departments after the grading policy took effect. There is evidence that the credit/non election changed differentially. Appendix Figure 1 shows how the credit/non option changed. There was a precipitous drop in the credit/non election in the departments that were *unaffected* by the grade cap. This, it turns out, can be traced to a policy adopted in Spring of 2003 – also designed to make outcomes in courses more informative – that moved the date by which students had to declare whether they were electing to take something "credit/non" much earlier in the semester, often before they would have received exam results. Not surprisingly, before this policy change, students were more likely to take courses in the low-grading departments credit/non. Once they had to decide prior to receiving concrete information on their grades in the course, their election of credit/non converges to the same rates as in the higher grading departments. The effect of this (slightly) earlier policy would likely be to push grades down in the departments that were "untreated" by the grade cap, because low performance was less likely to be masked by credit-non election. Thus, if

anything, the relative decline in grades in the departments that were affected by the 3.33 grade cap is likely to be understated.

ii. Subgroups

AC, like most other elite colleges and universities, shows persistent differences in GPA across different groups of students, even when controlling for characteristics like SAT scores that are meant to capture differences in preparation for college. In this section, we examine the impact of the grading policy on these gaps. Table 3 presents the results for the numeric value of students' grades, and the unit of analysis is at the student-course-semester level. The first column shows the overall differences-in-differences estimate with student characteristics held constant. In each of the rest of the columns, "Treated," "Post," and "Treated*Post," are interacted with a dummy variable indicating that the student was in a particular "Group." The Group in question is listed in the column headings.

Column 1 shows that grades fell by about a sixth of a letter grade in treated departments after the adoption of the average grade cap. The second column allows this estimate to be different for African American students. These students' grades fell by an *additional* 0.188 points. In other words, for African American students, the drop in GPA in treated departments was almost twice that for the rest of the students, and was about a third of a letter grade. There was a relative increase in the grades for Latino students, which may be driven by the fact that these students do relatively well in Spanish classes – which was the department that had to impose the largest decline in grades in order to comply with the policy.

We also examined the impact on grades for students who began AC with relatively low test scores. Column 4 uses as the "group" those students who received scores in the bottom 5% of a quantitative reasoning (QR) assessment administered to first year students during their orientation period at AC. Column 5 uses as the "group" students in the bottom 5% of SAT verbal scores (among AC matriculants, not in the national distribution of SATV scores).[9] Note that column 4 indicates that there is no differential effect of the policy on students with very low quantitative reasoning (QR) scores. However, there is a large and statistically significant effect for students with very low verbal SAT scores.

We believe this pattern emerges because the policy removed much of the grade compression at the A- threshold in the grade distributions in the treated departments. Imagine a world in which there is a latent performance measure of which the faculty member is aware. However, in an era with a great deal of grade inflation and grade compression, this latent performance is mapped onto a very narrow range of letter grades. Now the policy imposes a grade cap, and so the faculty member moves those with the lowest latent performance farther down the scale of letter grades. Put differently, if African American students were receiving the "lowest A-," for example, then when giving an A- or higher to most people in the class is no longer an option, their grades will fall more than those who were receiving the "second lowest A-." The evidence for students with the lowest SAT

[9] The numbers of observations are different because in the overall regressions, missing values are replaced with zeros and a dummy variable is included to account for this imputation. We throw out observations with missing values when we are defining group member based on these scores. QR scores are missing because the test was not given to the classes of 1999, 2000, and 2001. SATV scores are missing because some students apply using ACT scores.

verbal scores is consistent with this interpretation. Students with low verbal scores are the ones who would likely be at the bottom of the reading and writing intensive humanities classes that make up the bulk of the classes in the treated group, and again, when mapping disparate class performance onto a very tight distribution of high letter grades is not an option, they are the ones whose letter grades move down the most.[10] [11]

The existence of a GPA gap between African American and other students at elite colleges and universities, conditional on their SAT scores, is cause for concern. Whether one thinks an increase in the gap due to the anti-grade inflation policy adopted by AC is cause for *additional* concern depends on why the gap increases in treated departments. If the imposition of the grading caps caused professors to search for someone to give a bad grade, and this was independent of actual performance in the class, then the increasing GPA gap would be due to professorial bias that led them to favor one group over another. Another possible interpretation

[10] The fact that students with the lowest QR scores do not also experience a large drop may be less surprising when one considers that the treated departments are not mainly those with classes that rely on quantitative reasoning, and these students, given that they are relatively bad at quantitative reasoning, must be relatively good at something else or they would not have gotten into a highly competitive academic institution.

[11] We have also done the analyses for types of letter grades given, receipt of academic honors, probability of credit non-election, and probability of withdrawal. The probability of receiving high grades dropped and the probability of receiving low grades increased for African Americans and students with low QR and SAT test scores. Latin honors, which are closely tied to GPA, fell in ways consistent with the grade data. There was no differential effect of the policy on the probability that students from different subgroups elected to take a course credit-non. There is some evidence that African American and low-test score students were relatively less likely to withdraw from courses in treated departments after the policy change (and this compositional change may be part of why their grades were relatively more affected).

is that the discussion about grades and grade caps created a situation of "stereotype threat," such that one group of students' performance in the treated classes was adversely affected by the policy. Both of these interpretations indicate an adverse impact of the policy. A third interpretation is that outcomes in classes for African American and other students always differed, but the extent of this difference was more evident in departments where the grade distribution was not compressed at the top. Once the compression at the top was changed, the differences in grades across departments are more similar. The fact that grades for students with very low SAT verbal scores – which is a racially and ethnically diverse group of students at AC -- behave in the same way as grades for African American students, lends support to the idea that the policy induced faculty in treated departments to use a wider distribution of grades which revealed a wider distribution of performance.

Those concerned about grade inflation often argue that the compression in grades masks important information and that misleadingly high grades lead to poor resource allocation decisions by students choosing where to allocate their efforts, and by employers and post-graduate programs choosing whom to allocate their admissions slots. Assuming that the increase in the GPA gap that resulted from the AC anti-grade inflation policy is due to the change in compression in the treated departments – and not due to either increased bias or stereotype threat – AC provides a clear illustration of the resource allocation problem: student academic support services were more heavily targeted toward those departments with larger gaps. For example, more tutors, teaching assistants, and supplemental instruction classes were available for more technical courses because that is where gaps in

performance appeared stark. If similar gaps in performance exist in high-grading departments, but are masked, then students in those departments have not been allocated the support services that might benefit their learning.

iii. Honors at Graduation

One of the frequently cited reasons for being concerned about persistent differences in grade levels across departments is that students majoring in high grading departments are disproportionately rewarded with Latin honors at graduation. Given the changes in relative grades described in the previous section, one might expect a relative change in the receipt of Latin honors after AC adopted the grading cap.

Table 4 presents linear probability models where the outcome is whether or not the individual received a particular type of honors upon graduation. Here the unit of observation is a student, assessed at the time of graduation, thus the number of observations is only 6,738. The regressions control for student characteristics, department fixed effects, and graduation year effects. The "post" period includes students graduating in Fall 2005 or later, so those who were graduated at least a year after the policy took effect. Column 1 presents results for *Summa Cum Laude*; column 2 for *Magna Cum Laude*; column 3 for *Cum Laude*, and column 4 for *Phi Beta Kappa*. The results are consistent with the evidence presented on grades. The probability of graduating Summa cum laude, the highest level of honors, was not (significantly) differentially changed across departments by the policy. In order to graduate *Summa*, a student needs to get a GPA of 3.9 or above, and at AC, where there are substantial distribution requirements, this means that a student needs to

get straight As in many different types of courses. The *Summa* students were not differentially affected by grading policies across departments, and thus not differentially affected by the imposition of the grading cap.

The probability that a student graduated *Magna Cum Laude*, on the other hand, was markedly affected by the policy. For the treated departments there was a four percentage point drop in the probability that a student received a *Magna* designation. About 20% of students overall received a Magna designation in the pre-policy era.

There was no statistically significant differential change across treated and untreated departments in the probability that a student was graduated *Cum Laude*. Presumably, many of those who would have previously received a *Magna* designation slipped into the *Cum Laude* category, off-setting any declines from that category. The overall probability of a student receiving any type of Latin honors fell over this period.

Interestingly, there was no significant change in the relative probability in the treated departments that a student was elected *Phi Beta Kappa (PBK)*. This is likely because, unlike Latin honors, PBK election is not strictly based on a GPA calculation. A committee selects students for this honor, and the committee may always have been adjusting for what it viewed as a grade-inflated record.

B. Student Sorting and Choices

i. Sorting on Measures of Quality

In Table 5 we examine evidence on whether the type of student taking classes in treated and untreated departments changed. We have derive measures of student "type" from SAT scores and scores on the quantitative reasoning assessment (QR Score).[12] Column 1 uses each student's (in a course/semester) QR score as an outcome. Column 2 presents a linear probability model for whether a student in a course/semester had a score in the bottom 5% of the QR score distribution. Column 3, similarly, presents a linear probability model for whether a student (in a course/semester) had an SATV score that fell in the bottom 5% of distribution (among AC students). The regressions control for department and semester fixed effects, and student characteristics. We find that the QR scores fell in treated departments relative to untreated departments by a statistically insignificant and small amount: -0.185 points (on an 18 point scale). Similarly, we do not find statistically significant evidence that students with either the lowest QR scores or the lowest SATV scores shifted away from courses in the now relatively lower grading departments. There is little evidence that the policy changed the sorting of *types* of students across departments. One should keep mind, however, that although these departments lowered their grades after the policy, they still had higher grades on average than the untreated departments.

[12] A quantitative reasoning test was given to every student entering after the graduating class of 2001, during the orientation period for incoming students. The exam is used to determine whether students have the necessary background to access the curriculum in the quantitative classes offered at AC. It is an 18 point test, and if students do not pass, they take a semester long course on quantitative reasoning. A passing score is 9.5, and the average score is 12.6.

ii. Student Choices: Enrollments and Majors

Grade inflation is often blamed for distorting students choices across fields by giving them misinformation about where their relative strengths lie. Thus, we might expect that a change in grading policy would lead students to make different choices about which course to take or in which departments to major. We find no evidence that choice of major was affected by the policy change (results available, but not shown). However, we do find some evidence that enrollments were affected. Table 6 presents four specifications where the outcome is total departmental enrollments, in either levels or logs. There are a number of very small departments and programs at AC and these were eliminated from this analysis keeping the 15 largest departments, resulting in 342 observations on total enrollments by department for each semester. The results suggest a substantial decline in total enrollments for departments affected by the policy: enrollments fell a statistically significant 20 percent in treated departments.

It may seem contradictory that major choice was not affected by the policy, but enrollments were. It seems plausible that major choice is deeply tied to students' interests and these interests are not shaken by changes in grading policy, but a substantial amount of college course work at a liberal arts institution is comprised of elective courses outside one's major. Perhaps when students majoring in an untreated department learn that adding an elective in a treated department is

no longer guaranteed to boost one's GPA, they choose an additional course in the their major or an elective in another untreated department.

C. Students' Evaluations of their Professors

How to evaluate teaching at the College and University level is controversial, with recent research suggesting that those professors whose students get the highest scores, or professors who receive the highest evaluations, are not those whose students go on to demonstrate greater knowledge in subsequent courses (Braga et. al. 2011, Carrell and West 2010). Nonetheless, student evaluations of professors are a nearly universal feature of the higher education landscape in the United States, and form the main way in which teaching is evaluated at most institutions. Many observers contend that students' evaluations of their professors set up an implicit quid pro quo with student grades, and that this system of evaluating professors has contributed to grade inflation.

Student evaluations of professors are a critical component in promotion and tenure decisions at AC. Students submit evaluations electronically and their ability to see their grades in a timely fashion is tied to submitting an evaluation during a specified period. Thus, nearly 100% of students submit evaluations. Over the period studied, students gave both professors and courses separate ratings on a four point scale, with a "1" showing the strongest endorsement.[14] There is also a qualitative component to the evaluation, but we only have access to the numeric component of the evaluation. Appendix Table 1d shows the fraction of students giving ratings in

[14] 1="Strongly Recommend"; 2="Recommend"; 3="Neutral"; 4="Do Not Recommend"

each category before and after the policy change. Students are generally well-satisfied with their professors at AC, with over 60 percent "strongly recommending" their professors.

Tables 7A through 7D present evidence of the impact of the anti-grade inflation policy on student evaluations of their professors. The unit of observation is a faculty member in a course in a given semester. The outcome variable is the fraction of students issuing a rating within a given category. Table 7A presents results for the fraction of students giving their professor a "strongly recommend" rating. The first column shows the raw difference-in-difference estimate. Column 2 includes controls for faculty characteristics, including demographics and rank. Column 3 adds controls for level of the class, class size, and the average course GPA. Course GPA is, clearly, endogenous to the policy, but we were interested both in the coefficient on GPA itself in these regressions, and to see whether evaluations of courses in treated departments were affected even when compared to courses where students (nominally) achieved the same grade. Column 4 adds department and semester fixed effects to the regression.

In column 1, there is about a five percentage point decline in the percent of students who strongly recommend their professors in the treated departments after the policy change, and the change is highly statistically significant. It is robust to the inclusion of controls for faculty characteristics. The point estimate declines in absolute value to about -4 percentage points, but remains statistically significant, in columns 3 and 4. The fact that the effect is statistically significant even when course GPA is held constant suggests that students' expectations about their grades play an

important role in their evaluations of their professors. Even when compared to courses where the average grade is the same, those in the treated departments in the post-policy era are less satisfied with their experience. Note also that the coefficient on GPA itself is positive and significant, indicating that courses with higher grades have higher rated professors.[15]

Tables 7B through 7D show that not only is there a decrease in the percentage of students giving their professors a "strongly recommend" rating, there is also an increase in the percent of students giving their professors the lowest rating. Across the columns in Table 7D, there is an increase of about 2 percentage points in the percent of students giving their professors a "do not recommend" rating. On average, about 4 percent of students give this rating to their professors, so this is a large increase in the prevalence of this negative rating.

VII. Discussion and Conclusion

Grades at institutions of higher learning in the United States, particularly at private institutions, have risen markedly over the last decades, and continue to rise. Many argue that this is a crisis in education (see for example, Johnson 2003). Since grades are persistently higher in some fields than others, concerned observers claim grade inflation leads students to choose to major in some fields, not because they are particularly good at them, but because they receive the impression that they are good at them. This leads to a misallocation of resources since students are not

[15] Of course, higher grades may indicate that the professor believes the students learned a great deal in the class, and students' evaluations of the professor may simply corroborate that.

working in their areas of comparative advantage and society ends up training too many *social* scientists and not enough *scientists*. Further, employers and graduate programs may end up choosing the wrong students to hire and train, because they take the Humanities major with the 3.5 GPA over the Science major with the 3.2 GPA, even though both of these are average at their undergraduate institutions.

These concerns are grave, but there are powerful forces driving grade inflation. Faculty, whose promotion and tenure depend on student evaluations of them, have an incentive to increase students' grades quid pro quo, and these pressures may be particularly intense in disciplines where it is more time-consuming to evaluate work and defend that evaluation. Departments are competing for enrollments and have an incentive to entice students to take their courses by offering a higher GPA for less effort. In an environment where it is hard for employers and graduate programs to really know whether differences in grades for students from different institutions are due to differences in students' abilities or the grading environment at the institutions, the colleges themselves have an incentive to inflate grades relative to their peer institutions. With all these incentives lining up to inflate grades, it is unlikely that grades will fall on their own without an explicit policy to bring them down.

Anonymous College (AC) adopted just such a policy in the Fall semester of 2004. Average grades in courses were capped at 3.33, or a B+ average. There were high-grading departments that were affected by the cap, and low-grading departments that were not. We compare outcomes across the treated and untreated departments before and after the change in the policy. It is clear that faculty

responded to the mandate and reduced grades to comply with the policy. This relieved compression at the top of the grade distribution. Other changes, such as the fraction of students graduating *Magna Cum Laude*, changed in ways consistent with the changes in grades.

Some have recommended grade targets as a way to change students' choice of major, perhaps to encourage more students to major in the sciences. The AC policy did not have this effect. Although there is some evidence that enrollments experienced a relative decline in the treated departments, there is no evidence that choice of major was affected. This may be because although the policy did bring grades closer together in the high-grading and low-grading departments, the former still has significantly higher grades than the latter. If changing students' major choice is the desired goal of the policy, more stringent targets may be needed.

We also find that the grading policy change had different effects on the grades of different students. African American students and students with very low verbal SAT scores saw their grades in treated departments fall more than other students. Our preferred interpretation is that this is due to the fact that these were the students receiving the "lowest A-" in a very compressed grade distribution, and when the policy forced grades to spread out, these groups' grades decreased more than those for students previously receiving the "second lowest A-." To the extent that the policy helped reveal existing achievement gaps, and academic support services can now be more effectively allocated where they are needed, the policy may have beneficial effects for learning.

This paper finds robust evidence that student evaluations of professors are tied to students' grades, and when the faculty reduced grades, students reported being less satisfied with their learning experience. Whether this implies a quid pro quo relationship between faculty and student grades is not something that this framework can answer. Students may (correctly) interpret lower grades in treated departments as indicating that she did not learn as much as she might have, and the student may hold her professors responsible for that lack of success. The estimated impact of the policy on students' evaluations of professors does highlight the fact that there is little incentive for individual faculty, in the absence of an explicit policy, to deflate their own grades.

This paper leaves unanswered several of the most interesting questions about the impact of an anti-grade inflation policy. First, we have no way of knowing in this framework whether student learning was enhanced. Second, we cannot say whether there were long-term adverse consequences of this policy of unilateral disarmament for students of AC relative to students at other institutions. Clearly, prospective students, current students, and recent alums all worry that this is the case, leading to pressure to reverse the policy. If grade inflation is a systemic problem leading to inefficient allocation of resources on scales both large and small, then systemic policies are in order.

VI. References

Achen, Alexandra C. and Paul N. Courant, "What Are Grades Made Of?," *Journal of Economic Perspectives*, vol. 23 no. 3, Summer 2009, pp. 77-92.

Babcock, Philip, "Real Costs of Nominal Grade Inflation? New Evidence from Student Course Evaluations," *Economic Inquiry*, vol. 48, no. 4, October 2010, pp. 983-996.

Bar, Talia, Vrinda Kadiyali, and Asaf Zussman, "Grade Information and Grade Inflation: The Cornell Experiment," Journal of Economic Perspectives, vol. 23, no. 3, Summer 2009, pp. 93-108.

Braga, Michela, Marco Paccagnella, and Michele Pellizzari, "Evaluating Students' Evaluations of Professors," CEPR working paper no. 384, 2011.

Carrell, Scott E., and James E. West, "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors," Journal of Political Economy, vol. 118, no. 3, 2010, pp. 409-432.

Chan, William, Li Hao, and Wing Suen, "A Signalling Theory of Grade Inflation, *International Economic Review*, vol. 48, no. 3, August 2007, pp. 1065-1090.

Elzinga, Kenneth G., and Daniel O. Melaugh, "35,000 Principles of Economics Students: Some Lessons Learned," *Southern Economic Journal*, vol. 76, no. 1, 2009, pp. 32-46.

Espenshade, Thomas, J., Chang Y. Chung, and Joan L. Walling, "Admissions Preferences for Minority Students, Athletes, and Legacies at Elite Universities," Social Science Quarterly, vol. 85, no. 5, 2004.

Franz, Wan-Ju Iris, "Grade Inflation Under the Threat of Students' Nuisance: Theory and Evidence," *Economics of Education Review*, vol. 29, 2010, pp. 411-422.

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

Krautman, Anthony C., and William Sander, "Grades and Student Evaluations of Teachers, "Economics of Education Review, vo. 18, 1999, pp. 59-63.

Love, David A. and Matthew J. Kotchen, "Grades, Course Evaluations, and Academic Incentives," *Eastern Economic Journal*, vol. 36, 2010, pp. 151-163.

McEwan, Patrick and Kristen A. Soderberg, "Roommate Effects on Grades: Evidence from First-Year Housing Assignments, Research in Higher Education, vol. 47, no. 3, May 2006, pp. 347-370.

Nelson, Jon P., and Kathleen A. Lynch, "Grade Inflation, Real Income, Simultaneity, and Teaching Evaluations, Journal of Economic Education, Winter 1984, pp. 21-37.

Pressman, Steven, "The Economics of Grade Inflation," *Challenge*, vol. 50, no. 5, 2007, pp. 93-102.

Rojstaczer, Stuart and Christopher Healy, "Grading in American Colleges and Universities," *Teachers College Record*, March 4, 2010. http://www.tcrecord.org ID Number 15928

Rosovsky, Henry and Matthew Hartley, "Are We Doing the Right Thing? Grade Inflation and Letters of Recommendation," Occasional Paper, The American Academy of Arts and Sciences, 2002. http://www.amacad.org/publications/monographs/Evaluation_and_the_Academy.pdf

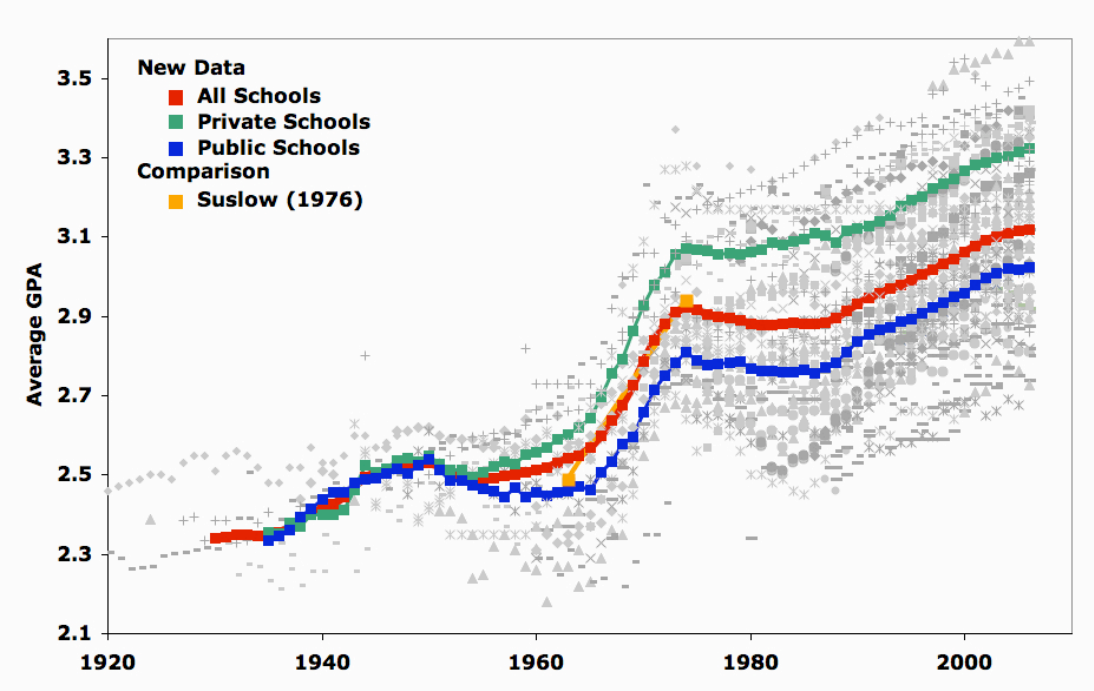
Sabot, Richard and John Wakeman-Linn, "Grade Inflation and Course Choice, " *Journal of Economic Perspectives*, vol. 5, no. 1, Winter 1991, pp. 159-170.

Wongsurawat, Winai, "Does Grade Inflation Affect the Credibility of Grades? Evidence from US Law School Admissions," *Education Economics*, vol. 17, no. 4, December 2009, pp. 523-534.

Zangenehzadeh, Hamid, "Grade Inflation: A Way Out," *Journal of Economic Education*, 1988, pp. 217-226.

Zimmerman, David J., "Peer Effects in Academic Outcomes: Evidence from a Natural Experiment," Review of Economics and Statistics, vol. 85, no. 1, 2003 pp. 9-23.

Figure 1:



Source: Rojstaczer and Healy (2010)

Figure 1. Average GPA over the time period 1930-2006 as a function of school type. Grey dots represent individual data points. Colored squared represent the mean GPA for each school type over time. Suslow (1976) shown for comparison.

Figure 2

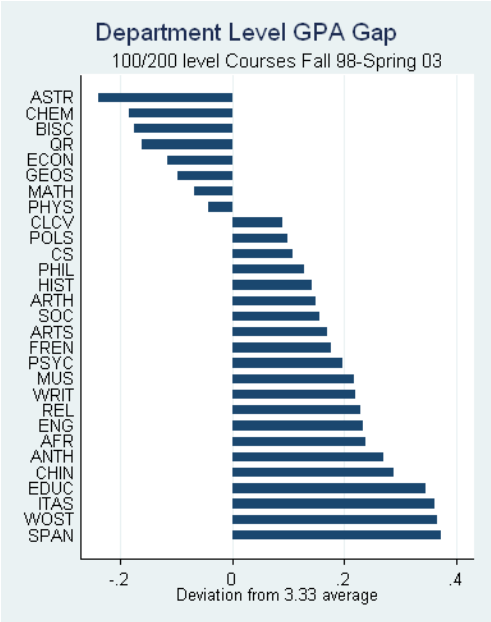


Figure 3:

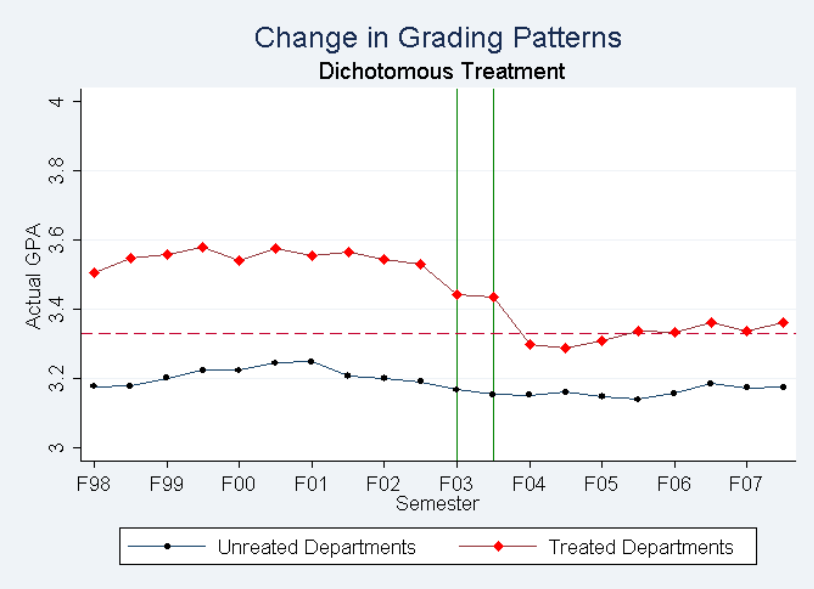

Change in Grading Patterns
Dichotomous Treatment
Actual GPA
3
3.2
3.4
3.6
3.8
4
F98
F99
F00
F01
F02
F03
F04
F05
F06
F07
Semester
Unreated Departments
Treated Departments

Table 1: Impact of Policy on Grades Awarded to Students

	(1)	(2)	(3)
Post Policy	-0.048***	-0.094***	
	(0.013)	(0.014)	
Treated	0.341***	0.300***	
	(0.026)	(0.033)	
Treated*Post	-0.175***	-0.172***	-0.174***
	(0.020)	(0.020)	(0.021)
SATM/100		0.0720***	
		(0.018)	
SATV/100		0.0356***	
		(0.00867)	
Black		-0.235***	
		(0.036)	
Latino		-0.170***	
		(0.031)	
Foreign		0.104***	
		(0.016)	
Asian		-0.0576***	
		(0.0143)	
Non-Traditional		0.084	
		(0.200)	
First Generation		-0.003	
		(0.009)	
Legacy		-0.028***	
		(0.005)	
Class Size/10		-0.009**	-0.014**
		(0.003)	(0.006)
Constant	3.208***	1.863***	3.090***
	(0.018)	(0.139)	(0.036)
Observations	104,454	104,454	104,454
R-squared	0.074	0.136	0.469
Other Controls	NO	YES	NO
Department FE	NO	NO	YES
Student FE	NO	NO	YES
Semester FE	NO	NO	YES

Robust s.e. (clustered by department) in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Other controls include indicator variables for
Humanities, Social Science or Science/Math courses
as well as an indicator variable for 200-level courses.

TABLE 2: Impact of Policy on Faculty Grading Behavior

VARIABLES	(1) Straight A	(2) A or A-	(3) B+	(4) B+, B or B-	(5) C- or Below
Post Policy	0.0197	0.0170	-0.0246*	-0.0277	0.0108*
	(0.0151)	(0.0144)	(0.0139)	(0.0186)	(0.00571)
Treated	0.150***	0.258***	-0.00299	-0.153***	-0.0347***
	(0.0242)	(0.0262)	(0.0211)	(0.0220)	(0.0106)
Treated*Post	-0.140***	-0.179***	0.0712***	0.166***	0.00842
	(0.0211)	(0.0178)	(0.0123)	(0.0215)	(0.00635)
Constant	0.175***	0.406***	0.207***	0.464***	0.0618***
	(0.0254)	(0.0332)	(0.0254)	(0.0250)	(0.0137)
Observations	104,454	104,454	104,454	104,454	116,119
R-squared	0.291	0.351	0.123	0.234	0.219
Other Controls	YES	YES	YES	YES	YES
Student Fixed Effects	YES	YES	YES	YES	YES

Robust s.e. (clustered by department) in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Other controls include indicator variables for
Humanities, Social Science or Science/Math courses
as well as an indicator variable for 200-level courses.

TABLE 3: Differential Impact, by Race, of Policy on Student Grades

	(1)	(2)	(3)	(4)	(5)
GROUP	All	Black	Latino	Low QR 5%	Low SATV 5%
Post Policy	-0.001	-0.005	0.004	0.001	-0.002
	(0.017)	(0.016)	(0.017)	(0.019)	(0.016)
Treated	0.352***	0.330***	0.344***	0.333***	0.343***
	(0.040)	(0.040)	(0.040)	(0.048)	(0.039)
Treated*Post	-0.167***	-0.156***	-0.173***	-0.157***	-0.162***
	(0.021)	(0.020)	(0.020)	(0.022)	(0.021)
Treat*Group		0.417***	0.203***	0.294***	0.142***
		(0.046)	(0.039)	(0.093)	(0.041)
Post*Group		0.103	-0.075	-0.072	0.100**
		(0.070)	(0.055)	(0.084)	(0.046)
Treat*Post*Group		-0.188***	0.089***	-0.086**	-0.122***
		(0.040)	(0.031)	(0.041)	(0.038)
Constant	3.231***	3.232***	3.228***	3.228***	3.243***
	(0.054)	(0.054)	(0.054)	(0.062)	(0.051)
Observations	104,454	104,454	104,454	90,196	98,186
R-squared	0.136	0.141	0.139	0.136	0.136
Other Controls	YES	YES	YES	YES	YES

Standard errors (clustered by department) in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Other controls include indicator variables for Humanities, Social Science
or Science/Math courses as well as an indicator variable for 200-level courses.

Table 4: Impact of Policy on Academic Honors

	(1)	(2)	(3)	(4)
	Summa Cum Laude	*Magna Cum Laude*	*Cum Laude*	*Phi Beta Kappa*
Treated*Post	-0.005	-0.037***	-0.008	-0.003
	(0.014)	(0.010)	(0.024)	(0.021)
SATM/100	0.022***	0.016*	-0.002	0.043***
	(0.005)	(0.009)	(0.010)	(0.008)
SATV/100	0.013	0.038***	0.003	0.047***
	(0.008)	(0.011)	(0.018)	(0.006)
Black	-0.009	-0.095***	-0.128***	-0.031**
	(0.006)	(0.015)	(0.022)	(0.012)
Latino	-0.003	-0.091***	-0.075***	-0.032*
	(0.009)	(0.012)	(0.021)	(0.018)
Asian	-0.040***	-0.057***	-0.011	-0.067***
	(0.012)	(0.016)	(0.010)	(0.016)
Foreign	0.049***	0.017	0.017	0.078***
	(0.012)	(0.018)	(0.025)	(0.019)
Non-Traditional Aged	0.012	-0.061*	-0.075*	-0.002
	(0.020)	(0.032)	(0.042)	(0.025)
First Generation	0.014	0.010	-0.005	0.011
	(0.009)	(0.014)	(0.018)	(0.008)
Legacy	0.010	-0.032**	-0.032*	-0.022
	(0.009)	(0.014)	(0.016)	(0.013)
Constant	-0.468***	-0.552***	-0.291*	-0.979***
	(0.149)	(0.118)	(0.139)	(0.099)
Observations	6,738	6,738	6,738	6,738
R-squared	0.054	0.069	0.037	0.079

Robust s.e. (clustered by graduation year) in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Because the standard for receiving Academic Honors was changed in 1999-2000, data from students graduating prior to the Class of 2000 were omitted in constructing this table

Table 5: Impact of Policy on Student Sorting by Quality

	(1)	(2)	(3)
	QR Score	Lowest 5% on QR Score	Lowest 5% on SATV
Treated*Post	-0.185	-0.003	0.002
	(0.129)	(0.007)	(0.003)
SATM/100	2.112***	-0.085***	0.002
	(0.039)	(0.006)	(0.002)
SATV/100	0.167***	-0.012***	-0.133***
	(0.021)	(0.002)	(0.006)
Black	-0.658***	0.123***	0.022***
	(0.085)	(0.010)	(0.008)
Latino	-0.390***	0.034***	-0.003
	(0.060)	(0.007)	(0.004)
Foreign	0.565***	-0.010***	0.088***
	(0.046)	(0.003)	(0.005)
Asian	0.082***	0.002	-0.008***
	(0.026)	(0.002)	(0.001)
Non-Traditional Aged	-1.868	0.182	-0.074**
	(1.449)	(0.202)	(0.033)
First Generation	0.078**	0.001	0.005
	(0.0308)	(0.003)	(0.003)
Legacy	-0.410***	0.011***	0.005**
	(0.027)	(0.003)	(0.002)
Constant	-7.434***	0.973***	1.088***
	(0.421)	(0.059)	(0.047)
Observations	116,374	99,947	109,179
R-squared	0.745	0.155	0.277
Other Controls	YES	YES	YES
Department FE	YES	YES	YES
Semester FE	YES	YES	YES

Robust s.e. (clustered by department) in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Other controls include indicator variables for Humanities, Social Science
or Science/Math courses as well as an indicator variable for 200-level courses.

Table 6: Impact of Policy on *Department Enrollments*

	(1) Total Enrollment	(2) Total Enrollment	(3) ln(Total Enrollment)	(4) ln(Total Enrollment)
Post Policy	25.73		0.0837	
	(20.21)		(0.0557)	
Treated Department	-65.78***		-0.202***	
	(8.723)		(0.0274)	
Treated*Post	-51.76***	-54.75***	-0.191***	-0.203***
	(12.68)	(13.30)	(0.0332)	(0.0346)
Constant	374.1***	347.1***	5.868***	5.778***
	(12.24)	(2.252)	(0.0380)	(0.00595)
Observations	342	342	342	342
R-squared	0.108	0.901	0.121	0.883
Clustered SE	YES	YES	YES	YES
Department FE	NO	YES	NO	YES
Semester FE	NO	YES	NO	YES

Standard errors (clustered by term) in parentheses
*** p<0.01, ** p<0.05, * p<0.1

To avoid giving undue influence to enrollments in smaller departments, we used the following categories:
HUMANITIES: English, Spanish, Art History, French, Studio Art, Other Languages, Other Humanities
SOCIAL SCIENCES: Economics, Psychology, Political Science, Philosophy, Religion, Other Social Sciences
SCIENCES: Biological Sciences, Mathematics, Chemistry, Computer Science, Other Sciences

TABLE 7A: Impact of Policy on "Strongly Recommend"

	(1)	(2)	(3)	(4)
Post Policy	0.052***	0.059***	0.062***	
	(0.011)	(0.010)	(0.010)	
Treated	0.100***	0.102***	0.0649***	
	(0.023)	(0.021)	(0.022)	
Treated*Post	-0.048***	-0.047***	-0.036**	-0.0384**
	(0.015)	(0.015)	(0.016)	(0.015)
Asian Faculty		-0.019	-0.024	-0.040*
		(0.023)	(0.024)	(0.022)
Black Faculty		-0.145***	-0.158***	-0.151***
		(0.032)	(0.035)	(0.044)
NTT Faculty		-0.000	0.009	-0.017
		(0.032)	(0.034)	(0.036)
Visiting Faculty		-0.100***	-0.101***	-0.104***
		(0.023)	(0.024)	(0.027)
Tenured Faculty		-0.027*	-0.031*	-0.031*
		(0.015)	(0.017)	(0.017)
Female Faculty		-0.008	-0.005	0.007
		(0.017)	(0.018)	(0.018)
Course GPA			0.094***	0.087***
			(0.033)	(0.032)
200 Level			0.036***	0.031***
			(0.011)	(0.011)
Class Size/10			-0.003	-0.002
			(0.003)	(0.003)
Constant	0.533***	0.568***	0.260**	0.382***
	(0.0204)	(0.0183)	(0.103)	(0.103)
Observations	5,416	5,413	5,277	5,277
R-squared	0.021	0.057	0.072	0.131
Clustered SE	YES	YES	YES	YES
Semester FE	NO	NO	NO	YES
Department FE	NO	NO	NO	YES
Robust standard errors (clustered by department) in parentheses				

*** p<0.01, ** p<0.05, * p<0.1

TABLE 7B: Impact of Policy on "Strongly Recommend" or "Recommend"

	(1)	(2)	(3)	(4)
Post Policy	0.049***	0.054***	0.056***	
	(0.009)	(0.008)	(0.008)	
Treated	0.073***	0.077***	0.054***	
	(0.015)	(0.015)	(0.016)	
Treated*Post	-0.040***	-0.039***	-0.031**	-0.032***
	(0.011)	(0.011)	(0.012)	(0.012)
Asian Faculty		-0.004	-0.007	-0.021
		(0.015)	(0.015)	(0.013)
Black Faculty		-0.097***	-0.100***	-0.095***
		(0.023)	(0.024)	(0.028)
NTT Faculty		-0.011	-0.001	-0.017
		(0.017)	(0.018)	(0.019)
Visiting Faculty		-0.076***	-0.074***	-0.080***
		(0.016)	(0.016)	(0.017)
Tenured Faculty		-0.0138	-0.0157	-0.0194**
		(0.00966)	(0.0101)	(0.00945)
Female Faculty		0.000	0.001	0.009
		(0.008)	(0.008)	(0.009)
Course GPA			0.059***	0.058***
			(0.017)	(0.017)
200 Level			0.026***	0.022***
			(0.009)	(0.007)
Class Size/10			0.000	0.001
			(0.003)	(0.003)
Constant	0.797***	0.816***	0.614***	0.696***
	(0.014)	(0.015)	(0.055)	(0.061)
Observations	5,416	5,413	5,277	5,277
R-squared	0.026	0.062	0.075	0.123
Clustered SE	YES	YES	YES	YES
Semester FE	NO	NO	NO	YES
Department FE	NO	NO	NO	YES

Robust standard errors (clustered by department) in parentheses

*** p<0.01, ** p<0.05, * p<0.1

TABLE 7C: Impact of Policy on "Neutral"

	(1)	(2)	(3)	(4)
Post Policy	-0.020***	-0.023***	-0.024***	
	(0.004)	(0.0030)	(0.003)	
Treated	-0.038***	-0.039***	-0.028***	
	(0.009)	(0.008)	(0.010)	
Treated*Post	0.017***	0.016***	0.013*	0.015**
	(0.00585)	(0.00558)	(0.00641)	(0.00620)
Asian Faculty		-0.000	0.001	0.008
		(0.008)	(0.009)	(0.008)
Black Faculty		0.046***	0.048***	0.045***
		(0.013)	(0.013)	(0.016)
NTT Faculty		0.001	-0.003	0.006
		(0.010)	(0.010)	(0.011)
Visiting Faculty		0.042***	0.042***	0.045***
		(0.010)	(0.010)	(0.011)
Tenured Faculty		0.010*	0.011*	0.013**
		(0.006)	(0.006)	(0.006)
Female Faculty		-0.001	-0.002	-0.007
		(0.006)	(0.006)	(0.006)
Course GPA			-0.0283**	-0.0255**
			(0.011)	(0.011)
200 Level			-0.013**	-0.009**
			(0.005)	(0.005)
Class Size/10			0.000	0.000
			(0.001)	(0.001)
Constant	0.126***	0.115***	0.211***	0.161***
	(0.008)	(0.008)	(0.037)	(0.037)
Observations	5,416	5,413	5,277	5,277
R-squared	0.017	0.040	0.048	0.087
Clustered SE	YES	YES	YES	YES
Semester FE	NO	NO	NO	YES
Department FE	NO	NO	NO	YES

Robust standard errors (clustered by department) in parentheses
*** p<0.01, ** p<0.05, * p<0.1

TABLE 7D: Impact of Policy on "Not Recommend"

	(1)	(2)	(3)	(4)
Post Policy	-0.029***	-0.031***	-0.032***	
	(0.006)	(0.006)	(0.006)	
Treated	-0.035***	-0.038***	-0.026***	
	(0.007)	(0.007)	(0.008)	
Treated*Post	0.023***	0.022***	0.019***	0.017**
	(0.007)	(0.007)	(0.007)	(0.007)
Asian Faculty		0.004	0.006	0.014**
		(0.007)	(0.007)	(0.007)
Black Faculty		0.027	0.036	0.042*
		(0.020)	(0.023)	(0.024)
NTT Faculty		0.010	0.004	0.011
		(0.008)	(0.008)	(0.009)
Visiting Faculty		0.034***	0.032***	0.036***
		(0.007)	(0.008)	(0.008)
Tenured Faculty		0.004	0.004	0.007
		(0.004)	(0.005)	(0.004)
Female Faculty		0.001	0.001	-0.002
		(0.004)	(0.004)	(0.004)
Course GPA			-0.031***	-0.032***
			(0.008)	(0.009)
200 Level			-0.013***	-0.013***
			(0.004)	(0.003)
Class Size/10			0.000	-0.001**
			(0.001)	(0.0006)
Constant	0.077***	0.069***	0.175***	0.144***
	(0.007)	(0.008)	(0.026)	(0.033)
Observations	5,416	5,413	5,277	5,277
R-squared	0.021	0.052	0.062	0.100
Clustered SE	YES	YES	YES	YES
Semester FE	NO	NO	NO	YES
Department FE	NO	NO	NO	YES

Robust standard errors (clustered by department) in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Appendix Table 1a: Summary Statistics

Observation Level: Student-Course-Semester						
	Pre Policy			Post Policy		
	Mean	Std. Dev	N	Mean	Std. Dev	N
Course Type						
Humanities Course	0.348	0.476	84,780	0.345	0.475	64,937
Social Science Course	0.403	0.491	84,780	0.411	0.492	64,937
Science Course	0.229	0.420	84,780	0.226	0.418	64,937
Interdepartmental Course	0.020	0.139	84,780	0.019	0.135	64,937
100-Level Course	0.375	0.484	84,780	0.351	0.477	64,937
200-Level Course	0.459	0.498	84,780	0.475	0.499	64,937
300-Level Course	0.167	0.373	84,780	0.174	0.379	64,937
Enrolled in Classes of Size						
< 10	0.069	0.253	84,780	0.077	0.266	64,937
$>= 10$ and $<= 25$	0.541	0.498	84,780	0.599	0.490	64,937
> 25	0.390	0.488	84,780	0.325	0.468	64,937
Grading Choice						
Took for Grade	0.900	0.299	84,780	0.884	0.320	64,937
Credit-Non Election	0.082	0.274	84,780	0.079	0.270	64,937
Withdraw	0.012	0.111	84,780	0.021	0.144	64,937
Incomplete	0.005	0.073	84,780	0.015	0.123	64,937
Course Grade	3.482	0.544	76,337	3.326	0.564	57,409
Grade Distribution						
A	0.293	0.455	76,337	0.174	0.379	57,409
A-	0.291	0.454	76,337	0.253	0.435	57,409
B+	0.203	0.402	76,337	0.252	0.434	57,409
B	0.115	0.319	76,337	0.176	0.381	57,409
B-	0.047	0.216	76,337	0.073	0.260	57,409
C+	0.021	0.142	76,337	0.031	0.173	57,409
C or below	0.031	0.174	76,337	0.041	0.199	57,409
Grades by Race						
Black	3.199	0.689	4,194	3.013	0.667	2,850
Asian	3.454	0.549	18,067	3.323	0.554	14,582
Biracial	3.456	0.567	2,576	3.245	0.597	2,768
International	3.530	0.537	4,428	3.454	0.524	4,387
Latino	3.310	0.636	3,829	3.110	0.647	3,220
White	3.546	0.491	36,493	3.373	0.533	24,513
Grades by Division						
Humanities	3.593	0.430	26,850	3.393	0.475	18,881
Social Sciences	3.496	0.507	31,645	3.339	0.532	24,641
Science	3.264	0.703	16,307	3.197	0.706	12,780

NOTE: Pre-Policy is data from Fall 1999 to Spring 2003. Post Policy is data from Fall 2004 to Spring 2008. We omit the two semesters (Fall 2003 and Spring 2004) during which the grading policy at AC was being proposed and voted on

Appendix Table 1b: Summary Statistics

Observation Level: Student						
	Pre Policy			**Post Policy**		
	Mean	Std. Dev	N	Mean	Std. Dev	N
Demographics						
Black	0.055	0.228	4,473	0.051	0.220	3,663
Asian	0.220	0.415	4,473	0.247	0.432	3,663
International	0.055	0.228	4,473	0.079	0.270	3,663
Latina	0.054	0.227	4,473	0.055	0.228	3,663
Native American	0.004	0.056	4,473	0.004	0.061	3,663
White/Caucasian	0.500	0.500	4,473	0.406	0.491	3,663
First Generation	0.078	0.261	4,473	0.111	0.315	3,663
Legacy	0.105	0.306	4,473	0.115	0.319	3,663
Non-Traditional Aged	0.034	0.182	4,473	0.018	0.132	3,663
Academic Background/Performance						
Cumulative GPA	3.499	0.351	4,464	3.315	0.384	3,662
SAT-Verbal	667.40	75.11	4,218	689.31	68.63	3,410
SAT-Math	659.50	69.86	4,218	679.26	65.28	3,410
SAT-Writing	659.14	89.69	3,890	688.28	74.72	2,202
ACT	28.82	3.04	878	29.34	2.85	1,012
QR Score	12.52	2.99	2,689	12.69	2.97	3,622
Academic Honors at Graduation						
Summa Cum Laude	0.052	0.221	3,870	0.023	0.149	2,518
Magna Cum Laude	0.205	0.404	3,870	0.097	0.296	2,518
Cum Laude	0.248	0.432	3,870	0.160	0.367	2,518
Phi Beta Kappa	0.106	0.308	3,870	0.112	0.316	2,518

NOTE: Students are only observed once (in the last semester for which we have data) for the purposes of constructing this table. To allow the policy to materially impact outcomes, Pre-Policy now counts anyone who graduated in the Class of 2004 or prior. Post Policy counts students in the Class of 2005 or later

Appendix Table 1c: Summary Statistics

Observation Level: Course						
	Pre Policy			Post Policy		
	Mean	Std. Dev	N	Mean	Std. Dev	N
Course Type						
100-Level	0.304	0.460	4,722	0.289	0.453	3,966
200-Level	0.449	0.497	4,722	0.456	0.498	3,966
300-Level	0.247	0.431	4,722	0.255	0.436	3,966
Humanities	0.409	0.492	4,722	0.408	0.492	3,966
Social Sciences	0.369	0.483	4,722	0.374	0.484	3,966
Science/Math	0.201	0.401	4,722	0.196	0.397	3,966
Enrollments						
Total Enrollment	18.16	11.60	4,722	17.08	9.53	3,966
Minority Enrollment	1.974	2.301	4,722	1.817	1.95	3,966
Classes of Size						
< 10	0.203	0.402	4,722	0.200	0.400	3,966
$>= 10$ and $<= 25$	0.594	0.491	4,722	0.630	0.483	3,966
> 25	0.203	0.402	4,722	0.170	0.375	3,966
Grades						
Course Grade	3.518	0.258	4,677	3.372	0.240	3,849
Conforming Course	0.296	0.457	3,230	0.575	0.494	2,811
Humanities	0.122	0.327	1,299	0.435	0.496	1,152
Social Sciences	0.278	0.448	1,141	0.591	0.492	976
Science/Math	0.653	0.476	735	0.813	0.389	627

NOTE: A Conforming course is a 100-level or 200-level course with more than 10 enrolled students that has a mean grade that does not exceed 3.33.

Appendix Table 1d: Summary Statistics

Observation Level: Course Evaluation of a Faculty Member						
	Pre Policy			**Post Policy**		
	Mean	Std. Dev	N	Mean	Std. Dev	N
Fraction Giving Professor a Rating of						
"Strongly Recommend"	0.627	0.260	4,317	0.649	0.244	4,050
"Recommend" + "Strongly Recommend"	0.865	0.179	4,317	0.886	0.151	4,050
"Neutral"	0.090	0.121	4,317	0.080	0.109	4,050
"Do Not Recommend"	0.045	0.097	4,317	0.034	0.074	4,050
Fraction Giving Course a Rating of						
"Strongly Recommend"	0.535	0.238	4,317	0.554	0.230	4,050
"Recommend" + "Strongly Recommend"	0.851	0.162	4,317	0.868	0.145	4,050
"Neutral"	0.107	0.121	4,317	0.100	0.113	4,050
"Do Not Recommend"	0.043	0.082	4,317	0.032	0.066	4,050

Appendix Figure 1: