Journal of Economic Perspectives—Volume 28, Number 3—Summer 2014—Pages 189–204

The Effects of an Anti-Grade-Inflation Policy at Wellesley College[†]

Kristin F. Butcher, Patrick J. McEwan, and Akila Weerapana

Average grades in colleges and universities in the United States are markedly higher than they were several decades ago. In 1960, the average grade point average for all private and public institutions was about 2.4, or a little above a C+. By 2006, this number was about 3.0, or roughly a B, and even higher in private institutions (Rojstaczer and Healy 2010). Courses in the humanities usually have the highest grades, science and math courses have the lowest grades, and social sciences fall somewhere in the middle (Rojstaczer and Healy 2010). Since the highest available grade is usually an A, this means that grade inflation has gone hand-in-hand with compression at the top of the distribution, particularly in the humanities.

If grades are the fundamental way in which students, administrators, graduate schools, and employers receive information about an individual's absolute and relative abilities, then grade inflation and compression masks valuable information and distorts choices. Based in part on grades, students make choices about how hard to work (Babcock 2009), courses (Sabot and Wakemann-Linn 1991), majors, and careers. Administrators make choices about where to allocate academic support services, graduate schools make choices about whom to admit (Wongsurawat 2009), and employers make choices about whom to hire (Chan, Hao, and Suen 2007).

In the early 2000s, the faculty and administration at Wellesley College—a selective, small, private, women's liberal arts college outside of Boston,

■ *Kristin F. Butcher is the Marshall I. Goldman Professor of Economics, Patrick J. McEwan is Professor of Economics, and Akila Weerapana is Associate Professor of Economics, all at Wellesley College, Wellesley, Massachusetts. Butcher is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Butcher is the corresponding author at kbutcher@wellesley.edu.*

[†] To access the data Appendix and disclosure statements, visit
http://dx.doi.org/10.1257/jep.28.3.189

doi=10.1257/jep.28.3.189

Massachusetts—concluded that grade inflation and compression was causing a number of these problems, potentially undermining the institution's credibility and reputation. Thus, the College implemented the following policy in Fall 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+. The rule has some latitude. If a professor feels that the students in a given section were particularly meritorious, that professor can write a letter to the administration explaining the reasons for the average grade exceeding the cap. Grades by department are reported to administrators and faculty during Academic Council meetings, the main governing body at the institution, so that peers can see if some departments are regularly violating the policy. Penalties are left to the discretion of the administration. The grading policy is detailed on the Registrar's web page, and is known to students, prospective students, and alums.

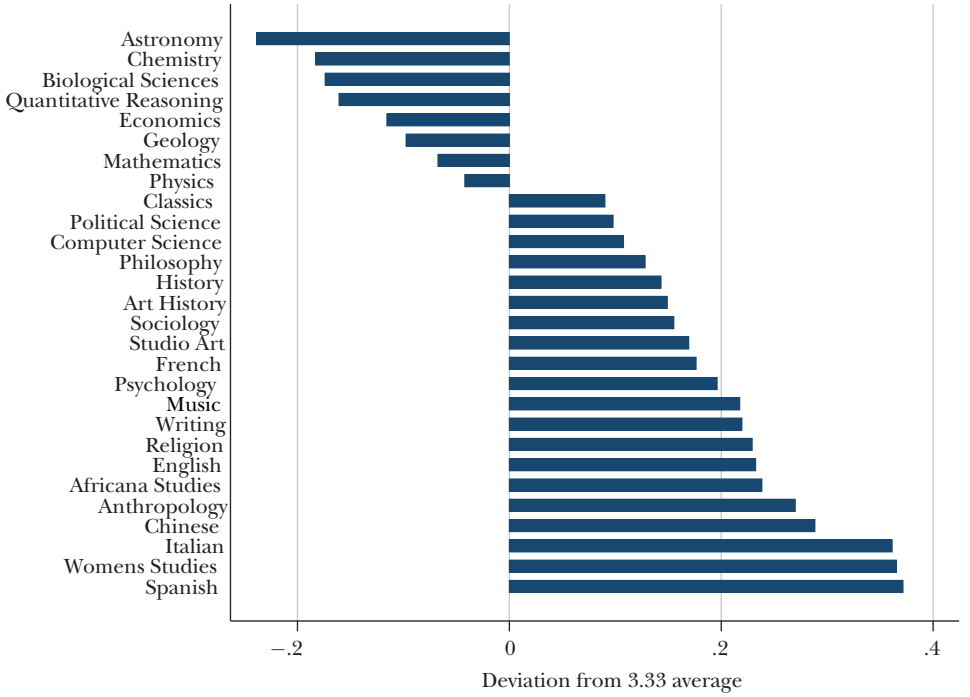
Since grading patterns at Wellesley College were similar to those across academia, only courses in high-grading departments in the humanities and social sciences (except economics) needed to change grading practices in order to comply. In this paper, we evaluate the consequences of the policy by comparing outcomes in departments that were obligated to lower their grades with outcomes in departments that were not. The policy had an immediate effect, bringing average grades down in the previously high-grading departments. Faculty complied by reducing compression at the top of the grade distribution, but there is little evidence that they increased the use of very low grades. We also examine the impact of the change in grading policy for different subgroups. For African-American students and students with low initial test scores, the gap in grade point average with other students increased in the departments where grades were reduced. There is little evidence of a change in sorting by student quality—as measured by students' initial test scores—into capped and uncapped courses. However, the number of students enrolled in courses in capped departments and the number of majors in these departments show relative declines. It also appears that students reduced their evaluations of their professors' performance in response to the change in the grading policy. The implications of these changes will be discussed in the concluding section.

How the Cap Affected Grades

Figure 1 shows how department and program average grades deviated from the policy target in the years prior to the adoption of the 3.33 cap (in 100- and 200-level courses with more than 10 students, from Fall 1998 to Spring 2003). At Wellesley, as at other institutions, it is largely the sciences that were low-grading and other departments that were high-grading.[1] Average grades were below the cap

[1] Given that these patterns persist across institutions, it is plausible that there are inherent differences in the cost of giving lower grades across disciplines. Achen and Courant (2009) and Franz (2010) have suggested that if typical assessment mechanisms are costlier in the humanities than in the sciences—for

Figure 1
Pre-policy Grade Differences across Departments



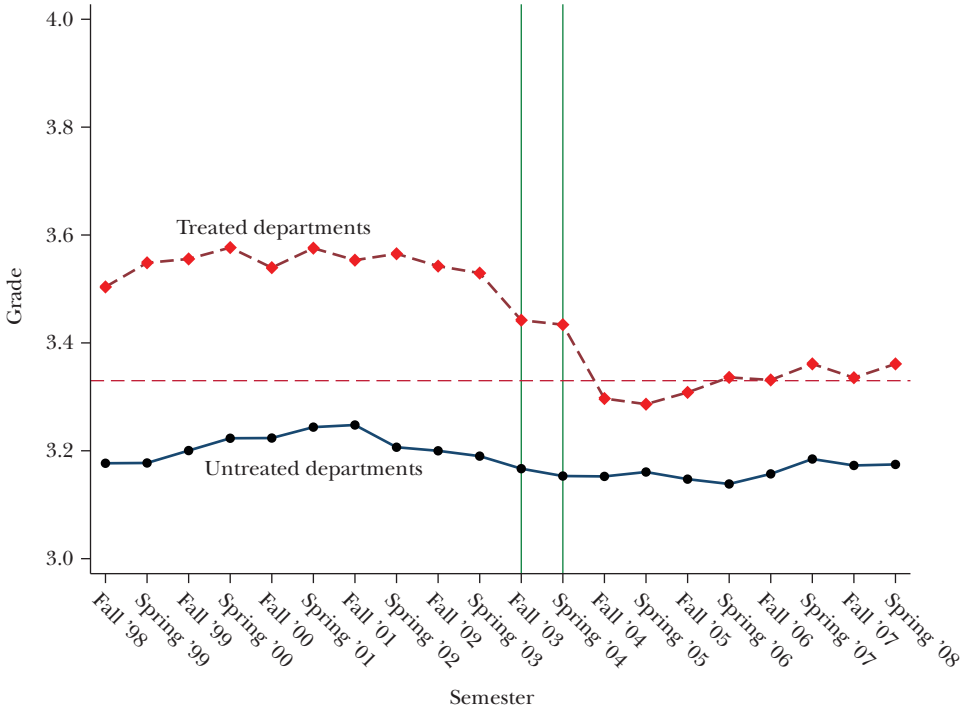
Source: Authors using student-transcript-level data from Wellesley College.
Note: Figure 1 shows how department and program average grades at Wellesley deviated from the policy target in the years prior to the adoption of the 3.33 cap (in 100- and 200-level courses with more than 10 students, from Fall 1998 to Spring 2003).

in Astronomy, Chemistry, Biological Sciences, Quantitative Reasoning, Economics, Geology, Mathematics, and Physics. Other departments had grades above the cap, with at least three departments exceeding it by a third of a letter grade.

Figure 2 shows what happened to average grades in "treated" and "untreated" departments from Fall 1998 to Spring 2008. Grades dropped a bit in the soon-to-be-treated departments in Fall 2003 when the policy was publicly debated (and so data from the academic year 2003–2004 are dropped from subsequent analyses). After the adoption of the policy, treated departments lowered their grades enough to comply.

example, if it is more difficult both to see that an argument in a paper is unsupported than to see that an answer to a math problem is incorrect, and to communicate that to students—then faculty in some disciplines will have more incentive to inflate grades. Smaller classes, and the nature of the work, may make a low grade feel more personal, and thus more difficult to both give and receive, in a humanities class than in a science class. 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, potentially by giving higher grades.

Figure 2
Change in Grading Patterns over Time
(Fall 1998 to Spring 2008)



Source: Authors using student-transcript-level data from Wellesley College.
Notes: The two vertical lines indicate the semesters when the policy was introduced and then implemented. The horizontal line shows the 3.33 average grade that the policy used as the new standard.

There is some evidence of backsliding in treated departments: since Spring 2006, average grades in the treated departments slightly exceeded the cap. Nonetheless, by and large, the policy was effective in its basic goal of lowering average grades in high-grading departments to the agreed-upon target.

To assess the effect of the policy on academic outcomes, we use a standard difference-in-differences methodology: that is, we compare the change in outcomes in the treated departments to the change in outcomes in the untreated departments. Under the assumption that factors affecting outcomes—other than the policy change—are the same in the treated and untreated departments, the approach will estimate the effect of the grading policy on a given outcome. It is a problem for this methodology if, during this period, there are differential changes across departments affecting the outcomes of interest that are separate from the grading policy. For example, if Wellesley attracted a particularly talented group of young scientists for the Fall 2004 term, then the underlying quality of students in the different

departments would be changing at the same time as the policy, and these changes may have their own independent effect on grades and other outcomes.

Alternatively, if the policy resulted in differential sorting across departments—perhaps because some students are more grade-sensitive than others—then there would, again, be an underlying change in the quality of students that might exacerbate the changes in outcomes across departments.[2] To the extent that these potential changes in student quality across departments are observable—for example, in baseline test scores—we can control for them by estimating the following regression model:

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

where Y is the outcome of interest; i indexes the individual, d the department, and t the semester or year; *PostPolicy* is a dummy for Fall 2004 and beyond; and *Treated* is a dummy variable equal to 1 for those departments with average grades above the cap prior to the policy. The coefficient of interest is β_3, because it indicates whether the outcomes for students in courses in the treated departments changed differentially before and after the policy was implemented.[3] X_{idt} is a vector of fixed and time-varying individual, course, and department characteristics. Standard errors are clustered at the department level.

Our main dataset is transcript-level data on student grades and courses from Fall 1998 to Spring 2008.[4] Wellesley College has a diverse and high-quality student body. Over the time period, 45 percent of the students identified as white, 22.8 percent identified as Asian, 5 percent as African American/black, and 5 percent as Latina. About 10 percent of students are first-generation college students, and about 10 percent are legacy students, related to an alum of the institution. Average SAT scores are high, consistent with Wellesley's status as a highly-selective institution. Roughly one-third of the courses were taken in the humanities and languages, 40 percent in the social sciences, and 23 percent in sciences/math, with the remainder in other programs. Wellesley has distribution requirements such that all students must take 100- and 200-level courses in all divisions in the college

[2] For example, work by Goldin (2013) shows that women students are less likely to major in economics than their male counterparts if they receive a low grade in introductory courses. This point will be discussed further in the conclusion, but it is worth noting here that grades remain higher in the humanities and non-economics social sciences even after the policy, and so students who were sensitive to low grades might still prefer to take courses in these disciplines.

[3] Figure 1 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.

[4] Summary statistics for students and courses pre- and post-policy adoption are shown in online Appendix Tables 1 and 2. There are 116,374 student-course-semester observations covering outcomes for over 8,000 students. The average numeric value for grades before the policy was about 3.5. About 8 percent 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, 1.3 percent 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.

Table 1
Impact of Policy on Grades Awarded to Students

	All (1)	*All* (2)	*All* (3)	*Black* (4)	*Latina* (5)	*Lowest 5% SATVerbal* (6)	*Low 5% QR* (7)
Treated	0.341***	0.300***					
	(0.026)	(0.033)					
PostPolicy	−0.048***	−0.094***					
	(0.013)	(0.014)					
*PostPolicy*Treated*	−0.175***	−0.172***	−0.174***	−0.363***	−0.081**	−0.299***	−0.334***
	(0.020)	(0.020)	(0.021)	(0.043)	(0.038)	(0.039)	(0.037)
SATMath/100		0.0720***					
		(0.018)					
SATVerbal/100		0.036***					
		(0.009)					
Observations	104,454	104,454	104,454	5,476	5,497	4,217	4,814
R^2	0.074	0.136	0.469	0.501	0.494	0.497	0.507
Demographics	No	Yes	No	No	No	No	No
Other controls	No	Yes	Yes	Yes	Yes	Yes	Yes
Department fixed effects	No	No	Yes	Yes	Yes	Yes	Yes
Student fixed effects	No	No	Yes	Yes	Yes	Yes	Yes
Semester fixed effects	No	No	Yes	Yes	Yes	Yes	Yes

Source: Authors using student-transcript-level data from Wellesley College.
Notes: In columns 3 through 7, department and semester fixed effects absorb the effects of *Treated* and *PostPolicy*, and student fixed effects absorb student demographics. *Black* is an indicator variable set equal to 1 for African American students and for students from other countries who self-identify as black. "Demographics" include indicator variables for race/ethnicity and for non-traditional-aged students, legacy students, and first-generation college students. "Other controls" include class size and an indicator variable for 200-level courses. Column 2 also includes indicator variables for humanities, social science, or science/math. Samples sizes for the last two columns differ because SAT Verbal scores are not available for all students and the Quantitative Reasoning test was not administered to the first set of graduating students in our sample. Robust standard errors are clustered by department.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

(humanities, science, and social sciences), meaning that all students will have experiences in both the treated and untreated departments.

Table 1 reports the estimated coefficients for regression models in which the dependent variable is the numeric value of the course grade. The first column reports the unadjusted difference-in-differences estimate (corresponding to the results in Figure 2). The second column controls for math and verbal SAT scores. It also includes dummy variables for black, Latina, international, Asian, non-traditional-aged student, first-generation college student, and legacy student as well as other controls for class size, whether a course is humanities, social science, or science/math and an indicator variable for whether it is a 200-level course. In column 3, student, department, and semester fixed effects are added. These control

for all fixed observable and unobservable differences across students, departments, and semesters; these fixed effects absorb the *Treated* and *PostPolicy* variables, as well as fixed student characteristics like race or SAT scores. Adding these controls has little effect on the estimate of β_3, the key parameter. Across columns 1–3, the coefficient on *PostPolicy*Treated* estimates the impact of the policy on grades to be about −0.17, or a relative drop of about a sixth of a letter grade in courses in treated departments.

Columns 4–7 report the estimated effect of the policy for various subgroups, using the same model as in column 3. The estimated drop in grades in treated departments is smaller for Latina students but much larger than average for black students (including African-Americans and foreign students who self-identify as black), those with low SAT verbal scores, and those with low Quantitative Reasoning scores. The results in columns 6 and 7 of Table 1 are very similar when black students are dropped from the sample, indicating that the results for those with low SAT verbal or low Quantitative Reasoning scores are not being driven by that group being disproportionately likely to be black. The estimated drops in grades for these groups are statistically different from the average drop at 1 percent level of significance, except for the Latina group, where the drop is statistically significant at the 5 percent level. If black students and those with low SAT verbal and Quantitative Reasoning scores have lower average outcomes than other students, and if, prior to the policy, those differences were being masked by grade compression in the treated departments, then the result is a natural outcome of the reduction in grade compression. If all students receive an A− or higher before the policy, and after the policy that is no longer possible, then the student who was receiving the "lowest A−" is likely to have her grades hit harder by the policy.[5] Whether that is a good or bad consequence of the policy will be taken up in the conclusion.

We now turn to examining in more detail how faculty complied with the new rule. Table 2 shows a series of linear probability estimates, 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 column 5, the outcome is equal to 1 if the student got a C− or below (including students who took the course "credit/non" and received no credit), and zero otherwise. Column 6 reports results for whether students withdrew or not, and column 7 for whether they elected to take the course credit/non. (The sample sizes are larger in columns 5–7 because students electing credit/non are included in the results.) The specifications control for department, semester, and student fixed effects and class size. The results are robust to the other specifications as well.

After the policy change, students were about 14 percentage points less likely to get a straight A in the treated departments. On average before the policy was

[5] Like black students, Latina students also tend to have lower grades on average than white students in the pre-policy period (even controlling for test scores), but the gaps are smaller, and Latina students and black students have a different distribution of courses. Since treatment "intensity" is different across departments (see Figure 1), the fact that Latina students occupied a different relative position in the grade distribution and have a different distribution of courses, may explain why their grades were less affected by the policy change.

Table 2
Impact of Policy on Faculty Grading Behavior

	Straight A (1)	*A or A−* (2)	*B+* (3)	*B+, B, or B−* (4)	*C− or below* (5)	*Withdraw* (6)	*Credit/non* (7)
*PostPolicy*Treated*	−0.141***	−0.184***	0.072***	0.167***	0.011	−0.007	0.069***
	(0.022)	(0.019)	(0.013)	(0.023)	(0.007)	(0.006)	(0.012)
Constant	0.077***	0.250***	0.243***	0.608***	0.049***	0.007	−0.039**
	(0.016)	(0.025)	(0.015)	(0.017)	(0.008)	(0.005)	(0.016)
Observations	104,454	104,454	104,454	104,454	116,374	116,374	116,374
R^2	0.317	0.370	0.132	0.250	0.227	0.114	0.136
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Department fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Student fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Semester fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Source: Authors using student-transcript-level data from Wellesley College.
Notes: Other controls include class size as well as an indicator variable for 200-level courses. C− or below includes students who took a course "credit/non" but did not pass. Samples sizes for the last three columns differ because students who did not take the course for a letter grade are included in the subsamples. Robust standard errors (clustered by department) in parentheses.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

enacted, about 29 percent of the grades in these departments were straight As, 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 essentially no change in the incidence of very low grades, C− or below, and no change in the share of students withdrawing from courses. This evidence, then, suggests that the policy alleviated 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 grades for other students while still meeting the policy goal of a 3.33 average.

There is evidence that the difference in the probability of taking a course credit/non changed between the treated and untreated departments. The change is driven by a precipitous drop in the credit/non elections in the departments that were *unaffected* by the grade cap. This effect, it turns out, can be traced to a policy adopted in Spring 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. Before this policy change, students were more likely to take courses in the low-grading departments using the credit/non option. Once students had to decide earlier, their election of credit/non dropped in the lower-grading

departments, converging to the same rates as in the higher-grading departments. The overall effect of this change 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 taking a course for credit only. Thus, the relative decline in grades in the departments that were affected by the 3.33 grade cap may be understated.

One oft-cited concern about persistent differences in grade levels across departments is that students majoring in high-grading departments are disproportionately rewarded with Latin honors. When we investigated these patterns, we found that the probability of graduating *summa cum laude* was not (significantly) differentially changed across departments by the policy. In order to graduate *summa*, a student needs a grade point average of 3.9 or above. At Wellesley, with its substantial distribution requirements, this standard means that a student needs to get nearly straight As in many different types of courses. These students are top performers across disciplines, and thus the *summa* students were not differentially affected by grading policies across departments and so were not differentially affected by the imposition of the grading cap.

However, the probability that a student graduated *magna cum laude* was notably affected by the policy. For the treated departments, about 20 percent of students received this designation in the pre-policy era; this fell to 16 percent after the policy was enacted. 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. These patterns are consistent with less grade compression in the upper portion of the grade distribution.[6]

Students' Choices of Courses and Majors

Grade inflation is often blamed for distorting students' choices across fields by misinforming them about their relative strengths. 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.[7] How did students alter their behavior in response to the grade cap policy?

[6] Details of these calculations are available from the authors. In addition, there was no relative change across treated and untreated departments in the probability of Phi Beta Kappa designation after the policy. Phi Beta Kappa is restricted to no more than 12 percent of a graduating class and is selected by a committee; the fact that there was no change suggests that the committee was implicitly taking into account that it was harder to reach a certain grade point average in some departments than others.

[7] One might also expect differential choices to lead to differential sorting across courses by student type. We investigated this by using initial test scores (for example, lowest 5 percent among Wellesley students on SAT verbal scores and lowest 5 percent on Quantitative Reasoning scores) as the outcome variable, and asking whether scores among students electing to take courses in treated and untreated departments changed after the policy. There was no statistically significant relative change in initial test scores. These results are available in Table 4 of the online Appendix available with this paper at http://e-jep.org.

Table 3
Impact of Policy on Enrollments and Majors

	Enrollment (1)	*ln(Enrollment)* (2)	*300-Level Enroll* (3)	*ln(300-Level Enroll)* (4)	*Majors* (5)	*ln(Majors)* (6)
*PostPolicy*Treated*	−51.26***	−0.186***	0.066	0.019	−8.406***	−0.307***
	(10.67)	(0.033)	(4.024)	(0.074)	(2.891)	(0.099)
Constant	348.8***	5.785***	66.16***	3.731***	35.80***	3.311***
	(11.93)	(0.041)	(2.819)	(0.105)	(1.784)	(0.077)
Observations	342	342	342	342	200	200
R^2	0.890	0.874	0.919	0.867	0.929	0.877
Department fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Semester fixed effects	Yes	Yes	Yes	Yes	Yes	Yes

Source: Authors using student-transcript-level data from Wellesley College.
Notes: To avoid giving undue influence to enrollments in smaller departments, we combined smaller departments by area creating the following 19 departments (and 20 majors): Art History, English, French, Spanish, Studio Art, Other Languages, Other Humanities, Economics, History, Philosophy, Political Science, Psychology, Religion, Other Social Sciences, Biological Chemistry (major only), Biological Sciences, Chemistry, Computer Science, Mathematics, and Other Sciences. We count the number of majors using what the student had listed as her major(s) in her final semester. In columns 5 and 6, the post-treatment period begins with the class of 2006 to account for lags in the policy's effect on the number of majors. Robust standard errors are in parentheses.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Here we use department- and program-level data to examine the number of enrollments and majors. Similar to the earlier results, we use a difference-in-differences methodology to examine whether those departments that were "treated" by the policy change (because they had grades above the 3.33 grade cap) had a differential change in enrollments and majors after the policy change. Table 3 presents results for the number of students enrolled in courses and the number of majors. Wellesley has some very small departments and programs, so we combined these enrollments into groups (see note to Table 3 for details), resulting in 342 semester-by-department observations on total enrollments. In columns 1 and 2, the outcome is enrollments in 100- and 200-level courses with at least 10 students (the courses subject to the grading policy), in either levels or logs. The results suggest a substantial decline in total enrollments for departments affected by the policy: total enrollments fell by about 51 students (levels) or by about 18.6 percent (logs) in treated departments.

As a placebo test of these results, columns 3 and 4 report whether enrollments in upper-level classes at the 300-level differed contemporaneously across treated and untreated departments. These enrollments should not have changed at this time because upper-level classes were not bound by the policy of the average grade being no higher than a B+. Since students typically take lower-level classes before moving on to upper-level classes, the policy should only affect upper-level enrollments with

a lag. Thus, if we find a statistically significant contemporaneous *PostPolicy*Treated* coefficient for department enrollments in upper-level classes, that would be an indication that the result for 100- and 200-level enrollments is merely picking up differential trends in enrollments across departments. However, as columns 3 and 4 show, the *PostPolicy*Treated* coefficients are small and statistically insignificant for upper-level enrollments.

In columns 5 and 6, we examine whether the number of majors was affected. There are only 200 observations in these columns because we count the number of majors only among second-semester seniors. Only those graduating in 2006 or later are counted as being in the post-policy era because earlier cohorts would not have had time to change major in response to the policy. The results suggest that majors declined in the treated departments by about eight students, on average, representing a relatively large decline of about 30 percent. Which departments gained and which departments lost majors? The fraction of a graduating class majoring in economics (and to a lesser extent in the sciences) increased, and the fraction of a graduating class majoring in other social sciences fell, with the fraction remaining flat in the humanities.[8]

Students' Evaluations of their Professors

Despite the fact that students' evaluations of their professors are a contentious measure of teaching effectiveness, they are a nearly universal feature of the US higher education landscape.[9] Some observers contend that this system of evaluating professors has contributed to grade inflation because students' evaluations of their professors set up an implicit quid pro quo with professors offering higher student grades in exchange for higher evaluations—evaluations that are used in tenure, promotion, and merit reviews (Zangenehzadeh 1988; Pressman 2007). It is the case at Wellesley that students in courses with higher average grades also tend to have higher evaluations of the quality of their professors' instruction, but this correlation cannot be taken as evidence that higher grades yield higher evaluations since higher average grades may indicate the teacher was effective, students learned

[8] For additional evidence of this, see Figure 1 in the online Appendix available with this paper at http://e-jep.org.

[9] Recent research that relies on random assignment to classes tends to support how difficult it is to assess teaching effectiveness. Braga, Paccagnella, and Pellizzari (2011) and Carrell and West (2010) use the fact that students are randomly assigned to classes, and develop measures of teachers' effectiveness based on students' performance in *subsequent* classes. Carrell and West (2010) find that students whose professors' teaching efforts yielded higher grades on a common exam tended to do less well in subsequent courses, suggesting that "teaching to the test" resulted in worse learning outcomes later. Braga et al. (2011) find that students who evaluated their professors more highly in their randomly assigned compulsory courses tended to have worse outcomes in subsequent courses. 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.

more, and both students' grades and their evaluations of the professor reflect this.[10] Here we examine what happened to student evaluations of their professors when there is an exogenous reduction in grades created by the policy.

At Wellesley, 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.[11] Thus, nearly 100 percent of students submit evaluations. Over the period studied, students gave their professors ratings on a four-point scale: "Strongly recommend"; "Recommend"; "Neutral"; and "Do not recommend." (There is also a qualitative component to the evaluation, but we only have access to the numeric component.) Students are generally well satisfied with their professors at Wellesley, with over 60 percent "strongly recommending" their professors (see online Appendix Table 3).

Table 4 shows the impact of the policy on student evaluations of their professors. The data are at the professor-by-course-by-semester level. Column 1 uses the average rating for the professor as dependent variable;[12] column 2 uses the percent of the students in the course that "strongly recommended" the professor as the outcome; outcomes are defined analogously for columns 3–5.

Students' evaluations of their professors fell after the policy in those departments that were affected by the grade cap. On average, students' ratings on the four-point scale fell by a statistically significant 0.11. The percent of students "strongly recommending" their professors fell by about 5 percentage points. There were statistically significant increases in the "neutral" and "do not recommend categories." In particular, the share of student evaluations in the "do not recommend" category rose from about 5 percent to slightly over 7 percent in the treated departments. In short, the results strongly indicate that students were less pleased with their instructors when the grading policy lowered average grades.

[10] In the online Appendix available with this paper at http://e-jep.org, online Appendix Table 5 presents more detailed findings on students' ratings of their professors at Wellesley in the pre-policy era. For example, holding various observable characteristics constant, visiting faculty receive lower student ratings; male and female faculty are rated without significant difference, professors with more experience at Wellesley receive higher teaching evaluations (which is not surprising since one does not get the opportunity to continue working at Wellesley if evaluations are poor); and students rate professors more highly in 200-level courses than in 100-level courses (again, unsurprising because generally speaking students have more choice over their 200-level courses). Students in general give higher scores to their professors in the humanities than they do in the sciences and the social sciences. Courses that have more withdrawals, and more students electing the credit/non option, have worse evaluations of the professors. Higher average course grade point average is associated with better professor evaluations.

[11] The privacy of these responses is taken very seriously, and a student's evaluation cannot be linked to the individual's academic record.

[12] In Wellesley's evaluation metric a "1" is the best score, but we have flipped the scoring system in column 1 because it is more intuitive that a negative coefficient indicates a lower rating.

Table 4
Impact of Policy on Student Evaluations

	Rating (1)	*% Strongly recommend* (2)	*% Recommend* (3)	*% Neutral* (4)	*% Do not recommend* (5)
*PostPolicy*Treated*	−0.111***	−0.050***	0.010	0.018***	0.022***
	(0.030)	(0.015)	(0.011)	(0.006)	(0.007)
Age (in years)	−0.001***	−0.005***	0.001	0.002***	0.002***
	(0.003)	(0.001)	(0.001)	(0.001)	(0.000)
Male Faculty	0.024	0.009	0.001	−0.008	−0.004
	(0.029)	(0.018)	(0.009)	(0.006)	(0.004)
Years Worked	0.009***	0.004**	−0.000	−0.001**	−0.002***
	(0.003)	(0.002)	(0.001)	(0.001)	(0.001)
Non-Tenure-Track	−0.011	0.004	−0.010	−0.003	0.009
	(0.058)	(0.032)	(0.019)	(0.016)	(0.009)
Visiting Faculty	−0.160***	−0.075***	0.016	0.034***	0.025***
	(0.046)	(0.025)	(0.012)	(0.010)	(0.007)
Tenured Faculty	−0.056	−0.020	−0.001	0.006	0.015**
	(0.033)	(0.018)	(0.012)	(0.009)	(0.006)
200 level	0.069***	0.032***	−0.009	−0.010*	−0.014***
	(0.021)	(0.012)	(0.007)	(0.005)	(0.003)
Class size/10	−0.001	−0.003	0.003*	0.000	−0.001
	(0.001)	(0.000)	(0.000)	(0.002)	(0.000)
Pretreatment mean	3.410	0.608	0.245	0.097	0.051
Observations	5,378	5,378	5,378	5,378	5,378
R^2	0.145	0.133	0.132	0.092	0.108
Department fixed effects	Yes	Yes	Yes	Yes	Yes
Semester fixed effects	Yes	Yes	Yes	Yes	Yes

Source: Authors using anonymized professor-course-semester-level data from Wellesley College.
Notes: "Rating" is calculated on a scale of 1 through 4 using 4 for "Strongly Recommend," 3 for "Recommend" etc. Pretreatment means are shown in the table for comparison. Controls for racial/ethnic characteristics of faculty were included but not shown in the table. Column 2 uses the percent of the students in the course that "strongly recommended" the professor as the outcome; outcomes are defined analogously for columns 3–5.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Discussion and Conclusion

Grades provide information. Grade inflation, and the resulting compression of grades near the top, reduces the information content of that signal, especially when the degree of grade inflation and compression varies across departments. This distortion in information may lead to a misallocation of resources on scales small and large. On the smaller side, Wellesley (like other institutions) has tended to allocate academic support services where bigger gaps in grades are evident, because those gaps are taken to reflect gaps in learning. When the policy lowered

grades in the humanities and non-economics social sciences, gaps in grades by group appeared to be more similar across departments. If the increased gaps mean that administrators now have better information about student learning in different departments, academic support services can be targeted more efficiently to improve learning outcomes.[13]

On a larger scale, grades provide information to students about where their talents lie. If students interpret lower grades as an indication that they are not good at a given subject, then they may shy away from low-grading disciplines to their detriment. The results here indicate students' choices about courses and majors are sensitive to grades. Research has shown that students base decisions to take a subsequent course in a discipline more on absolute grades than relative rankings (and relative skill) in the class (Sabot and Wakeman-Linn 1991). Goldin (2013, or see discussion by Rampell 2014) points out that women's choice of major may be especially sensitive to grades; she also points out that choices about academic major can have profound consequences for future earnings. Thus, the costs of the distortion in information created by grade inflation may be disproportionately born by some groups.[14]

One hope of those who advocate addressing grade inflation is that such policies might encourage more students to enter the science, technology, engineering, and mathematics fields. Although we find an effect of grades on major choice, the switching of majors appears to be more of a function of switching between types of social sciences than from humanities and social sciences to the sciences fields. It is worth noting that economics is currently the biggest major at Wellesley, with nearly 20 percent of students declaring it as (at least one of) their major(s). Of course, Wellesley's policy did not force average grades to be equal across departments, and the high-grading departments continue to have substantially higher grades; bigger changes in grades might induce bigger changes in choice of major.

Given the pressures on individual faculty members to have high student evaluations, and the pressures on departments to maintain enrollments, reducing grade inflation and compression requires action at the institutional level. Even so, devising policies is fraught with potential unintended consequences. Institutions have usually followed one of two types of policies: implementing grade targets as at Wellesley,[15] or trying to give students, graduate schools, and employers more information about the content of grades. In 1996, Cornell University adopted the latter type of policy by making public the median grades in each course. As Bar, Kadiyali, and Zussman

[13] If the new bigger gaps between groups within treated departments reflect an adverse reaction to increased pressure over grades—as the literature on stereotype threat might suggest—then the interpretation of the bigger gaps post-policy is quite different, and less sanguine.

[14] Although this comment suggests that the responses to the grading policy that we find at Wellesley, a women's college, may be larger than they would be in other settings, the broader point holds: students who are sensitive to grades in their choice of academic major will distort their choices if there is differential grade inflation and compression across disciplines.

[15] Princeton University also has a targeting policy, which is that no more than 35 percent of grades in a department should be an A.

(2009) discussed in this journal, the policy accelerated grade inflation at Cornell as a higher fraction of students chose to take more leniently graded courses.

Any institution that attempts to deal with grade inflation on its own must consider the possibility of adverse consequences of this unilateral disarmament. At Wellesley College, for example, prospective students, current students, and recent alums all worry that systematically lower grades may disadvantage them relative to students at other institutions when they present their grades to those outside the college. They point to examples of web-based job application systems that will not let them proceed if their GPA is below at 3.5. The economist's answer that firms relying on poor information to hire are likely to fare poorly and to be poor employers in the long run proves remarkably uncomforting to undergraduates. These concerns lead to pressure to reverse the grade policy. If grade inflation is a systemic problem leading to inefficient allocation of resources, then colleges and universities may wish to consider acting together in response.

■ *We appreciate the support and feedback from administrative staff, faculty, and students at Wellesley College. We are grateful to the administration and institutional research staff for making the data available. We thank participants in seminars at Wellesley College, Rutgers University, Brandeis University, and participants in the NBER Education meetings for helpful comments. We thank YeJi Kee and Ashley Longseth for helpful research assistance. The research received support from Wellesley College, in the form of access to anonymized data on students and faculty. Administrators of Wellesley College had the right to review the manuscript before publication.*

References

Achen, Alexandra C., and Paul N. Courant. 2009. "What Are Grades Made Of?" *Journal of Economic Perspectives* 23(3): 77–92.

Babcock, Philip. 2010. "Real Costs of Nominal Grade Inflation? New Evidence from Student Course Evaluations." *Economic Inquiry* 48(4): 983–96.

Bar, Talia, Vrinda Kadiyali, and Asaf Zussman. 2009. "Grade Information and Grade Inflation: The Cornell Experiment." *Journal of Economic Perspectives* 23(3): 93–108.

Braga, Michela, Marco Paccagnella, and Michele Pellizzari. 2011. "Evaluating Students' Evaluations of Professors." Bank of Italy Temi di Discussione (Working Paper) 825. Available at SSRN: http://ssrn.com/abstract=2004361.

Carrell, Scott E., and James E. West. 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." *Journal of Political Economy* 118(3): 409–32.

Chan, William, Li Hao, and Wing Suen. 2007. "A Signaling Theory of Grade Inflation." *International Economic Review* 48(3): 1065–90.

Franz, Wan-Ju Iris. 2010. "Grade Inflation Under the Threat of Students' Nuisance: Theory and Evidence." *Economics of Education Review* 29(3): 411–22.

Goldin, Claudia. 2013. "Can 'Yellen Effect'

Attract Young Women to Economics?" BloombergView, October 14. http://www.bloombergview.com/articles/2013-10-14/can-yellen-effect-attract-young-women-to-economics-.

Love, David A., and Matthew J. Kotchen. 2010. "Grades, Course Evaluations, and Academic Incentives." *Eastern Economic Journal* 36(2): 151–63.

Pressman, Steven. 2007. "The Economics of Grade Inflation." *Challenge* 50(5): 93–102.

Rampell, Catherine. 2014. "Women Should Embrace the B's in College to Make More Later." *Washington Post,* March 10. http://www.washingtonpost.com/opinions/catherine-rampell-women-should-embrace-the-bs-in-college-to-make-more-later/2014/03/10/1e15113a-a871-11e3-8d62-419db477a0e6_story.html?hpid=z3.

Rojstaczer, Stuart, and Christopher Healy. 2010. "Grading in American Colleges and Universities." *Teachers College Record,* March 4.

Sabot, Richard, and John Wakeman–Linn. 1991. "Grade Inflation and Course Choice." *Journal of Economic Perspectives* 5(1): 159–70.

Wongsurawat, Winai. 2009. "Does Grade Inflation Affect the Credibility of Grades? Evidence from US Law School Admissions." *Education Economics* 17(4): 523–34.

Zangenehzadeh, Hamid. 1988. "Grade Inflation: A Way Out." *Journal of Economic Education* 19(3): 217–26.

This article has been cited by:

1. Andrei Ternikov, Mikhail Blyakher. 2024. Grade inflation and grading process: does faculty workload matter?. *Journal of Applied Research in Higher Education* **23**. . [Crossref]
2. James Thomas. 2024. What Do Course Offerings Imply about University Preferences?. *Journal of Labor Economics* **42**:1, 53-83. [Crossref]
3. Wolfgang Stroebe. 2023. If Student Evaluations of Teaching Are Invalid, Why Are They Still Being Used? Comments on Uttl (2023). *Human Arenas* **44**. . [Crossref]
4. Makoto Shimoji. 2023. Setting an exam as an information design problem. *International Journal of Economic Theory* **19**:3, 559-579. [Crossref]
5. Eric Floyd, Sorabh Tomar, Daniel J. Lee. 2023. Making the Grade (But Not Disclosing It): How Withholding Grades Affects Student Behavior and Employment. *Management Science* **9**. . [Crossref]
6. Bryan Engelhardt, Marianne Johnson, Sarinda Siemers. 2023. Business school grades, assessment scores, and course withdrawals in the Covid-19 pandemic. *Journal of Education for Business* **98**:4, 199-215. [Crossref]
7. Christopher Lalley, Lauren McInally. 2023. Secondary school grades and graduate returns to education in the UK. *Journal of Education and Work* **36**:3, 169-185. [Crossref]
8. Jeffrey T. Denning, Eric R. Eide, Kevin J. Mumford, Richard W. Patterson, Merrill Warnick. 2022. Why Have College Completion Rates Increased?. *American Economic Journal: Applied Economics* **14**:3, 1-29. [Abstract] [View PDF article] [PDF with links]
9. Apostolos Filippas, John J. Horton, Joseph M. Golden. 2022. Reputation Inflation. *Marketing Science* **41**:4, 733-745. [Crossref]
10. Guannan Wang, Aimee Williamson. 2022. Course evaluation scores: valid measures for teaching effectiveness or rewards for lenient grading?. *Teaching in Higher Education* **27**:3, 297-318. [Crossref]
11. Stephen L. Baglione, Zachary Smith. 2022. Grade inflation: undergraduate students' perspective. *Quality Assurance in Education* **30**:2, 251-267. [Crossref]
12. Eric Floyd, Sorabh Tomar, Daniel Lee. 2022. Making the Grade (But Not Disclosing It): How Withholding Grades Affects Student Behavior and Employment. *SSRN Electronic Journal* **133**. . [Crossref]
13. Glen Waddell, Jenni Putz. 2022. What Can We Learn from Student Performance Measures? Identifying Treatment in the Presence of Curves and Letter Grades. *SSRN Electronic Journal* **30**. . [Crossref]
14. Amanda L. Griffith, Veronica Sovero. 2021. Under pressure: How faculty gender and contract uncertainty impact students' grades. *Economics of Education Review* **83**, 102126. [Crossref]
15. Martin Gregor. 2021. Electives Shopping, Grading Policies and Grading Competition. *Economica* **88**:350, 364-398. [Crossref]
16. Julie Berry Cullen, Cory Koedel, Eric Parsons. 2021. The Compositional Effect of Rigorous Teacher Evaluation on Workforce Quality. *Education Finance and Policy* **16**:1, 7-41. [Crossref]
17. Marianne Johnson, Bryan Engelhardt, Sarinda Taengnoi. 2021. Business School Grades, Assessment Scores, and Dropout Rates in the Covid-19 Pandemic. *SSRN Electronic Journal* **59**. . [Crossref]
18. Xuan Jiang, Kelly Chen, Zeynep Hansen, Scott Lowe. 2021. A Second Chance at Success? Effects of College Grade Forgiveness Policies on Student Outcomes. *SSRN Electronic Journal* **70**. . [Crossref]
19. Veronica Minaya. 2020. Do Differential Grading Standards Across Fields Matter for Major Choice? Evidence from a Policy Change in Florida. *Research in Higher Education* **61**:8, 943-965. [Crossref]

20. Dawn Apgar. 2020. The Fate of the Master's in Social Work (MSW) Degree: Will the Practice Doctorate Replace It as the Profession's Flagship Credential?. *Journal of Teaching in Social Work* **40**:5, 411-430. [Crossref]
21. Wolfgang Stroebe. 2020. Student Evaluations of Teaching Encourages Poor Teaching and Contributes to Grade Inflation: A Theoretical and Empirical Analysis. *Basic and Applied Social Psychology* **42**:4, 276-294. [Crossref]
22. Ian Burn, Michael E. Martell. 2020. The role of work values and characteristics in the human capital investment of gays and lesbians. *Education Economics* **28**:4, 351-369. [Crossref]
23. Shana K. Carpenter, Amber E. Witherby, Sarah K. Tauber. 2020. On Students' (Mis)judgments of Learning and Teaching Effectiveness. *Journal of Applied Research in Memory and Cognition* **9**:2, 137-151. [Crossref]
24. Martin Nordin, Gawain Heckley, Ulf Gerdtham. 2019. The impact of grade inflation on higher education enrolment and earnings. *Economics of Education Review* **73**, 101936. [Crossref]
25. Zombor Berezvai, Gergely Dániel Lukáts, Roland Molontay. 2019. A pénzügyi ösztönzők hatása az egyetemi oktatók osztályozási gyakorlatára. *Közgazdasági Szemle* **66**:7-8, 733-750. [Crossref]
26. William Walstad, William Bosshardt. 2019. Grades in Economics and Other Undergraduate Courses. *AEA Papers and Proceedings* **109**, 266-270. [Abstract] [View PDF article] [PDF with links]
27. Laura J. Ahlstrom, Carlos J. Asarta. Navigating the Economics Major: The Effect of Gender on Students' Degree Pathways 115-136. [Crossref]
28. Chenoa S. Woods, Toby Park, Shouping Hu, Tamara Betrand Jones. 2018. How High School Coursework Predicts Introductory College-Level Course Success. *Community College Review* **46**:2, 176-196. [Crossref]
29. Apostolos Filippas, John J. Horton, Joseph Golden. 2018. Reputation Inflation. *SSRN Electronic Journal* . [Crossref]
30. Martin Gregor. 2018. Electives Shopping, Grading Competition, and Grading Norms. *SSRN Electronic Journal* . [Crossref]
31. Diyi Li, Cory Koedel. 2017. Representation and Salary Gaps by Race-Ethnicity and Gender at Selective Public Universities. *Educational Researcher* **46**:7, 343-354. [Crossref]
32. Donghun Cho, Joonmo Cho. 2017. Does More Accurate Knowledge of Course Grade Impact Teaching Evaluation?. *Education Finance and Policy* **12**:2, 224-240. [Crossref]
33. Devon Gorry. 2017. The impact of grade ceilings on student grades and course evaluations: Evidence from a policy change. *Economics of Education Review* **56**, 133-140. [Crossref]
34. Amanda Bayer, Cecilia Elena Rouse. 2016. Diversity in the Economics Profession: A New Attack on an Old Problem. *Journal of Economic Perspectives* **30**:4, 221-242. [Abstract] [View PDF article] [PDF with links]
35. Wolfgang Stroebe. 2016. Why Good Teaching Evaluations May Reward Bad Teaching. *Perspectives on Psychological Science* **11**:6, 800-816. [Crossref]
36. William B. Walstad, Laurie A. Miller. 2016. What's in a grade? Grading policies and practices in principles of economics. *The Journal of Economic Education* **47**:4, 338-350. [Crossref]
37. Brandon Lehr. 2016. Information and Inflation: An Analysis of Grading Behavior. *The B.E. Journal of Economic Analysis & Policy* **16**:2, 755-783. [Crossref]
38. J.G. Altonji, P. Arcidiacono, A. Maurel. The Analysis of Field Choice in College and Graduate School 305-396. [Crossref]

39. Sam Allgood, William B. Walstad, John J. Siegfried. 2015. Research on Teaching Economics to Undergraduates. *Journal of Economic Literature* **53**:2, 285-325. [Abstract] [View PDF article] [PDF with links]

40. Tatyana Deryugina, Olga Shurchkov. 2015. DOES BEAUTY MATTER IN UNDERGRADUATE EDUCATION?. *Economic Inquiry* **53**:2, 940-961. [Crossref]