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

Working Paper

Making the (letter) grade: The incentive effects of mandatory pass/fail courses

Working Paper, No. WP 2022-55

Provided in Cooperation with:
Federal Reserve Bank of Chicago

Suggested Citation: Butcher, Kristin F.; McEwan, Patrick J.; Weerapana, Akila (2022) : Making the (letter) grade: The incentive effects of mandatory pass/fail courses, Working Paper, No. WP 2022-55, Federal Reserve Bank of Chicago, Chicago, IL, https://doi.org/10.21033/wp-2022-55

This Version is available at:
https://hdl.handle.net/10419/272815

Making the (Letter) Grade:

The Incentive Effects of Mandatory Pass/Fail Courses

Kristin Butcher, Patrick J. McEwan, and Akila Weerapana

November 2022

WP 2022-55

https://doi.org/10.21033/wp-2022-55

FEDERAL RESERVE BANK of CHICAGO

Making the (letter) grade: The incentive effects of mandatory pass/fail courses*

Kristin Butcher
Federal Reserve Bank of Chicago, Wellesley College, and NBER

Patrick J. McEwan
Wellesley College

Akila Weerapana
Wellesley College

November 2022

JEL Codes: I23
Keywords: Higher Education, Grading Policy

Abstract: In Fall 2014, Wellesley College began mandating pass/fail grading for courses taken by first-year, first-semester students, although instructors continued to record letter grades. We identify the causal effect of the policy on course choice and performance, using a regression-discontinuity-in-time design. Students shifted to lower-grading STEM courses in the first semester, but did not increase their engagement with STEM in later semesters. Letter grades of first-semester students declined by 0.13 grade points, or 23% of a standard deviation. We evaluate causal channels of the grade effect—including sorting into lower-grading STEM courses and declining instructional quality—and conclude that the effect is consistent with declining student effort.

* We are grateful to many administrators and staff of Wellesley College for facilitating this research; to our Economics colleagues and students for their ideas; and to Josh Hyman and AEFP conference attendees for their comments. The authors did not receive grant support for the research, although they are employees of Wellesley College. The views here represent those of the authors and do not necessarily reflect those of the Federal Reserve Bank of Chicago, the Federal Reserve System, or any other entity.

1. Introduction

College grades have risen since the 1960s, especially in the humanities and non-Economics social sciences (Sabot and Wakeman-Linn, 1991; Rojstaczer and Healy, 2010; Butcher, McEwan, and Weerapana, 2014). The literature has expressed two concerns about letter-grade inflation and compression.[1] First, students are responsive to letter grades in their choices of courses and fields, perhaps in ways that nudge students away from lower-grading but higher-labor-market-return fields in STEM and Economics (Ahn et al., 2019).[2] Indeed, a literature shows that students' course and major choices are sensitive to exogenous variation in letter grades.[3] At Wellesley College, for example, an anti-grade-inflation policy in 2004 capped mean grades at a B+, leading to sharp declines in grades in the humanities and non-Economics social sciences, but not in lower-grading fields like STEM and Economics (Butcher et al., 2014). Course enrollments and major elections declined in capped departments, relative to already-complying ones.

Second, grades are performance incentives that may induce higher effort and learning among extrinsically motivated students (Becker and Rosen, 1992; Betts, 1998). The incentive is plausibly dampened in courses with a compressed grade distribution, since there is a lower risk

[1] In addition to the concerns mentioned in the text, a recent paper shows that rising college completion rates in the United States can be largely attributed to grade inflation, rather than changing student attributes or other explanations (Denning et al., 2022). A later section explores this hypothesis in the context of our data.

[2] On the one hand, students sort into courses in which they *anticipate* a higher grade for a given effort (Bar, Kadiyali, and Zussman, 2009). On the other hand, students' choices *respond* to higher or lower letter grades, perhaps because students use grades to update their beliefs about field-specific abilities (Stinebrickner and Stinebrickner, 2014).

[3] Owen (2019) analyzed a university policy that awarded higher grades in Economics for the same underlying course performance, finding that it increased course and major choices (but not in the subsample of women). In addition, several regression-discontinuity studies use an alternate source of exogenous grade variation, by comparing the course and major choices of students in Economics courses with course averages falling just above or below letter-grade cutoffs. McEwan, Rogers, and Weerapana (2021) find that Wellesley College students with higher letter grades are substantially more likely to major in Economics. The results are mixed in two earlier discontinuity studies with smaller samples (Owen, 2010; Main and Ost, 2014). Despite the lack of statistical significance of some coefficients, confidence intervals include large effects on major choice.

of receiving a lower grade. Influential research in K-12 schools showed that teachers with higher grading standards improved students' test scores (Betts and Grogger, 2003; Figlio and Lucas, 2004).[4] In higher education, inflated grades are associated with lower study effort (Babcock, 2010; Babcock and Marks, 2011), and graded assignments in Economics courses tend to increase effort and course performance relative to ungraded work (Grove and Wasserman, 2006; Grodner and Rupp, 2013; Artés and Rahona, 2013; Bonesrønning and Opstad, 2015). But, to our knowledge, there is no causal evidence that grading standards in higher education systematically affect student effort and/or learning.[5]

This paper focuses on identifying the effects of a related and potentially important source of grade compression: the use of pass/fail grading instead of letter grades. In the typical policy, students with very low grades receive a "fail" or "unsatisfactory" on their official transcripts, while other letter grades are awarded a "pass" or "satisfactory." Among 20 highly ranked liberal arts colleges, for example, nearly all allow students to voluntarily designate at least three or four courses for pass/fail grading which are exempted from official grade point averages.[6] All the colleges implemented some form of pass/fail grading during the pandemic semester of Spring 2020.

[4] Also see Bonesrønning (2004) and Gershenson (2020). More recently, Hvidman and Sievertsen (2021) analyzed a Danish high school reform that arbitrarily recoded grade point averages. Students who received exogenously lower grade signals—despite similar prior performance—responded by increasing their performance on later assessments. A Swedish experiment among sixth-graders found no effects of letter-grading on performance relative to a control group (Jalava, Joensen, and Pellas, 2015). However, the grades awarded in the experimental task were not high stakes, unlike the Danish setting in which grades and performance strongly affected higher education access.

[5] Theory and evidence in Oettinger (2002) show that students respond to incentives of a non-linear (or cutoff-based) grading system by strategically adjusting effort. The paper is consistent with the idea that fewer or lower cutoffs may induce further strategic responses that could further lower effort and performance.

[6] Middlebury College allows only two. Many colleges allow more frequent elections, or otherwise creatively encourage the use of pass/fail grading among students and instructors. At Amherst, students can declare the pass/fail option after receiving the final letter grade. At Haverford and Swarthmore, students can rescind the option after receiving the final letter grade. At Vassar, students can declare a letter-grade threshold, above which a letter grade is reported and below which the course converts to pass/fail reporting. At Wesleyan, instructors have discretion in declaring a course's grading scheme to be pass/fail, rather than letter-graded.

There is little empirical evidence on the phenomenon, presumably because of barriers to causal identification. The first is related to missing data, since instructors interval-censor letter grades and may not officially record a letter grade when issuing a pass/fail grade. The second is related to selection bias. In addition to reduced effort, lower grades in pass/fail courses relative to letter-graded ones could indicate that lower-ability students are more inclined to use the pass/fail option, or that students voluntarily elect the pass/fail option in courses with harsher grading standards, so to avoid a negative grade signal on a transcript.

Wellesley College provides a novel opportunity to identify the causal effect of pass/fail grading on course choices and performance. First, instructors are required to record letter grades for nearly all students, to monitor compliance with an anti-grade-inflation policy that caps mean course performance among all enrolled students at a B+ (Butcher et al., 2014).[7] Second, the College began mandating pass/fail grading for first-year, first-semester students in Fall 2014. Under the shadow grading policy, students are privately informed of letter grades, but the official transcript reports a pass for letter grades from A to D, and a fail otherwise.[8] The policy was based on related policies for first-year students at Swarthmore College, MIT, and Johns Hopkins.[9]

The implementation of the policy allows for a regression-discontinuity-in-time design, akin to an interrupted time series design (Hausman and Rapson, 2018; Shadish, Cook, and Campbell, 2002). In this setting, academic year is the running variable and Fall 2014 is the policy cutoff, after which first-semester courses are mandatory pass/fail. We use local linear regression

[7] This feature of the policy was designed in explicit recognition of the fact that pass/fail elections are more common among students with lower anticipated grades.
[8] As section 2 describes, there is also a voluntary pass/fail option, available before and after Fall 2014, in which a passing grade is a C or higher.
[9] Before 2002, MIT required pass/fail grading for Fall and Spring of the first year. Using observational data, Harris (2010) reports estimates of the subsequent shift to letter-grading in Spring. Like this paper, he found that the shift is associated with higher Spring grades.

specifications, complemented by visual identification, to test for policy effects in the vicinity of the cutoff. The key threat to internal validity is that College applicants and/or students non-randomly sort into or out of the sample, either in anticipation of the pass/fail policy or in response to it. We test for sorting bias in several ways, as described in a later section.

Our preferred estimates show that first-semester students exposed to the policy were 82 percentage points more likely to take a course as pass/fail than earlier groups. Students in some later semesters were slightly less likely to use a voluntary pass/fail option, providing a tacit recognition that "too many" pass/fail courses on official transcripts provides a negative signal to graduate schools or employers. The policy influenced course choices in the first semester, increasing the probability of taking a STEM course by 5 percentage points over a pre-policy mean of 25%. One interpretation is that student choices responded to potential grades. Consistent with this, departments with lower pre-policy grades had larger increases in enrollment. Another is that students' first-year advisors nudged them to take courses outside their usual preferences. Regardless, early engagement with introductory STEM classes might spur later engagement via course and major choices. However, we do not find evidence that the policy increased the cumulative number of STEM courses or the probability of graduating with a STEM degree. Rather, the policy's main effect on student choices was to shift introductory STEM courses earlier in college careers.

There are more consequential results for student grades. The policy lowered the average grade points of first-semester students by 0.13 or about 23% of a standard deviation in pre-policy grades, although effects on the cumulative grade point average were small and not statistically distinguishable from zero. There are several possible explanations for the grade effects on first-year students. The first is compositional: students sorted into lower-grading STEM courses

which mechanically lowered average grades of first-semester students. However, we show that compositional changes explain a reduction of less than 0.01, implying a substantial role for within-course effects.

We next consider the possibility that grade reductions are due to lower quality instruction. For example, some STEM courses experienced larger class sizes or declining peer ability, due to an influx of first-semester students. However, the estimates are not sensitive to the inclusion of controls for quality proxies. We further show that the policy did not affect the grade performance of all students in a course, as might be expected if quality uniformly declined. Rather, it lowered the grades of first-semester students relative to later-semester students enrolled in the same courses.

Finally, we rule out the possibility that the policy led instructors to arbitrarily modify their grading standards for first-semester students relative to later-semester students. For example, instructors might have sought an easy route to comply with the anti-grade-inflation policy by lowering unofficial letter grades awarded to first-semester students. This is unlikely, if only because instructors—even in relatively higher-grading departments—were already complying with the policy in the semesters leading up to 2014. This suggests few benefits to justify lowering the grades of first-semester students, especially when anonymous student course evaluations influence tenure, promotion, and compensation.

The remaining and most plausible explanation for the policy-induced reduction in grades is that students covered by mandatory pass/fail grading exerted less effort relative to letter-graded students. We do not have direct measures of student effort. However, we show that the policy had no effect on students' anonymous course evaluations. In prior research we found that an exogenous grade reduction—in which student effort likely remained the same—led students to

"punish" faculty with lower course evaluations (Butcher et al., 2014). One interpretation is that the mandatory pass/fail policy reduced *both* effort and grades. This interpretation is bolstered by descriptive evidence from a faculty survey in 2017, which provided detailed examples of how students reduced effort, including attendance and course preparation.

Grades are generally understood by economists and students as performance incentives that influence effort and the acquisition of human capital (Becker and Rosen, 1992; Betts, 1998). Despite an early literature on K-12 grading standards (Betts and Grogger, 2003; Figlio and Lucas, 2004), there is little work on higher education. Our paper provides important evidence on whether college-wide grading standards have causal effects on student learning in higher education. It is also the first to report causal effects of pass/fail grading on student outcomes. Pass/fail grading is widely used in American higher education—especially during the pandemic semester of Spring 2020—but has been subjected to little empirical study.

2. Pass/Fail Grading at Wellesley College

Wellesley College students can take an unlimited number of courses as voluntary credit/credit-non, described in this paper using the generic term of pass/fail. Pass appears on students' transcripts if they receive a letter grade of C or above, while fail appears for lower grades. Students are required to declare this grading option before the end of the 4th week of the semester. Between Fall 2004 and Spring 2013, 9% of Wellesley course grades were assigned with this grading option (see Table 1).

Beginning in Fall 2014, the College further implemented a shadow grading policy for first-year, first-semester students. Under this policy, transcripts record a pass if students receive a letter grade of D or above, and a fail if they receive an F. However, students are privately

notified of the letter grade. The policy has two objectives. The first is to encourage students to take courses outside their usual preferences. Curricular exploration might foster—in the short- and longer-run—increased student engagement with fields in which they are under-represented (such as women in mathematically-intensive STEM majors). The second is to promote successful transitions from high school to college, thereby preventing leaves of absence and drop out.

Figure 1 (panel A) describes the use of pass/fail grading over time.[10] The cohort of students who enrolled in Fall 2004 took just under 10% of its Wellesley courses with the voluntary pass/fail option. The Fall 2014 cohort—the first one exposed to shadow grading—took more than 25% of its courses with the pass/fail option. The Fall 2016 cohort experienced another sharp increase, given mandatory pass/fail grading during the pandemic semester of Spring 2020. Across all cohorts, students took just over 25 courses at Wellesley College (with the balance of 32 degree-counting courses transferred from elsewhere, including cross-registration and study-abroad programs).

Panel B further illustrates changes in grades during the same period. The solid line describes mean grades *only in letter-graded courses* (where an A is 4.0, an A- is 3.67, and so on). The dashed line includes these grades, in addition to letter grades assigned in pass/fail courses (i.e., the larger sample used to monitor compliance with the anti-grade-inflation policy). It has a slight upward trend, but does not substantially exceed the policy cap of B+ (3.33).[11] The dashed line is consistently below the solid line, indicating that letter grades in pass/fail courses are lower.

[10] The sample in the figure—and the entire paper—is limited to cohorts that were exposed to the college-wide anti-grade-inflation policy, implemented in Fall 2004 (Butcher et al., 2014).

[11] The policy applied only to courses at the 100- and 200-level (but not 300-level courses taken by majors) enrolling more than 10 students. Thus, an apparent weakening of compliance can also indicate a shift in the distribution of enrollments across small or advanced courses.

However, the identification of causal effects is complicated by selection, since lower-ability students may be more inclined to use the pass/fail option. We shall compare the grades of cohorts just after and before the implementation of mandatory pass/fail for first-semester students in Fall 2014. Indeed, there is a small dip in mean grades for the cohort entering in Fall 2014, although it pools grades from all semesters for a given entering cohort. The next section describes the empirical strategy for isolating effects on first-year, first-semester students.

3. Effects on First-Semester Students

A. Estimation and Results

Let $S_{ic} = \{1, \ldots, 8\}$ denote the semester in which student *i* took course *c* at Wellesley College. The numbering of semesters is independent of students' choices to take leaves of absence or otherwise not enroll, such that $S_{ic} = 8$ always indicates spring courses taken in the eighth consecutive semester after Fall enrollment in $S_{ic} = 1$. Let $T_{ic} = \{4, \ldots, 19\}$ denote the academic year of the course, between 2004-2005 and 2019-2020. Beginning in Fall 2014, all first-year, first-semester courses were taken pass/fail, although students received advisory letter grades that were not reported on official transcripts.

We initially limit the sample to student-by-course observations in the first semester ($S_{ic} = 1$). The goal is to identify whether the policy caused immediate changes in: (1) whether courses used pass/fail grading; (2) whether courses were taken in a STEM department; and (3) the letter grade received in a course, recalling that grades are recorded even when pass/fail grading is used. We convert letters to grade points using the official scale. An A is 4.0, an A- is 3.67, a B+ is 3.33, a B is 3.0, a B- is 2.67, a C+ is 2.33, a C is 2.0, a C- is 1.67, a D is 1, and an F is 0.

To illustrate the empirical approach, consider the upper-left panel of Figure 2. The dots are unsmoothed proportions of courses taken pass/fail in each academic year T_{ic}, in the sample of 38,214 student-by-course observations for which $S_i = 1$. The proportion increases sharply in Fall 2014. (It was non-zero in earlier years because students could use the voluntary pass/fail option and because a mandatory writing course was required to be taken as pass/fail.)

The solid lines are fitted values obtained with a linear spline regression:

$$(1) \qquad O_{ic} = \alpha_1 + \beta_1(T_{ic} - 14) + \gamma_1 1\{T_{ic} \geq 14\} + \delta_1(T_{ic} - 14) \times 1\{T_{ic} \geq 14\} + \varepsilon_{ic},$$

where O_{ic}, a binary indicator of pass/fail grading, is regressed on a function of T_{ic}. The slope of T_{ic} is allowed to vary on either side of the cutoff at 14, while γ_1 measures the intercept shift at the cutoff. Table 2 reports $\hat{\gamma}_1$ and its standard error for each dependent variable. Given the likelihood of correlated errors, we adjust standard errors for multi-way clustering within 2,060 courses and 9,544 non-nested students.[12] When using asterisks to report statistical significance for multiple estimates in a table, we control the false discovery rates using a step-up procedure (Benjamini and Hochberg, 1995).

Figure 2 and Table 2 show that the policy dramatically increased the probability—by 82 percentage points—that first-semester students took a course with the pass/fail rather than a letter-graded option. The estimate is statistically different from zero and not sensitive to the inclusion of a rich set of pre-college covariates. (We shall later confirm that pre-college covariates do not vary sharply at the cutoff.)

[12] Note that we do not adjust for clustering within discrete values of T_{ic}, once a common approach to adjust for potential mis-specification of the functional form for the assignment variable in regression-discontinuity designs. Kolesár and Rothe (2018) show these standard errors have poor coverage properties. Indeed, when we use multi-way clustering by students and discrete values of T_{ic} (which nest unique courses), the standard errors for the grade variable models are more than 30% smaller than those in Table 2. As advised in Kolesár and Rothe (2018), we cluster standard errors only on variables suggested by the structure of the dataset (but not the discrete assignment variable), and also report models with smaller bandwidths to assess potential bias.

The upper-right panel of Figure 2 illustrates that the proportion of STEM courses taken by first-semester students also increased sharply in response to the policy. In Table 2, the estimate is about 5 percentage points, regardless of controls for pre-college covariates, and statistically distinguishable from zero. One interpretation is that students' choices reflect an aversion to grade-related risk in lower-grading STEM departments, and thus a desire to "cover" grades with pass/fail grading. To explore this, we repeatedly estimated equation 1, always including pre-college covariates, with 59 dummy dependent variables indicating the department of each course. In the lower-left panel of Figure 2, we plot the $\hat{\gamma}_1$ for each department against its pre-2014 mean grade.[13] The policy increased election of courses in both mathematics and computer science by over one percentage point, and both departments had mean grades well below the policy cap. All STEM-related departments fell into the upper-left quadrant indicating lower mean grades and positive effects on course election.[14] The fitted line indicates the negative relationship between pre-policy grades and the policy's effect on course election.

Another interpretation of this evidence is that first-year students were encouraged by their first-year advisors to choose courses outside their usual preferences, and that these happened to be in lower-grading STEM courses.[15] This, of course, was an explicit goal of the policy, as described in section 2, and we cannot unambiguously identify students' motives from the descriptive evidence.

[13] Specifically, we estimate the mean grade of a department in the sample of courses that were subject to the anti-grade-inflation policy: 100- and 200-level courses enrolling 10 or more students. First-year students are generally not eligible to take advanced (or 300-level) courses.

[14] We did not include the Economics Department in the STEM classification, though it offers mathematically-intensive courses with mean grades well below the cap. In the lower-left panel of Figure 2, Economics is the lowest grading non-STEM department, though the policy did not markedly increase the selection of Economics enrollments. The previous results are not sensitive to the classification of Economics as STEM.

[15] Each first-year student is assigned to a faculty advisor who consults with students prior to registration. Faculty advisors were informed of the policy and its objectives, although we have no data on their specific interactions with students.

Finally, Figure 2 and Table 2 report estimates for course grade points, recalling that instructors record letter grades even for pass/fail courses. Among first-semester students, the policy reduced mean grade points by 0.14 points, focusing on the specification that includes pre-college covariates. The same effect is evident in the unsmoothed means in Figure 2. As a caveat, the sample excludes 7% of observations for which pass/fail outcomes—but not letter grades—are available. Of these, 40% reflect a failure to complete course requirements (denoted "incomplete" on transcripts), or withdrawal from the course after the official deadline (indicated by "withdraw" on transcripts). For the remaining 60%, instructors simply neglected to report a letter grade. In these cases, grades are interval-censored rather than missing, and fall into either a lower (fail) or higher (pass) interval.[16]

We assess robustness to sample selection in the final column of Table 2, which reports estimates from a linear interval regression that treats grades as either exactly observed or interval-censored. The first-semester estimate (0.13) is similar in its magnitude and statistical significance to the prior estimate. We conclude that interval-censoring does not substantially influence the main finding: that first-semester grades decline sharply as a result of policy exposure. The effect represents a decline of 23% of a standard deviation in the pre-policy grade distribution (see Table 1). The findings in Table 2 are also robust to using a smaller bandwidth of 6 years. In Table A3, we exclude the academic years before 2008-2009, such that both linear splines are estimated with six academic years of data. The main coefficients are similar to those of Table 2.

[16] The voluntary pass/fail option—referred to as "credit/no-credit"—categorizes grades as C or above as "credit" and C- or below as "no credit" (see https://www.wellesley.edu/registrar/registration/creditnon). We assume that incomplete or withdrawal grades fall into the lower interval, since they are consistent with a student's failure to meet basic course requirements after substantial engagement with a course.

B. Threats to Internal Validity

The causal interpretation of the estimates in Table 2 hinges on whether there is balance near the cutoffs in observed and unobserved variables that affect O_i. Because this is a regression-discontinuity-in-time design, a key threat to internal validity is that applicants and/or students responded to the impending or realized policy by non-randomly sorting into or out of the sample (Hausman and Rapson, 2018). In this setting, there are two types of sorting: (1) non-random selection in students' initial enrollment decisions (whether before or after the policy), and (2) non-random attrition from the estimation sample after the initial enrollment decision.

The policy was approved by the Academic Council of Wellesley College in May 2013 for implementation in Fall 2014, although detailed knowledge of the policy was limited to a faculty committee tasked with implementation. In Spring 2014, information about the new policy was disseminated via the admissions website, although it is not clear whether this influenced students' decisions. As a simple test, Figure A1 plots the yield rate—the percentage of admitted students who enroll—of fall admission cohorts between 2005 and 2019. Despite an increasing trend and some noise, there is no evidence of a discontinuity in a practical or statistical sense, using a specification like equation 1 (see the note to Figure A1).

Non-random attrition from the grade data is possible if students: (1) take a leave of absence, (2) choose to study elsewhere for a semester or year (and therefore contribute no grade data in a semester), and/or (3) permanently transfer or drop out. There is zero attrition among initially-enrolled students in the first-semester sample ($S_{ic} = 1$), which eliminates concerns about the main results in Figure 2 and Table 2. In a later section, we assess whether the policy affected student attrition in later semesters.

Lastly, we can assess the smoothness of pre-college covariates near the cutoffs. Figure 3 plots unsmoothed means of covariates from each academic year, as well as fitted values obtained from estimates of equation (1) in the first-semester sample (using four pre-college covariates as dependent variables). Despite an increasing trend in the percentage of students with a pre-college interest in STEM majors, there is no discontinuity at the policy cutoff. Likewise, there are no discontinuities for a quantitative skills assessment given to entering cohorts before the start of first-semester classes, or for math SAT and composite ACT scores.[17]

Table A2 reports estimates from the pooled regression specification for each of 16 pre-college covariates, including the four in Figure 3. The other covariates include verbal and writing SAT scores; a global quality rating of each application by the College's admissions committee (vote total); seven dummy variables indicating race and ethnicity; and dummy variables indicating first-generation college student status and the receipt of any financial aid. There is no evidence of practical or statistical differences at the cutoffs, consistent with evidence in Table 2 that our main results are not sensitive to controls for pre-college variables.

C. Heterogeneity

Table 2 also examines how the policy affected the distribution of grades. It reports estimates of equation (1) for eight dummy variables indicating seven letter grades, as well as a combined category of C- or lower. Among first-semester courses, the probability of receiving an A or A- declines by 0.03 and 0.05 percentage points, respectively. There are no changes in the probability of receiving a B+ or B, and increased probabilities of receiving lower grades (including a C- or

[17] In contrast to the SAT and ACT, the quantitative skills assessment is available for all students in a given cohort and so it is a particularly credible test of whether pre-college ability is continuous at the cutoff. The assessment is missing in the 2018-2019 academic year and later due to a substantial modification and re-scaling of the test.

lower). One interpretation is that the policy caused modestly lower grades among many students, thus shifting students across the entire grade distribution. Another is that the policy caused especially large declines among fewer students of predominantly high ability, as A students became B- and below students.

Available theory does not predict which is the more likely response. Higher-ability students may respond to higher grading standards by increasing costly effort, while lower-ability students may fail to exert additional effort as the standard becomes unattainable (Becker and Rosen, 1992; Betts, 1998). In this view, a weaker grading standard might have produced large grade reductions concentrated among high-ability students. A contrasting view is that high-ability students might have greater intrinsic motivation than other students, perhaps leading them to exert higher effort regardless of the grading standards (and therefore less sensitive to extrinsic grading incentives).

We can further assess this by estimating effects within subsamples defined by the median of pre-college math skills.[18] Within each category of pre-college math skills, we further divide students by financial aid eligibility, a proxy for low household income. The results in Table 3 suggest three main findings. First, all four groups are substantially exposed to a greater proportion of pass/fail courses, as expected. Second, students with below-median math skills—regardless of financial aid status—are modestly more likely to choose STEM courses in the first semester. This is consistent with a greater perceived grade risk of taking STEM courses among students with weaker math skills. Third, the reduction in mean grades is also greater among students with below-median math skills, but mainly among higher-income students without

[18] We categorize students into "below-median" and "above-median" math skills using the quantitative reasoning assessment applied during first-year orientation (see Table 1). In the three most recent semesters—when the assessment is not comparable with prior years—we categorize students using mathematics SAT scores.

financial aid. Thus, there is little support for the notion that effects are primarily concentrated among high-achieving or high-ability students, in contrast to some predictions (Becker and Rosen, 1992; Betts, 1998). Nonetheless, we still find an appreciable effect of 0.12 among students with above-median math skills.

4. Causal Channels of First-Semester Grade Effects

This section explores evidence of four channels for the causal impact of the policy on average grades. First, the policy increased the proportion of students enrolled in lower-grading STEM courses by 5 percentage points, suggesting that sorting from higher-grading to lower-grading courses could partly explain the grade effect. In the sample of courses taken by first-semester students, the mean grade in non-STEM courses was 3.32 in the pre-policy period, compared with 3.20 in STEM courses. Given these grades, the implied change in average grades due to sorting alone is -0.006.[19] Thus, the larger decline of 0.13 in mean first-semester grades must be explained by changes within courses.[20]

Second, it is possible that the policy affected the quality of instruction, via its effects on the class size or the peer composition of a course. Because the policy affected the course choices of first-year students, some courses experienced larger classes or a higher percentage of first-semester students (since many courses also enroll older students). Larger classes could have reduced instructor availability in office hours, for example, while younger students could have

[19] In the pre-policy sample of first-semester grades, the mean grade in STEM and non-STEM courses is 3.20 and 3.32, respectively, and these categories account for 26.1% and 73.9% of courses. The weighted average is 3.28, which declines by 0.006 if the STEM proportion declines by 5 percentage points.

[20] In a related exercise, we estimated the pooled specification within subsamples defined by the STEM and non-STEM categories of courses. With a full set of pre-college covariates, $\hat{\gamma}_1 = -0.09$ in the non-STEM sample and $\hat{\gamma}_1 = -0.20$ in the STEM sample, with standard errors of 0.01 and 0.04, respectively. This provides additional evidence that full-sample estimates are driven by within-course changes in student or instructor behavior, with the caveat that selection into the STEM and non-STEM samples is endogenous to the policy.

affected the level of classroom instruction or peer interactions. Of course, some courses may have experienced the opposite effects since first-semester students also sorted out of some courses. As a straightforward test, we included controls for course enrollment and the percentage of first-year students in a course, in addition to pre-college covariates. The estimates and standard errors are very similar for grade points: including these controls changed the estimate from -0.129 (0.024) to -0.127 (0.0.23).

A related implication of the instructional quality hypothesis is that *all* students' grades—rather than just first-semester students—should have been affected by lower quality instruction. To assess this, we can estimate the policy effect on relative grades of first-semester and later-semester students within courses. In Figure 4, the sample includes student-by-course observations for first-semester students. The upper-left panel indicates the proportion of first-semester students enrolled in a course in which at least one later-semester student (i.e., students for whom $S_{ic} \neq 1$) received a grade. This has remained relatively steady across academic years, and the discontinuity in 2014 is small and not statistically different from zero. That is, the great majority of first-semester students are enrolled in courses with at least one later-semester student, both before and after the policy.

We next calculated the average grade of later-semester students in each course, and imputed this course-level mean for each first-semester student's observation in the same course. The upper-right panel suggests that the policy did not directly affect mean grades of later-semester students in the same courses as first-semester students (the point estimate and standard error are reported in the figure note). This is not unexpected, since later-semester students were not subject to mandatory pass/fail grades. Finally, we calculated the ratio of each first-semester student's grade points to the average grade points of later-semester students in the same course.

The lower-left panel shows that the ratio declined from 1.03 to 0.96 at the policy cutoff. The estimate of -0.07 is statistically different from zero at conventional levels. That is, grades of first-semester students exceeded course averages of later-semester students before the policy, but fell below the average after the policy. This is suggestive that the policy was not driven by a uniform reduction in the quality of instruction.

Third, it is possible that instructors used a harsher grading standard for first-semester students, even if instructional quality did not change. This is a plausible means of complying with the anti-grade-inflation policy. However, the mean grades in first-semester courses suggest that instructors—even in non-STEM courses—were complying with mean grade caps in the pre-policy period. Thus, there were no obvious benefits to applying stricter grading standards to first-semester students. Yet there is a potential cost, since students complete anonymous course evaluations that affect faculty tenure, promotion, and salary outcomes. In earlier work, we showed that an exogenous reduction in grades due to the anti-grade-inflation policy—enacted in Fall 2004—caused average course evaluations to decline (Butcher et al., 2014).

Fourth, student effort under pass/fail grading may have declined. Effort is costly to students, but is justified by additional learning or, at least, a credibly higher signal of learning on one's transcript. This, in turn, potentially impacts graduate school admissions and employment opportunities. The pass/fail option reduces the payoff to effort for extrinsically-motivated students, since the letter-grade signal no longer appears on transcripts.

We do not have quantitative measures of student effort that could be used to directly assess this causal channel. As indirect evidence, Figure A2 analyzes student course evaluations in the sample of 100-level courses offered in the fall semester of each academic year (i.e., the

introductory courses usually attended by first-semester students).[21] The figure shows that average evaluations in these courses were not affected by the policy, despite sharply lower grades and despite prior evidence that exogenous declines in grades can lower course evaluations (Butcher et al., 2014). A plausible interpretation is that students only punish faculty with lower course evaluations when student effort is held constant. The anti-grade-inflation policy exogenously reduced transcript grades for some students, even as they exerted similar effort levels. In contrast, the pass/fail policy plausibly induced lower effort, which in turn reduced letter grades that were available only to students.

As indirect evidence of student effort, we summarize qualitative data from a faculty survey conducted by the Provost's Office in 2017. Of 110 faculty members providing anonymous, written feedback, 48% have a generally negative opinion of the policy, 22% have mixed opinions, 17% are neutral, and 13% are generally positive. There are two common themes in the negative feedback, both consistent with our empirical results. First, instructors who also serve as academic advisors believed that first-year students were more inclined to game course selection, shifting required STEM courses into the first semester. Second, instructors commented that student motivation and effort were lower in first-semester sections of their introductory courses, relative to similar settings in pre-policy years and relative to non-first-semester students. As examples, instructors mention (1) course attendance; (2) the level of preparation for class discussions; (3) the quantity and quality of students' note-taking; (4) commitment to tasks such as peer review of writing assignments; (5) the take-up of options to re-write papers for a higher grade; and (6) exam preparation and performance. Some instructors commented specifically that

[21] The course evaluation results are only reported to instructors and administrators as course-specific averages, and so we cannot attach course evaluations to course-by-student observations in the estimation sample.

lower effort was more pronounced in the second half of the semester, possibly when students felt greater assurance of reaching a passing threshold.

5. Policy Effects After the First Semester

A. Effects in Semesters Two Through Seven

We next assess whether policy continued to affect student choices and performance even after the first semester. Table 4 presents a stylized description of students' exposure to the policy. The table rows indicate values of the running variable (T_{ic}), while column 1 indicates first-semester students ($S_{ic} = 1$) included in equation 1. Each cell indicates the fall entering cohort implied by values of T_{ic} and S_{ic}. Thus, first-semester students in the academic year 2004-2005 belong to the Fall 2004 entering cohort, and so on. The shaded boxes in column 1 indicate that first-semester students were impacted by the policy in 2014-2015 and thereafter.

Students in their second and later semesters may also be affected by the policy, albeit indirectly. As an example, consider how the analysis might differ in the sample of student-by-course observations in the third semester ($S_{ic} = 3$). Third-semester students were exposed to the policy in 2015-2016 and thereafter. Effects are likely smaller because they represent a delayed impact of the cohort's prior exposure. First, the policy may affect the use of voluntary pass/fail elections in a later semester, given mandatory pass/fail grading in the first semester. Second, students may choose a different mix of courses in later semesters, perhaps due to earlier engagement with STEM courses and, possibly, a different choice of majors. Third, the policy may affect grade performance in later semesters if student performance in a prerequisite course was influenced in the first semester.

To compactly estimate later-semester effects, we report a single regression specification that pools student-by-course observations for all academic years and semesters (and nests equation 1):

$$O_{ic} = \left[\sum_{s=1}^{7} 1\{S_{ic} = s\} \times [\alpha_s + \beta_s(T_{ic} - t_s) + \gamma_s 1\{T_{ic} \geq t_s\} + \delta_s(T_{ic} - t_s) \times 1\{T_{ic} \geq t_s\}]\right] + \varepsilon_{ic}$$

where $t_1, t_2 = 14$; $t_3, t_4 = 15$; $t_5, t_6 = 16$; and $t_7 = 17$ (consistent with the shaded boxes in Table 4 which indicate the exposure of each semester's sample to the policy). The γ_s coefficients represent seven semester-specific discontinuities for a given dependent variable. Note that eighth-semester observations are excluded from the estimation sample for a practical reason. In the pandemic semester of Spring 2020 all course grades were reported as pass/fail, and no letter grades were reported by instructors. However, this leaves only two values of T_{ic} for eighth-semester students (see Table 4), which is insufficient to estimate a slope without overfitting the data.

As before, we control for pre-college covariates, and employ multi-way clustering of standard errors for students and courses. Table 5 reports the results. The coefficients in the first row replicate the first-semester results from Table 2.[22] Among second-semester students, the policy increased the probability of pass/fail election by 0.05, though it is not precisely estimated. This is explained by a quirk of implementation. First-year students must take a writing course in either fall or spring semester. Spring writing courses were thus included in the new policy to avoid excess demand for fall sections.[23] After the second semester, the estimates are all slightly negative. Only the third-semester coefficient—implying a decline of 2 percentage points—is

[22] The minor differences in first-semester coefficient estimates between Tables 2 and 5 are due to the use of pre-college controls in a pooled rather than a first-semester sample.

[23] The spring semester policy was abandoned after the policy's first review in Fall 2018 due to faculty dissatisfaction, perhaps foreshadowing our conclusions on grades and effort.

statistically different from zero. Despite the absence of a cap on the use of the voluntary pass/fail option, the negative coefficients suggest that students sought to avoid "too many" pass/fail courses in the wake of the first-semester policy.

The next column does not show any evidence that students substantially increased their engagement with STEM courses in subsequent semesters. The point estimates are all negative, although imprecisely estimated. The next column assesses whether later-semester grades were affected by the policy. After the first semester, there is only one coefficient that is statistically distinguishable from zero, but it becomes smaller and not statistically different from zero after accounting for interval-censored observations in the final column.

In addition to delayed effects on course choice and performance, we are interested in whether the policy affected students' decisions to take a temporary or permanent leave of absence from the College. There is no attrition in the first-semester sample, as described earlier. When $S_{ic} = 2$, 1.2% of initially-enrolled students are missing from the data. In semester 3 to 8, respectively, attrition is 5.4%, 6.9%, 32.6%, 33.8%, and 11.0%. Semesters 3 and 4 largely reflect the choice to transfer to another college, while semesters 5 and 6 reflect the large number of students that choose to study abroad during at least one semester of junior year.

Our primary concern is whether student attrition is affected by exposure to the policy. In Table A3, we limit the pooled sample to 49,932 student-by-semester observations—including 8,932 students—for semesters 2 to 7.[24] We report estimates from equation (2) with a dummy dependent variable indicating whether an initially-enrolled student is missing from the sample. There is no evidence of discontinuities in the probability of sample attrition at any of the semester-specific cutoffs. On the one hand, this provides evidence that non-random sample

[24] The number of unique students is lower than the full estimation sample because it necessarily excludes the cohort of students enrolling in Fall 2019 (for whom we only observe results in the first semester).

attrition does not introduce bias into estimates reported in Table 5. On the other hand, it provides evidence that the policy did not affect the probability that students left Wellesley College, particularly after the first year. This might have occurred had the policy significantly affected student well-being by providing an easier transition between high school and college.

B. Cumulative Outcomes

Rather than semester-specific effects, Figure 5 reports effects on cumulative outcomes across a student's full academic career at Wellesley College. In this case, the assignment variable is a student's entering fall cohort, beginning with fall 2004 and concluding with fall 2016 (since later cohorts have yet to graduate). The upper-left panel confirms, as expected that the policy sharply increased the total number of courses taken pass/fail by the 2014 cohort from approximately 3 courses to 6. Table 6 reports a point estimate of 2.9 courses, which is statistically different from zero. The increase is smaller than the 4 pass/fail courses mandated by the first-semester policy, but consistent with evidence from Table 2 that students modestly reduced their use of the voluntary pass/fail option in later semesters. Since Wellesley College does not cap the number of pass/fail elections, a plausible interpretation is that students do not believe that voluntary pass/fail elections are without cost, since they may be justifiably perceived as masking lower grades.

Figure 5 and Table 6 further show that the policy has no measurable effect on other cumulative outcomes in the full sample, including the total number of STEM courses taken by students. There is a negative coefficient on students' cumulative grade point averages (which includes all instructor-reported grades, even those with pass/fail options). However, it is not statistically distinguishable from zero at conventional levels, and we can rule out effects less than

-0.04 on the cumulative average. Thus, there is no evidence that first-semester reductions in grades had sustained effects on grade performance. Finally, the policy did not affect the probability of graduating with a STEM degree, or the probability of graduating with any degree from Wellesley College in the full sample. However, the confidence intervals are wide enough to include substantially positive effects. This provides some reason for caution, although positive effect would be consistent with the observed effects on student attrition from the college, or the cumulative number of STEM courses.

As in Table 2, Table 6 reports effects within subsamples defined by pre-college math skills and financial aid status. Unsurprisingly, the coefficients are estimated with less precision, and many confidence intervals include meaningfully positive and negative effects. For example, among students with above-median math skills who receive financial aid, there are positive coefficients on both STEM course-taking and the chances of receiving a STEM degree. The 95% confidence intervals include zero as well as positive effects as large as 1.2 courses and 11 percentage points in STEM graduation. While encouraging, more research is needed to justify stronger conclusions.

Similarly, there is a positive and significant coefficient on the probability of receiving a degree among students with higher-income students with above-median math skills. This deserves more scrutiny, given recent evidence that rising grades can explain the rising college completion rates of U.S. college students (Denning et al., 2022). It is possible that the increasing use of pass/fail grading is similarly bolstering graduation rates for at least some students at the College. However, we reserve judgment until additional cohorts of data are available.

6. Conclusions

In Fall 2014, Wellesley College began mandating the use of pass/fail grading in courses taken by first-year, first-semester students. This paper identified the causal effect of the policy, which substantially increased students' exposure to pass/fail grading in the first semester. The policy increased the probability of taking a lower-grading STEM courses in the first semester by 5 percentage points (a 20% increase over the pre-policy proportion), although it had no apparent effect on STEM courses or majors in later semesters. The policy lowered mean grades by 0.13, or 23% of a standard deviation. The paper explores evidence on several causal channels that might explain the grade effect. Sorting of students in STEM courses with lower grades can explain only a very small portion of the effect. Neither can the grade effect be explained by uniformly lower instructional quality or arbitrarily lower grading standards for first semester students.

In prior work, we showed that students' course evaluations sharply declined in the wake of exogenous reductions in grades, but holding student effort constant (Butcher et al., 2014). We find no such decline in course evaluations in response to declining grades in this setting. A natural interpretation is that effort declined, making students less inclined to "punish" faculty with lower ratings. Consistent with this, descriptive evidence from a qualitative survey of faculty suggests margins along which effort was reduced, such as attendance and class preparation.

It is possible that the policy produced benefits that our estimates did not fully capture. For example, one of the policy's stated objectives was to facilitate a successful transition from high school to college. We measure and rule out the most severe consequences of an unsuccessful transition, such as an early leave of absence, though it is possible that the grading policy reduced anxiety and produced related social and academic benefits for students (beyond the measurable outcomes related to course and major choice).

Finally, we add two caveats about the generalizability of our estimates. First, Wellesley College admits only women, and so we provide no evidence on whether the incentive effects of pass/fail grading might be larger or smaller among men. Second, it bears emphasis that our paper is most informative about the effects of a mandatory pass/fail option in which all courses in a semester are subject to the same grading standard. On the one hand, this may be quite informative about recent semesters, such as Spring 2020, when many colleges and universities implemented some version of pass/fail grading (including the 20 liberal arts colleges referenced earlier in the text). On the other hand, it may be less informative about voluntary pass/fail elections in which a single course is taken pass/fail. One hypothesis is that effort reductions would be even more pronounced in single courses, given the competing demands of letter-graded courses. However, this is clearly a topic worthy of additional research, given the prominence of pass/fail options in American higher education.

References

Ahn, T., Arcidiacono, P., Hopson, A., & Thomas, J. R. (2019). Equilibrium grade inflation with implications for female interest in STEM majors. Working Paper No. 26556. Cambridge, MA: National Bureau of Economic Research.

Artés, J., & Rahona, M. (2013). Experimental evidence on the effect of grading incentives on student learning in Spain. *Journal of Economic Education, 44*(1), 32–46.

Babcock, P. (2010). Real costs of nominal grade inflation? New evidence from student course evaluations. *Economic Inquiry, 48*(4), 983–996.

Babcock, P., & Marks, M. (2011). The falling time cost of college: Evidence from half a century of time use data. *Review of Economics and Statistics, 93*(2), 468–478.

Bar, T., Kadiyali, V., & Zussman, A. (2009). Grade information and grade inflation: The Cornell experiment. *Journal of Economic Perspectives, 23*(3), 93–108.

Becker, W., & Rosen S. (1992). The learning effect of assessment and evaluation in high school. *Economics of Education Review, 11,* 107–118.

Benjamini, Y., & Hochberg, A. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological), 57,* 289–300.

Betts, J. R. (1998). The impact of educational standards on the level and distribution of earnings. *American Economic Review, 88*(1), 266–275.

Betts, J. R., & Grogger, J. (2003). The impact of grading standards on student achievement, educational attainment, and entry-level earnings. *Economics of Education Review, 22,* 343–352.

Bonesrønning, H. (2004). Do the teachers' grading practices affect student achievement? *Education Economics, 12*(2), 151–167.

Bonesrønning, H., & Opstad, L. (2015). Can student effort be manipulated? *Applied Economics, 47*(15), 1511–1524.

Butcher, K. F., McEwan, P. J., & Weerapana, A. (2014). The effects of an anti-grade-inflation policy at Wellesley College. *Journal of Economic Perspectives, 28*(3), 189–204.

Denning, J., Eide, E. R., Mumford, K. J., Patterson, R. W., & Warnick, M. (2022). Why have college completion rates increased? *American Economic Journal: Applied Economics, 14*(3), 1–29.

Figlio, D., & Lucas, M. (2004). Do high grading standards affect student performance? *Journal of Public Economics, 88,* 1815–1834.

Gershenson, S. (2020). Great expectations: The impact of rigorous grading practices on student achievement. Thomas B. Fordham Institute.

Grodner, A., & Rupp, N. G. (2013). The role of homework in student learning outcomes: Evidence from a field experiment. *Journal of Economic Education, 44*(2), 93–109.

Grove, W. A., & Wasserman, T. (2006). Incentives and student learning: A natural experiment with Economics problem sets. *American Economic Review: Papers and Proceedings, 96*(2), 447–452.

Harris, G. A. (2010). *The impact of hidden grades on student decision-making and academic performance: An examination of a policy change at MIT.* Unpublished PhD dissertation, Harvard University.

Hausman, C., & Rapson, D. (2018). Regression discontinuity in time: Considerations for empirical applications. *Annual Review of Resource Economics, 10,* 533–552.

Hvidman, U., & Sievertsen, H. H. (2021). High-stakes grades and student behavior. *Journal of Human Resources, 56*(3), 821–849.

Jalava, N., Joensen, J. S., & Pellas, E. (2015). Grades and rank: Impacts of non-financial incentives on test performance. *Journal of Economic Behavior and Organization, 115,* 161–196.

Kolesár, M., & Rothe, C. (2018). Inference in regression discontinuity designs with a discrete running variable. *American Economic Review, 108*(8), 2277–2304.

Main, J. B., & Ost, B. (2014). The impact of letter grades on student effort, course selection, and major choice: A regression-discontinuity analysis. *Journal of Economic Education, 45*(1), 1–10.

McEwan, P. J., Rogers, S., & Weerapana, A. (2021). Grade sensitivity and the Economics major at a women's college. *AEA Papers and Proceedings, 111,* 102–106.

Oettinger, G. S. (2002). The effect of nonlinear incentives on performance: Evidence from "ECON 101." *Review of Economics and Statistics, 84*(3), 509–517.

Owen, A. L. (2010). Grades, gender, and encouragement: A regression-discontinuity analysis. *Journal of Economic Education, 41*(3), 217–234.

Owen, S. (2022). Ahead of the curve: Grade signals, gender, and college major choice. Working paper.

Rojstaczer, S., and Healy, C. (2010). Grading in American colleges and universities. *Teachers College Record,* March 4.

Sabot, R., & Wakeman-Linn, J. (1991). Grade inflation and course choice. Journal of Economic Perspectives, 5(1), 159–170.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental design for generalized causal inference.* Boston: Houghton Mifflin.

Stinebrickner, R., & Stinebrickner, T. R. (2014). A major in science? Initial beliefs and final outcomes for college major and dropout. *Review of Economic Studies, 81,* 426–472.

Table 1: Descriptive statistics for outcomes and pre-college variables

	Academic years (T_{ic})		
	2004-2005 to 2013-2014	2014-2015 to 2019-2020	Full sample
Panel A: Outcome variables			
Pass/fail course (1/0)	0.09	0.27	0.16
STEM course (1/0)	0.25	0.29	0.27
Grade points in course (0-4)	3.31	3.39	3.34
	(0.57)	(0.60)	(0.58)
Letter grade is missing (1/0)	0.07	0.08	0.07
Panel B: Pre-college variables			
STEM major preference (1/0)	0.33	0.39	0.35
Math SAT	684.01	695.22	687.66
	(67.22)	(68.61)	(67.88)
Verbal SAT	693.51	699.50	695.46
	(67.83)	(64.80)	(66.91)
Writing SAT	696.08	703.41	698.91
	(64.11)	(62.08)	(63.43)
Composite ACT	29.86	31.09	30.42
	(2.81)	(2.62)	(2.79)
Quantitative skills assessment	-0.01	0.03	0.00
	(1.00)	(1.00)	(1.00)
First-generation college student status (1/0)	0.11	0.13	0.12
Receives any financial aid (1/0)	0.59	0.60	0.59
Admissions committee vote total	4.15	4.20	4.17
	(0.39)	(0.40)	(0.40)
African-American (1/0)	0.06	0.06	0.06
Asian-American (1/0)	0.25	0.24	0.24
International (1/0)	0.09	0.12	0.10
Hispanic (1/0)	0.09	0.12	0.10
Biracial (1/0)	0.04	0.07	0.05
Other (1/0)	0.00	0.00	0.00
Unreported race and ethnicity (1/0)	0.03	0.00	0.02

Notes: Means are reported for all variables; standard deviations are in parentheses for continuous variables. The full sample includes 202,669 student-by-course observations, with 9,544 unique students and 6,651 unique courses. Although observations are missing for some pre-college variables; later regressions include dummy variables for observations with missing values.

Table 2: Student outcomes at the policy cutoff in the first-semester sample

	(1)	(2)	(3)
Pass/fail course (1/0)	0.823**	0.822**	—
	(0.017)	(0.018)	
N	38,214		
STEM course (1/0)	0.051*	0.049*	—
	(0.022)	(0.021)	
N	38,214		
Grade points in course	-0.156**	-0.139**	-0.129**
	(0.022)	(0.023)	(0.024)
N	34,063		38,214
A (1/0)	-0.037**	-0.030*	—
	(0.012)	(0.012)	
N	34,645		
A– (1/0)	-0.049**	-0.051**	—
	(0.012)	(0.013)	
N	34,645		
B+ (1/0)	-0.003	0.004	—
	(0.012)	(0.013)	
N	34,645		
B (1/0)	0.017	0.016	—
	(0.010)	(0.011)	
N	34,645		
B– (1/0)	0.010	0.017*	—
	(0.007)	(0.008)	
N	34,645		
C+ (1/0)	0.019**	0.010*	—
	(0.005)	(0.005)	
N	34,645		
C (1/0)	0.011*	0.012*	—
	(0.004)	(0.005)	
N	34,645		
C– or lower (1/0)	0.032**	0.024**	—
	(0.006)	(0.007)	
N	34,645		
Method	OLS	OLS	Interval
Controls	N	Y	Y

Notes: For each dependent variable, the table reports an estimate of γ_1 from equation (1). Heteroskedasticity-consistent standard errors are clustered by students and courses. Additional controls are given in panel B of Table 1; they also include dummy variables indicating missing values of those variables. ** (*) indicates statistical significance at 1% (5%), after controlling the false discovery rate across 23 coefficient estimates (see the text for details). The final column includes interval-censored observations for which letter grades were not reported.

Table 3: Student outcomes at the policy cutoff in the first-semester sample (in subsamples defined by math skills and financial aid status)

Sample	Pass/fail course (1/0)	STEM course (1/0)	Grade points in course	
Below-median math; no financial aid	0.822**	0.058*	-0.235**	-0.208**
	(0.023)	(0.028)	(0.056)	(0.056)
N of students	6,534	6,534	5,878	6,534
Below-median math; financial aid	0.792**	0.066**	-0.113**	-0.119**
	(0.024)	(0.024)	(0.040)	(0.042)
N of students	11,432	11,432	9,906	11,432
Above median math; no financial aid	0.849**	0.035	-0.123**	-0.106**
	(0.019)	(0.027)	(0.038)	(0.039)
N of students	9,197	9,197	8,316	9,197
Above median math; financial aid	0.832**	0.038	-0.118**	-0.108**
	(0.020)	(0.023)	(0.036)	(0.037)
N of students	11,043	11,043	9,958	11,043
Method	OLS	OLS	OLS	Interval
Controls	Y	Y	Y	Y

Notes: For each dependent variable, the table reports an estimate of γ_1 from equation (1), within subsamples. Heteroskedasticity-consistent standard errors are clustered by students and courses. Additional controls are given in panel B of Table 1; they also include dummy variables indicating missing values of those variables. ** (*) indicates statistical significance at 1% (5%), after controlling the false discovery rate across 16 coefficient estimates (see the text for details). The final column includes interval-censored observations for which letter grades were not reported.

Table 4: Entering fall cohorts by academic year (T_{ic}) and semester (S_{ic})

Academic year T_{ic}	Students attending during semester S_{ic}							
	1	2	3	4	5	6	7	8
2004-05	04	04	—	—	—	—	—	—
2005-06	05	05	04	04	—	—	—	—
2006-07	06	06	05	05	04	04	—	—
2007-08	07	07	06	06	05	05	04	04
2008-09	08	08	07	07	06	06	05	05
2009-10	09	09	08	08	07	07	06	06
2010-11	10	10	09	09	08	08	07	07
2011-12	11	11	10	10	09	09	08	08
2012-13	12	12	11	11	10	10	09	09
2013-14	13	13	12	12	11	11	10	10
2014-15	14	14	13	13	12	12	11	11
2015-16	15	15	14	14	13	13	12	12
2016-17	16	16	15	15	14	14	13	13
2017-18	17	17	16	16	15	15	14	14
2018-19	18	18	17	17	16	16	15	15
2019-20	19	—	18	—	17	—	16	—
N	38,214	35,379	33,828	30,495	21,408	19,051	24,294	—

Notes: The numbers in cells are the entering fall cohort of students (implied by the academic year and semester). "—" indicates that a cell is not included in the estimation sample in Table 5 (see text for details). Shaded cells indicate observations exposed to the policy. The final row indicates the number of student-by-course observations in each column (the N for column 8 is not reported because it is not part of the estimation sample). The total number of student-by-course observations in semesters 1 to 7 is 202,669.

Table 5: Student outcomes at the policy cutoffs in the all-semester sample

	Pass/fail course (1/0)	STEM course (1/0)	Grade points in course	
Semester 1 (γ_1)	0.822**	0.055*	-0.141**	-0.135**
	(0.017)	(0.021)	(0.022)	(0.023)
Semester 2 (γ_2)	0.048	-0.003	-0.029	-0.032
	(0.023)	(0.017)	(0.020)	(0.022)
Semester 3 (γ_3)	-0.020*	-0.002	0.015	0.017
	(0.008)	(0.018)	(0.019)	(0.021)
Semester 4 (γ_4)	-0.015	-0.023	-0.010	0.003
	(0.008)	(0.022)	(0.021)	(0.024)
Semester 5 (γ_5)	-0.014	-0.020	0.009	0.015
	(0.011)	(0.030)	(0.024)	(0.029)
Semester 6 (γ_6)	-0.021	-0.012	0.072*	0.048
	(0.010)	(0.030)	(0.023)	(0.029)
Semester 7 (γ_7)	-0.013	-0.014	0.004	-0.002
	(0.014)	(0.033)	(0.027)	(0.030)
N	202,669	202,669	187,762	202,669
Method	OLS	OLS	OLS	Interval
Controls	Y	Y	Y	Y

Notes: Each column reports seven discontinuity estimates of γ_1 to γ_7 from equation (2), using the dependent variable in the top row. Heteroskedasticity-consistent standard errors are clustered by students and courses. Additional controls are given in panel B of Table 1; they also include dummy variables indicating missing values of those variables. ** (*) indicates statistical significance at 1% (5%), after controlling the false discovery rate across 28 coefficient estimates (see the text for details). The final column includes all interval-censored observations for which letter grades were not reported.

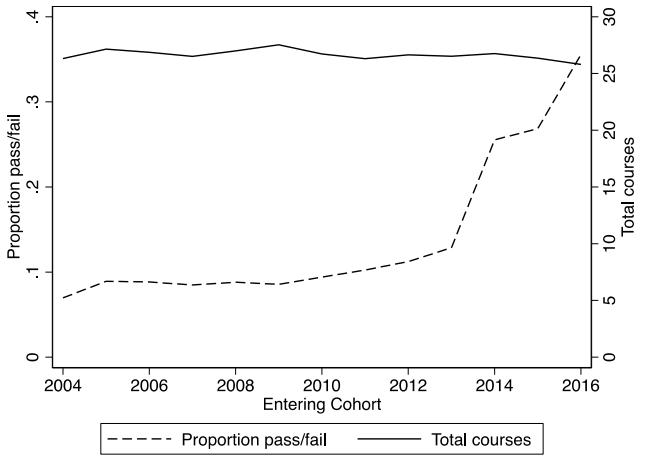
Table 6: Cumulative student outcomes at the policy cutoff, in the full sample and subsamples

Sample		Dependent variable: Cumulative number of pass/fail courses	Cumulative number of STEM courses	Cumulative grade point average	Received degree	Received a STEM degree
Full sample		2.949**	0.111	-0.011	0.023	0.013
		(0.087)	(0.237)	(0.015)	(0.013)	(0.021)
	N	7,716	7,716	7,700	7,716	7,109
Below-median math; no financial aid		2.514**	-0.216	-0.071	-0.003	0.002
		(0.214)	(0.500)	(0.036)	(0.034)	(0.043)
	N	1,443	1,443	1,442	1,443	1,307
Below-median math; financial aid		2.831**	0.143	0.013	0.007	0.039
		(0.158)	(0.389)	(0.029)	(0.023)	(0.035)
	N	2,354	2,354	2,346	2,354	2,173
Above-median math; no financial aid		3.233**	0.063	0.036	0.070*	-0.039
		(0.180)	(0.529)	(0.030)	(0.026)	(0.048)
	N	1,736	1,736	1,733	1,736	1,595
Above-median math; financial aid		3.069**	0.312	-0.038	0.017	0.031
		(0.155)	(0.477)	(0.028)	(0.022)	(0.041)
	N	2,183	2,183	2,179	2,183	2,034

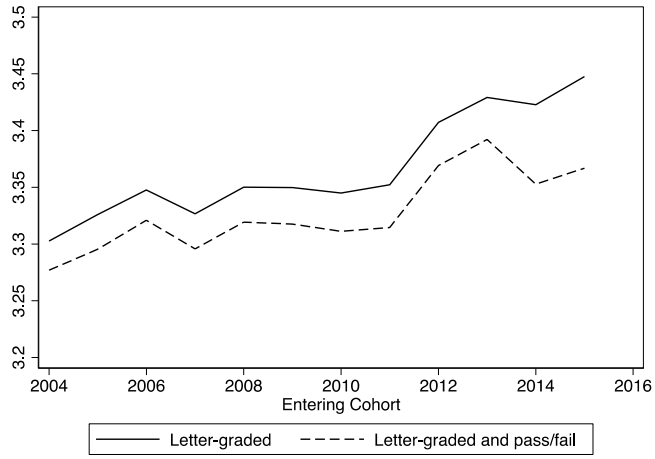
Notes: Each cell reports a discontinuity estimate using the entering cohort as the assignment variable and the dependent variable in the top row, within various subsamples. Heteroskedasticity-consistent standard errors are in parentheses. Additional controls are given in panel B of Table 1; they also include dummy variables indicating missing values of those variables. ** (*) indicates statistical significance at 1% (5%), after controlling the false discovery rate across 25 coefficient estimates (see the text for details).

Figure 1: Proportion of pass/fail courses and mean grades for entering cohorts

Panel A: Proportion of courses taken pass/fail and total courses taken

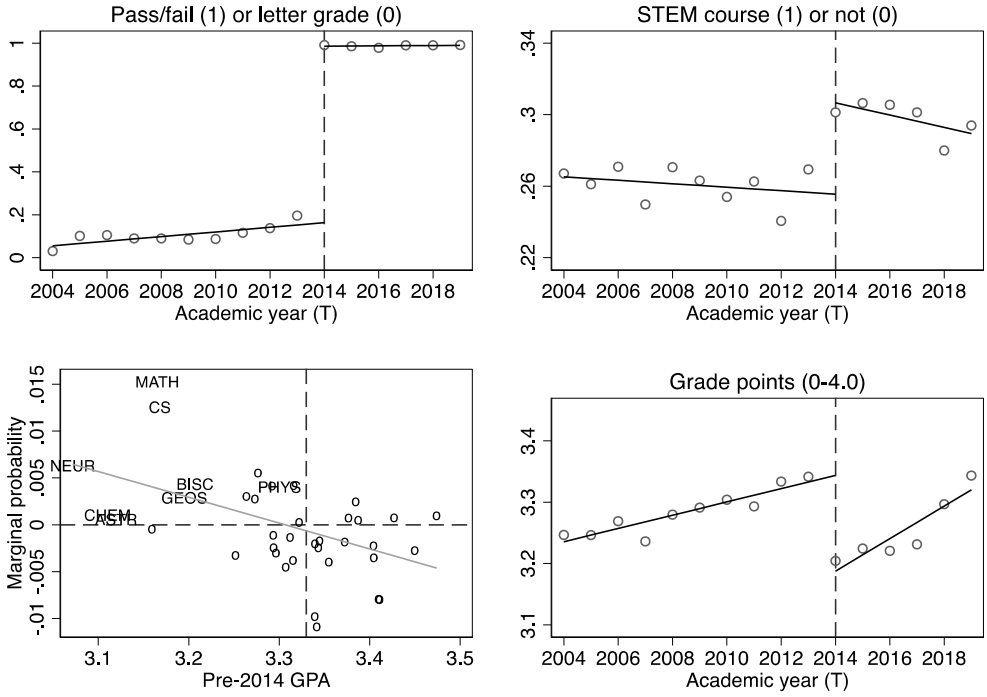


Panel B: Mean grades (letter-graded courses and all courses)



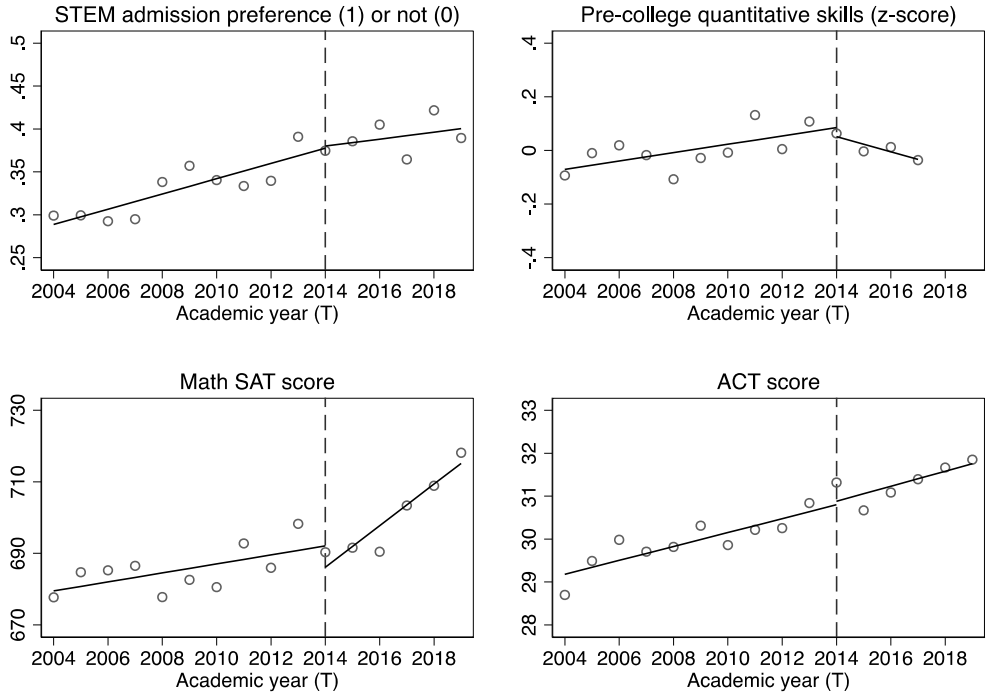
Notes: Panel A reports the proportion of courses taken pass/fail for any reason, including (1) voluntary pass/fail elections; (2) mandatory shadow grading for first-year, first-semester courses for the 2014 and later cohorts; and (3) mandatory pass/fail in the Spring 2020 semester. Panel B reports mean grade points, both for letter-graded courses and all courses (even if they used pass/fail grading). The latter includes the letter grade recorded by the instructor, for purposes of determining compliance with the grade inflation policy or—for the 2014 and later cohorts—for reporting on unofficial student transcripts.

Figure 2: Outcome variables in the sample of first-semester students



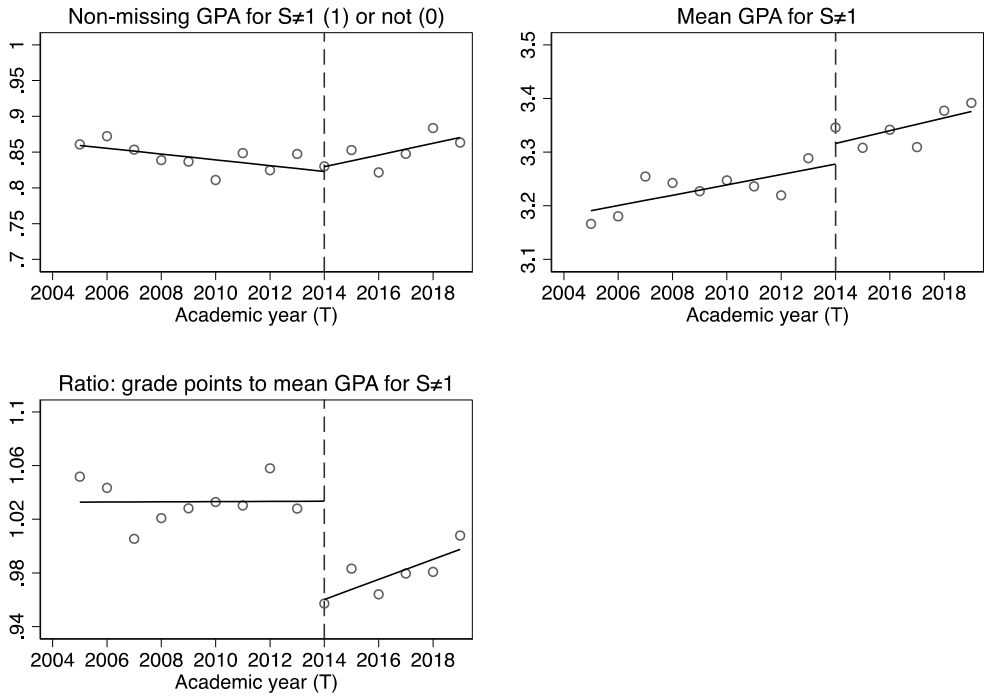
Notes: The sample in all panels includes student-by-course observations for first-semester students ($S_{ic} = 1$). Excepting the lower-left panel, the circles indicate unsmoothed means of student-by-course observations within academic years, and the lines are fitted values from estimates of equation (1), using the specified dependent variable. See the text for details on the lower-left panel.

Figure 3: Pre-college variables in the sample of first-semester students



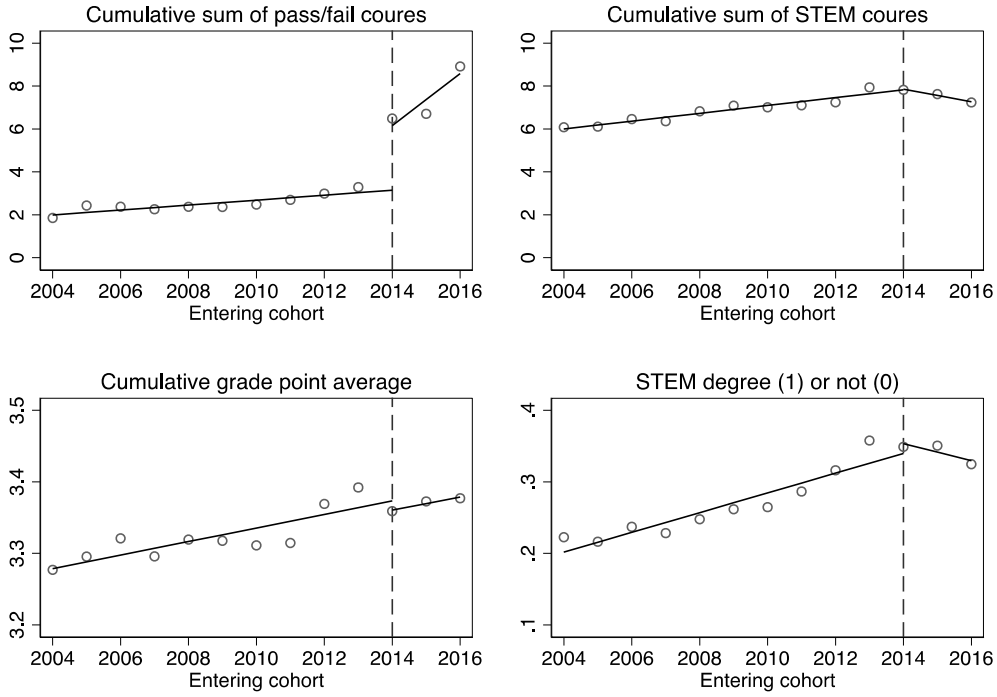
Notes: The sample in all panels includes student-by-course observations for first-semester students ($S_{ic} = 1$). The circles indicate unsmoothed means of student-by-course observations within academic years, and the lines are fitted values from estimates of equation (1), using the specified dependent variable.

Figure 4: Variables for non-first-semester students in the sample of first-semester students



Notes: The sample in all panels includes student-by-course observations for first-semester students ($S_{ic} = 1$). The circles indicate unsmoothed means of student-by-course observations within academic years, and the lines are fitted values from estimates of equation (1), using the specified dependent variable (see text for further explanation of the dependent variables). In the upper-left panel (N=31,686), the estimate (standard error) at the discontinuity is 0.007 (0.027). In the upper-right panel (N=26,812), it is 0.039 (0.027). In the lower-left panel (N=26,812), it is -0.073 (0.014). Standard errors adjust for clustering within students and courses.

Figure 5: Cumulative outcome variables for cohorts entering between Fall 2004 and Fall 2016



Notes: The sample in all panels includes student-level observations. The circles indicate unsmoothed means of student observations within entering fall cohorts, and the lines are fitted values from estimates of equation 1, using the specified dependent variable (see text for further explanation of the dependent variables).

Online Appendix

Table A1: Student outcomes at the policy cutoff in the first-semester sample (bandwidth=6)

	(1)	(2)	(3)
Pass/fail course (1/0)	0.796**	0.795**	—
	(0.025)	(0.025)	
N	26,147		
STEM course (1/0)	0.055*	0.055*	—
	(0.025)	(0.024)	
N	26,147		
Grade points in course	-0.153**	-0.128**	-0.120**
	(0.025)	(0.025)	(0.026)
N	23,057		26,147
A (1/0)	-0.034*	-0.019	—
	(0.015)	(0.015)	
N	23,442		
A– (1/0)	-0.050**	-0.048**	—
	(0.015)	(0.015)	
N	23,442		
B+ (1/0)	-0.001	0.001	—
	(0.014)	(0.015)	
N	23,442		
B (1/0)	0.008	0.001	—
	(0.012)	(0.013)	
N	23,442		
B– (1/0)	0.018	0.018	—
	(0.009)	(0.009)	
N	23,442		
C+ (1/0)	0.017**	0.009	—
	(0.005)	(0.006)	
N	23,442		
C (1/0)	0.011*	0.009	—
	(0.005)	(0.005)	
N	23,442		
C– or lower (1/0)	0.029**	0.028**	—
	(0.007)	(0.007)	
N	23,442		
Method	OLS	OLS	Interval
Controls	N	Y	Y

Notes: For each dependent variable, the table reports an estimate of γ_1 from equation (1). Heteroskedasticity-consistent standard errors are clustered by students and courses. Additional controls are given in panel B of Table 1; they also include dummy variables indicating missing values of those variables. ** (*) indicates statistical significance at 1% (5%), after controlling the false discovery rate across 23 coefficient estimates (see the text for details). The final column includes interval-censored observations for which letter grades were not reported.

Table A2: Covariate balance at the policy cutoff in the first-semester sample

	Admission STEM preference (1/0)	Math SAT	Verbal SAT	Writing SAT	ACT	Quantitative Reasoning Score	First-generation status (1/0)	Any financial aid (1/0)
γ_1	0.002	-5.955	0.902	-5.993	0.080	-0.034	-0.024	0.011
	(0.021)	(3.506)	(3.249)	(3.328)	(0.174)	(0.047)	(0.014)	(0.020)
N	38,214	30,095	30,099	24,959	15,418	33,263	38,214	38,214
	Application vote total	African-American (1/0)	Asian-American (1/0)	Inter-national (1/0)	Hispanic (1/0)	Biracial (1/0)	Other (1/0)	Not reported (1/0)
γ_1	0.040	-0.001	0.016	-0.005	0.002	-0.018	0.002	-0.007
	(0.018)	(0.010)	(0.018)	(0.014)	(0.013)	(0.010)	(0.001)	(0.005)
N	35,711	38,214	38,214	38,214	38,214	38,214	38,214	38,214

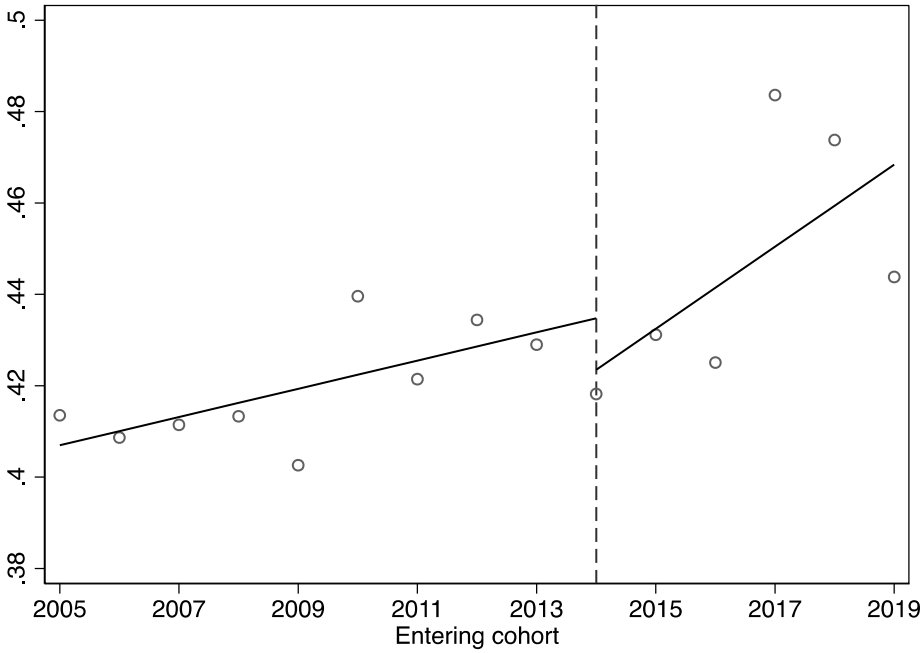
Notes: For each dependent variable, the table reports a discontinuity estimate from equation (1). Heteroskedasticity-consistent standard errors are clustered by students and courses. ** (*) indicates statistical significance at 1% (5%), after controlling the false discovery rate across 16 coefficient estimates (see the text for details).

Table A3: Student attrition and missing grade data at the policy cutoffs

	Initially-enrolled student is missing (1) or not (0)
Semester 2 (γ_2)	0.002
	(0.004)
Semester 3 (γ_3)	-0.003
	(0.010)
Semester 4 (γ_4)	-0.004
	(0.011)
Semester 5 (γ_5)	0.012
	(0.020)
Semester 6 (γ_6)	-0.032
	(0.022)
Semester 7 (γ_7)	-0.014
	(0.014)
N	49,932

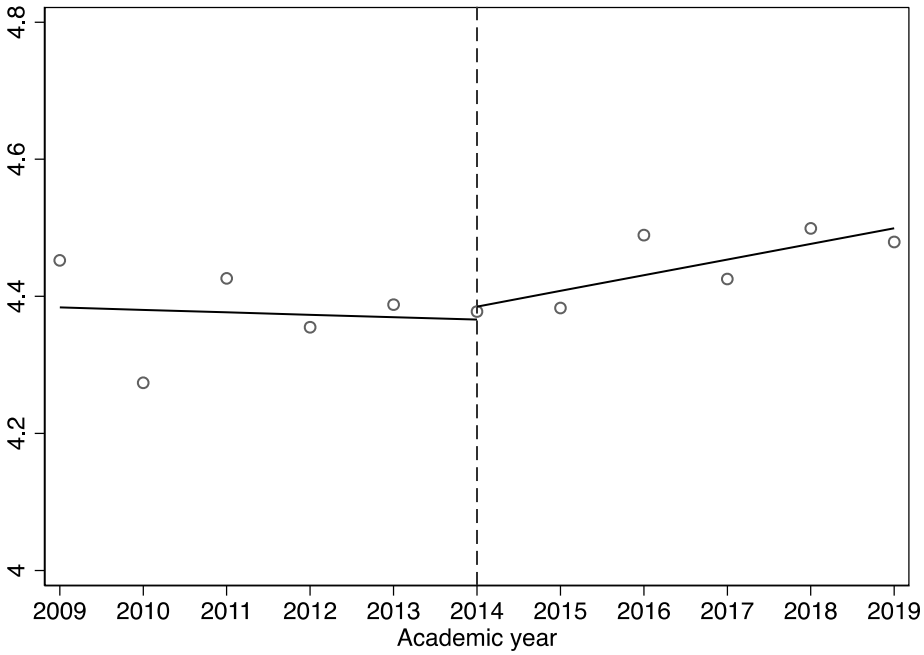
Notes: The table reports six discontinuity estimates from equation (2) for the dependent variable. Observations from semester 1 are excluded from the regression because no initially-enrolled students are absent from the dataset in semester 1. Heteroskedasticity-consistent standard errors are clustered by students. ** (*) indicates statistical significance at 1% (5%).

Figure A1: Yield rate among admitted applicants



Notes: The yield rate is the number of enrolled students (first-year, first-semester) divided by the number of admitted applicants. They are obtained from the Common Data Set (https://www.wellesley.edu/oir/instdata). The estimate (robust standard error) of the discontinuity at the dashed line is -0.011 (0.010).

Figure A2: Instructor evaluations in 100-level courses



Notes: The sample includes 1,936 observations for fall-semester, 100-level courses taught at Wellesley College. The dependent variable is the course average of instructor evaluations on a 5-point scale. The estimate (robust standard error) at the discontinuity is 0.020 (0.049).