Do Grade Signals Drive the Gender Gap in STEM? Evidence From a Regression Discontinuity

Please click here for the most recent version of this paper.

Adam Bestenbostel*

Department of Economics

Texas A&M University

November, 2021

Abstract

A common hypothesis for the gender gap in STEM, which is believed to drive much of the gender gap in earnings, is that women are more responsive to the negative grade signals that are common in STEM. I test this hypothesis by applying a regression discontinuity design to the underlying numerical scores of more than 21,000 university students in feeder courses in STEM and economics. Results indicate that letter grade thresholds (with no plus/minus modifiers in this context) have no effect on STEM major choice for either women or men. This is true even for susceptible subgroups, within particular majors, or at particular grade thresholds.

*adam.bestenbostel@tamu.edu

1 Introduction

In 2017, U.S. women earned 80 cents to the dollar relative to men (Fontenot, Semega, and Kollar, 2017). While there are several potential explanations for this, gender differences in occupational choice are estimated to be responsible for 44 percent of the wage gap (Goldin, Kerr, Olivetti, and Barth, 2017). For example, women held 14 percent of full-time architecture and engineering jobs and 25 percent of full-time computer and math jobs in 2016, which pay 25 and 36 percent more than the national average for bachelor's degree graduates, respectively (Bureau of Labor Statistics, 2017, 2019).

The gender disparity in STEM occupations can be traced back to gender differences in college majors. While women make up 58 percent of all bachelor's degree recipients, only 36 percent of bachelor's degrees in STEM are conferred to women (National Center for Education Statistics, 2019). Indeed, Speer (2021) attributes more than half of the STEM occupation gap to differences in college major, and Card and Payne (2021) suggest these differences in major explain up to a fifth of the wage gap. This is especially striking given that female students outperform their male peers in the relevant STEM courses at all levels of education (O'Dea, Lagisz, Jennions, and Nakagawa, 2018). A critical question, then, is why women are so much less likely to select into a STEM field. One recent hypothesis is that women are more deterred by the grade environment typical of STEM courses in college, where desirable grades are relatively scarce. For example, Goldin (2015) writes that "Grades in Principles are extremely important in determining whether females major in economics. But that is far less the case for males." The relative aversion to STEM majors for women could be driven by documented gender differences in preferences with respect to risk or competition, as STEM majors are typically viewed as both more difficult/higher risk and where good grades are harder to earn. Indeed, Niederle and Vesterlund (2010) suggest that "evidence of a large gender gap in mathematics performance at high percentiles in part may be explained by the differential manner in which men and women respond to competitive test-taking environments." The purpose of this paper is to document the extent to which

grade signals drive the gender gap in STEM.

To do so, I use a regression discontinuity design that compares the college major decisions of men and women just above and below the thresholds that determine final letter grades. I use administrative data from a large, public university linked to final number grade. This setting provides several important advantages for implementing this analysis. First, I am able to focus on only those courses where there is little scope for grade manipulation. For example, exams are multiple choice or final grades have been preemptively verified, and in all cases grading policies are strictly enforced and instructors have specifically stated they do not allow arbitrary revision of final grades. Second, I am able to test for differences across a wide range of courses, instructors, and cutoffs to see if grades matter anywhere. Third, I am able to do so in a setting where only full letter grades are given, without pluses or minuses. This generates much larger discontinuities than in the presence of grade modifiers. The identifying assumption of this approach is that the probability of majoring in STEM should be smooth across the grade threshold in the absence of grade signal effects.

In order to implement the regression discontinuity design, I use administrative data on more than 21,000 student-course observations covering a total of 7 distinct STEM courses taught by 16 different professors at a large university. These data include final number grade (e.g. 0 to 100), letter grade, gender, major, and other individual student characteristics such as prior academic performance, standardized test scores, first generation, transfer, and international status, as well as whether a student has applied for financial aid. These data come from a variety of core, compulsory, first and second-year STEM courses including biology, calculus, computer programming, economics,[1] engineering statics, and organic chemistry. Importantly, I include only those courses for which instructors explicitly declared that they did not allow for or engage in any manipulation of students around the grade thresholds.[2]

[1] I do not include economics in the main results since the major is not classified as STEM at this university, although I do use it for later analysis which has identical results. These economics data add an approximate 6,000 observations from 2 courses and 4 more instructors.

[2] As discussed later in section 3, instructors often choose to place thresholds where there were large gaps in the distribution and not at predetermined cutoffs such as 90.00%.

The major threat to identification in this context is manipulation around the cutoff. For example, a problematic behavior would be if instructors were to grade exams subjectively so that students judged to be different on unobservable characteristics (such as being more or less deserving) would land on the other side of the threshold. There are three reasons why I believe this is unlikely to have occurred. First, these are all very large classes with an average enrollment of 166, which leaves little opportunity for professors to know students or their unobservables. Second, because of these large class sizes, almost all the exams are multiple choice and machine-graded, which allows few opportunities for the type of subjective grading that would be problematic. Third, and perhaps most importantly, I screened for this behavior when I recruited instructors for their numerical scores. Specifically, I asked whether instructors would bump up scores for students they thought were particularly deserving of a higher grade. Nearly all said they did not; three who suggested they might do so were excluded from the sample.[3] A related threat to identification is if instructors were to draw the thresholds such that the students just above the threshold were unobservably (to the researcher) different than those just below the threshold. While it is clear both from the data and my conversations with the instructors that they do choose the thresholds – many said they put them where there are large gaps in the numerical distribution, rather than at predetermined cutoffs – none stated that they did so based on who the students were. Again, in these large classes, it is unlikely for instructors to know students well. I also note that for many of these courses, instructors described meeting with others and agreeing on a common grade distribution across sections, limiting the flexibility of any one instructor choosing letter thresholds. Nonetheless, I show that results are robust to an exercise where I exclude observations within up to 2 points of the threshold.

Results indicate that despite the widespread belief that grades matter differently for men and women's major choice, there is no evidence that letter grade signals affect graduating with a STEM major for either gender. This result is robust to a variety of specifications and

[3]One said, "I would never do that because when word got out, there would be a never-ending line of students at my door or emailing me to beg for a higher grade."

subsamples, including the addition of control variables, expanded bandwidths, or focusing only on male or female instructors, first generation college students, or freshman. Importantly, estimates enable me to rule out that a higher letter grade reduces the likelihood of graduating with a STEM major by more than 3.2 percentage points (6.0 percent) for women. This is smaller than other input factors to female STEM participation.[4]

In assessing the role of grade signals on college major, this study is most related to two others that examine the likelihood of majoring in economics across grade thresholds. Owen (2010) uses a similar methodology and data but is limited to data from one economics course with 1,300 students, and thus examines only majoring in economics as an outcome. Main and Ost (2014) also use a regression discontinuity on a sample of 2,126 students in micro and macro principles courses and find no effect. Their context involves grade modifiers, so the distinction between an A- and a B+ may not be strong enough to induce an effect. My paper differs from these in two key ways. First, my sample includes many more students and instructors. With more than 21,000 student observations, this is 10 to 16 times larger than the samples examined in these studies. Additionally and perhaps most importantly, I examine behavior beyond economics to cover science, technology, engineering, and mathematics. As a result, the main contribution of this paper is that it is the first to my knowledge to evaluate the impact of letter grades on the gender gap in STEM majors.

In addition, the paper also contributes to a small but growing literature providing evidence outside the laboratory setting on gender responses to environmental factors such as feedback signals. Johnson and Helgeson (2002) find that women were more likely to agree with employee performance evaluations, and their self-esteem was affected to a greater degree than that of men by both positive and negative reviews. Additionally, Mayo, Kakarika, Pastor, and Brutus (2012) show that women in an MBA program are more likely than men to align their self-ratings with those from peers. Lastly, Kugler, Tinsley, and Ukhaneva (2017)

[4]Carrell, Page, and West (2010) find having all female instructors for math and science increase graduation with a STEM major by 15.5 p.p. (∼40%) for women with above average math SAT scores, and Porter and Serra (2020) find that exposure to a "successful and charismatic" female role model can induce women to major in economics by 8 p.p. (∼100%).

use a switching model based on a selection-on-observables identification strategy to estimate that female students choose to switch out of STEM not solely because of grades, but only when they face multiple signals of "lack of fit" such as poor grades in addition to being in a male-dominated field or external stereotypes. My paper complements this literature by demonstrating that women do not adjust their major choices away from STEM as the result of performance feedback from instructors in the form of letter grades.

In short, this paper demonstrates that letter grade signals do not seem to explain any of the gender gap in STEM majors, and thus do not explain any of the gender wage gap that exists due to occupational differences between men and women. This also suggests that while there may indeed be behavioral responses that generate the gender gap in majors, it is not a result of grade signals. In the following sections, I will outline my emprical approach, data, and results.

2 Research Design

It is difficult to assess the impact of grade signals on college major because they are likely correlated with many unobserved factors such as preferences or determination, which impact major themselves. To overcome these concerns, I utilize a regression discontinuity approach in the context of letter grade cutoffs. For the main analysis I utilize a multiple-threshold "stacked" regression discontinuity design following Pop-Eleches and Urquiola (2013). The advantage of using the stacked approach in this context is greater statistical power to detect significance in results. Formally, I estimate the following model:

$$y_{iz} = \alpha + \beta \cdot \tilde{\text{score}}_{iz} + \gamma \cdot 1[\tilde{\text{score}}_{iz} \geq 0] + \delta \cdot \tilde{\text{score}}_{iz} \cdot 1[\tilde{\text{score}}_{iz} \geq 0] + \text{Cutoff}_z + X_i + \epsilon_{iz} \quad (1)$$

Here y_{iz} represents student i's outcome for cutoff z, and $\tilde{\text{score}}_{iz} = \text{score}_i - \text{score}_z$ is the standardized running variable. Cutoff_z represents a set of three dummy variables, one for each letter grade cutoff, and X_i represents a vector of individual-specific control variables.

I estimate the model with a local linear regression and uniform kernel. The coefficient of interest is γ, which indicates the magnitude of the discontinuity at the stacked threshold. Here, it can be interpreted as the effect of a higher letter grade in core STEM courses on the likelihood of a student graduating with a STEM major. Standard errors are two-way clustered at the instructor and term levels to account for correlation both within instructor and within terms.

There are a number of reasons to believe a final letter grade can be thought of as a signal which contains information, including that pertaining to student aptitude for a given subject. First, letter grades affect GPA which matters to potential employers and students know this. In some cases, such as for medical school applicants, letter grades for particular courses can matter even more. The students' internal decisions about college major should therefore be affected by their knowledge of the letter grade and its value as an external signal. Second, it stands to reason that the final letter grade and not the precise final number grade is more easily recalled by the student as a result of availability bias. Third, the letter grade can be thought of as validation of effort and knowledge over the entire semester. No matter what thoughts are in the minds of those students who just missed the higher letter grade cutoff, they are certainly different from those students who just *made* the same cutoff and are likely quite pleased with the result of all their hard work.

The identifying assumption of this approach is that all other determinants of major besides letter grade vary smoothly across the letter grade cutoff. The major threat to identification is the possibility that students or professors precisely manipulate student grades around the cutoff according to some unobserved determinant of the outcome. There are several reasons why this identifying assumption is likely to hold in this context. The first is that while students can affect their grade through additional effort, they cannot precisely control the exact number grade. More practically, they likely do not even know the exact letter grade thresholds since instructors often adjust them at the end of the semester in order to yield a certain letter grade distribution, or in order to draw cutoffs where there are gaps

in the numerical score distribution. A more significant concern, however, is that individual students could approach the instructor after final grades have been released, seeking to improve their score to achieve a better letter grade. For example, a student who is "more motivated" could be more likely to visit the instructor after final grades are released, and since this attribute is correlated with both desire to achieve a higher letter grade as well as persistence in STEM, this would bias estimates. Similarly, instructors could theoretically adjust scores in order to assign what they believed to be the appropriate grades to those near the threshold. Importantly, it will not invalidate my research design if instructors shift the entire grade distribution to bump up letter grades overall - only if they alter individual numerical scores. Knowing these issues in advance, I carefully selected courses and instructors based on reputation and then spoke with the instructors individually.

In interviews, instructors stated explicitly that they did not allow manipulation of grades. Three professors were unable to make this assertion or mentioned changing individual student scores to help them across grade thresholds - each of these are thus excluded from all data and analysis in this paper. There are also two specific facts gleaned from these interviews that mitigate concerns over grade manipulation. First, no instructor in my sample allows special extra credit opportunities for students close to grade cutoffs, which is one potential pathway for manipulation. Second, the vast majority of the courses in my sample use multiple choice exams, which are objectively machine-graded.[5] Most of the other assignments for the courses in my sample are also computer-graded through online learning management systems. Although this does not prevent a student coming to seek regrades, it does mitigate concerns if most grades are objective and machine-graded with little room for flexibility.

Overall, instructors reported that students seldom if ever even attempt to manipulate their way to a higher letter grade each semester. This is possibly a reflection of their reputation as being strict with respect to grading policies. Additionally, though I did not ask this question in interviews, three of the instructors reported that they take the extra precaution

[5]Many use a full multiple choice exam including biology, calculus, computer programming, and economics, although one calculus instructor reported not using a multiple choice exam.

of identifying students close to letter grade thresholds prior to releasing them, and preemptively regrade major assignments for these students. This is so that if students do come to complain about grades, nothing will change. It is also possible for students to contest grades at the department or college level, although that is rare in general and would not be a concern in this context since it would only affect the final letter grade, not the number grade.

In these discussions, many instructors said that while they did not move numerical scores, they would often look for natural gaps in the grade distribution to draw the actual letter grade cutoff. The reasons for this are typically twofold. First, a certain letter grade distribution is often targeted to match that of other classes in the same or previous semesters. Second, instructors know if the next highest score below a letter grade cutoff is further away, students will be less likely to complain about being close to it and less likely to plead for a higher letter grade. Looking for a break in the numerical distribution that yields the desired portion of each letter grade is a solution to both problems. This will likely lead to a density histogram which is not uniform due to a drop in frequency of observations close to the cutoff. Importantly, this does not invalidate the research design as long as they aren't drawing the cutoff where based on some unobservable-to-the-researcher characteristics, e.g. where there are "less motivated to succeed" students with the lower letter grade and "more motivated" students with the higher one. I address this concern in Section 4 by performing a "donut" RD that omits those observations within up to 2 grade percentage points of the threshold. It is also helpful that class sizes for my sample are typically quite large as is true with many core classes at public universities, so instructors are less likely to know individual students well.

I also perform the standard empirical tests of the identifying assumption. I plot a histogram of observations near the stacked cutoff in Figure 1 and check for abnormal heaping. I regress individual covariates in table A.2. And using student characteristics including pre-existing academic ability to predict whether a student has a STEM major, I test for and

find evidence of smoothness through the letter grade threshold in Figure 2. These tests are discussed in greater detail in section 4 with other results.

3 Background and Data

Data for this project come from a large, public university in the U.S. and include observations on students who were enrolled in core STEM courses including biology, calculus, computer programming, economics, engineering statics, and organic chemistry. The courses were selected because interviews with advisors and professors indicated that they were challenging first- or second-year classes critical to the curriculum of popular STEM majors including biology, pre-med, computer science, and engineering. In addition, each course selected featured significant enrollment of female students.

At this university, students are admitted with a declared major and are allowed to request a change of major at any time. This means that most students in my sample data are taking these classes to satisfy the requirements of a major they are already pursuing. However, they can also switch to a different major if they wish. Change of major requests are generally allowed as long as the student has met the degree requirements for the program they wish to switch into, though the decision to grant such requests can also depend on space and demand. Typically this means it is relatively easy to switch from a STEM major to a non-STEM major since the GPA and course requirements tend to be less strict. To transfer into another major, a student would have had to have already completed some of the core coursework (e.g. courses in my sample data) and achieved a minimum GPA.

A critical yet unique aspect of the data I use is that in addition to containing final letter grades, they also contains final numerical scores used to determine the letter grade. I collected these data by meeting with individual instructors and coordinating the secure uploading of these scores to the administrative office. That office then matched these scores to university records including final letter grade, major, gender, and other background characteristics. All

data were then carefully de-identified to preserve anonymity.

For my main analysis, I use 21,533 student records from 16 instructors representing seven different courses and numerous STEM majors including engineering and pre-med as well as biology, chemistry, and math. For each student-by-course, I observe final numerical grade, final letter grade, gender, and major as of graduation. Additionally, the data include information on prior academic ability such as SAT scores,[6] high school rank, and prior post-secondary GPA, student status as a first generation, transfer, or international student, and an indicator for whether the student applied for financial aid. The data span 2004-2019, depending on when each individual instructor was active in the course. Data on major only exist for students who have graduated. To account for this, all analyses are restricted to students for whom the six-year graduation rate can be determined, which leaves the sample of 21,533 observations. Additionally, the main outcome is defined as "graduated with a STEM degree within six years."[7]

In later analyses, I use an additional sample from two economics courses, principles of and intermediate microeconomics. This makes for a total of twenty instructors, nine courses, and 27,572 observations. These data are excluded from the main STEM analysis because the courses are commonly taken by many non-STEM majors, and because this university's economics major is not STEM according to the Department of Homeland Security definition.

Summary statistics are shown in Table 1. Women make up approximately 46% of my sample, which includes courses that are typically male-dominated, such as engineering, as well as those which are often female-dominated, such as biology. Along most characteristics, men and women look similar on average. Interestingly, women seem to have a better high school rank at 48.2 versus men at 64.6 (lower is better) but worse SAT scores with 1198 versus 1249. Women also graduate at a slightly higher rate within six years (90% vs. 86% for men) and are less likely to do so with a STEM major (50% vs. 66%), both of which are

[6]For cases where I observe ACT scores, I use official concordance tables to convert to SAT scores and then take the maximum of the ACT concordance score or the SAT score, if the latter exists.

[7]The sample with a six-year graduation rate does include some observations as recent as 2019 - likely from students who just needed to complete that one STEM course to complete their graduation requirements.

consistent with other literature.

The exact cutoffs between letter grades typically vary both across instructors and even within an instructor across semesters, in part because instructors say they can change the grade thresholds based on the overall distribution of grades and the gaps in the distribution. To determine the grade thresholds, I use an objective algorithm to determine the cutoff rule for each instructor-course-semester group of students separately. For each class, I find the lowest A and highest B. (In a small number of cases where the class groups are not known, I use instructor-term groups.) Then, for test cutoffs at each 0.01 numerical grade increment between these two scores, I regress a binary indicator for receiving the higher letter grade on a binary indicator for whether a numerical grade is equal to or greater than the incremented "cutoff." I set the RD cutoff for the class group equal to the test cutoff at the increment that generates the best fit R^2. I repeat this procedure for the B/C and C/D cutoffs.

While my data contain scores, letter grades, and major information for each student enrolled in the course, there are cases where I am missing data for certain covariates. Table A.1 shows to what extent missing variables exist in the data. Roughly half of all observations are missing at least one variable, and the most commonly missing variables are ACT score (46% missing), high school rank (30%), SAT score (14%), transfer hours (12%), and prior college GPA (10%). All other variables have missing data for fewer than 10% of observations. I use official concordance tables to convert ACT scores to SAT scores. I then take the maximum of each student's SAT score, if it exists, and their ACT concordance score, if it exists. In Table A.1, the max(SAT score, ACT concordance score) is only missing from 3% of the sample. To further address concerns regarding missing variables, I plot the number of missing variables across the threshold and verify that is smooth in Figure A.1. Specifically, I plot the number of missing variables per observation as well as the percent of observations missing at least one variable. In both figures, the binned data are smooth (if noisy) across the threshold. This suggests that the appearance and number of missing variables is unconcerning. I also demonstrate robustness of my main results using both a subsample of data containing no

missing values as well as with the full sample using multiple imputation (Tables A.4 and A.5, respectively). I discuss this robustness exercise in greater detail in section 4.

4 Results

4.1 Tests of the identifying assumption

With this regression discontinuity design, the identifying assumption is that all other determinants of major vary smoothly through letter grade thresholds. While I believe the institutional features described above indicate there is little reason to doubt this assumption *ex ante*, I also perform the standard statistical tests of this assumption. First, I present frequency histograms in Figure 1. Overall, observations appear smooth across the threshold, which supports the notion that there is no bunching above the threshold which could occur due to manipulation of the running variable. There is a dip near the cutoff for men and women, but this dip is roughly equal both above and below the cutoff. This is because many instructors look for a natural break or gap in the final grade distribution and draw letter grade cutoffs there, which would only invalidate my design if instructors are also choosing cutoffs to push specific students above or below a cutoff.

Additionally, I test whether observed covariates are smooth across the threshold. Specifically, I regress each variable on the left-hand side of equation 1 to verify there is no discontinuity at letter grade cutoffs. These include gender, transferred credit hours, prior GPA, high school rank, SAT scores, whether a student applied for financial aid, whether a student is new to the university or any college, same-semester GPA excluding courses in the sample, student classification (e.g. freshman), first-generation status, international student status, and transfer status. I plot these variables around the threshold in Figure A.2 and estimate them in Table A.2. While my main analysis focuses on men and women separately (panels B and C of the table, respectively), I also estimate for the pooled sample of all students together (panel A). I do this using the optimal bandwidth for the main outcome of the sample used

in each panel. By and large, the visual evidence and point estimates show no discontinuity across the threshold, with a few exceptions. From the table, estimates are significant for the variables gender, SAT max, freshmen, and sophomore for the pooled sample, and junior and transfer for women. Visually there is some evidence of a discontinuity for SAT max for both genders, and junior, senior, and transfer for women, but not for any of the other variables that appear significant in the table. This is roughly what one should expect to see due to random chance.

Relatedly, I also examine whether predicted outcomes vary smoothly across the threshold. Instead of considering individual covariates separately, this is a linear combination of these factors where the relative weights are chosen to best predict the outcome of interest. Specifically, I predict STEM major using all predetermined characteristics and plot the predicted outcomes in Figure 2. The figure shows that predicted outcomes vary smoothly through the letter grade threshold, which is consistent with the identifying assumption.

4.2 Effect of threshold crossing on letter grade

Following these tests of the identifying assumption, I transition to the effect of numerical grade crossing the threshold on letter grade. This result is plotted in Figure 3. There is clear, visually compelling evidence that threshold crossing is associated with a higher letter grade, and that this effect is close to 1. Some instructors have multiple sections within the same semester, so they could in theory have different cutoffs for each section. This means it's possible that the highest B is higher than lower A, for example, since I do not always observe different sections and group by instructor-term in these cases instead. Therefore, it is possible for there to be letter grades on the wrong side of the threshold and this can explain why the effect of threshold crossing is not exactly equal to 1. It is also possible that in rare cases, letter grades (but not number grades) are revised at the department or college level, or that there are administrative errors in the data.

4.3 Graduation

Graduation rate is an interesting outcome by itself, but it is also important because I only observe major for students who have graduated. It is possible that students respond to feedback signals at letter grade cutoffs in a way that impacts their probability of graduating. This could occur through choice of major or be independent from it. For example, students may be less likely to graduate overall as a result of missing the higher letter grade cutoff due to failing the class, losing a scholarship, or as the result of a poor GPA.

Results for six-year graduation rates are shown in Figure 4, and it does not appear that there is any effect on graduation for women. Results indicate there is no evidence of any effect on graduation for women. The figure for men is perhaps less clear, but I still do not find compelling evidence of a discontinuity.

4.4 STEM major

The main outcome of interest is whether students are more or less likely to graduate with a STEM major as the result of being just above or below a letter grade threshold. This result is plotted for men and women in Figure 5. This figure provides visual evidence that for any reasonable bandwidth selection or functional form, the estimated effect appears to be close to zero for both genders.

I formally estimate the effect of letter grade on propensity to major in STEM in Table 2. Estimates use a local linear regression and uniform kernel, with the optimal bandwidth determined according to procedures outlined by Calonico et al. (2014). Column (1) is the base specification with the optimal bandwidth. In column (2) I add instructor and term fixed effects, and introduce additional student-specific control variables in column (3). Because some variables are missing for some observations, I use mean substitution and also include a set of indicators for missing variables in order to include these observations in columns (3-6) and keep the sample consistently representative throughout the table. As one might expect, adding control variables does not result in significant changes to the point estimates.

In columns (4-6), I expand the bandwidth to verify that the estimates do not vary widely in a way that could indicate they depend on the choice of bandwidth. Since the estimates do not change dramatically with wider bandwidths, this does not appear to be of any concern. It is also worth noting that estimates are still not significant even with the added precision that wider bandwidths allow.

Estimates in Table 2 range from 1.8 to 3.0 percentage points for men and -0.8 to 1.5 p.p. for women, and none are statistically significantly different from zero. To interpret one estimate from column (2) as an example, a man achieving a higher letter grade at the margin in a core STEM course is 3.04 percentage points (4.2 percent) more likely to graduate with a STEM major within six years - although this is not statistically significant.[8]

While I do not find significant effects of letter grade on persistence in STEM majors for either men or women, I also find no evidence that results are different between genders. The point estimates for men are higher than for women across all columns of Table 2, but so are the baseline means. Looking again at column (2), the 95% confidence interval for men is (-1.3, 7.4) percentage points (-1.8%, 10.3%). This means the point estimate for women, 0.5 p.p. (0.9%), falls squarely within the confidence interval of the estimate for men.

4.5 Robustness

In this analysis, there might be some concern that the small number of instructors (sixteen in the main sample) means clustering in that dimension can lead to incorrect statistical inference. To address this possibility, I estimate wild bootstrap p-values and include them in Table 2 in the footer of each panel (Cameron, Gelbach, and Miller, 2008). For men the p-values range from .063 to .417 and only one is significant at the 10% level. The p-values of estimates for women are higher, ranging from .404-.998. It is clear from this analysis that the estimates are still not statistically significantly different from zero.

A second potential concern is that while instructors do look for a gap at which to draw

[8]I also verify that results are similar when using a 4 and 8 year graduation rate. Estimates range from 0.1 - 3.1 and -0.7 - 2.1 percentage points for men and women, respectively, and none are statistically significant.

grade cutoffs, the threshold may be drawn in part because of some student characteristic that is unobservable to the researcher. It is helpful to note that many of the classes in my sample have a large number of students and therefore it is less likely that instructors know individual students as well. Moreover, if the did this it would likely be limited to a student or two, which in this size sample is unlikely to generate bias of a meaningful magnitude. However, to address this issue, I perform a robustness exercise where I drop observations that are close to the threshold on either side, since these are the students who would be affected by such an issue. Specifically, I omit from the regression observations within 0.4, 0.8, 1.2, 1.6, and 2.0 percentage points of the threshold. Results are shown in Table A.3. Estimates range from -2.2 to 4.3 percentage points for men and -0.2 to 1.7 p.p. for women. Only one estimate is significant out of the 30 in this table, and that is a negative effect for men in column (7). I note this is inconsistent with all other estimates in this paper and even within this table, and given the number of coefficients is likely due to chance. Overall, estimates from this table are consistent with the main results in that neither men nor women are impacted by grade thresholds. This suggests that my main results are not affected by any issue of instructors selectively including or excluding certain students at letter grade thresholds due to unobservable characteristics.

I also show that my main results are robust using two alternative solutions to the missing variables in addition to mean substitution, although this only matters where controls are included since these are the only variables that are sometimes missing. First, I estimate effects using only the subsample of observations with no missing data in Table A.4. This subsample is obviously smaller with less statistical power, but estimates are roughly similar in magnitude and significance to the main results. However, a larger concern is that this may not be a representative subsample since it is not known whether variables are missing at random, even though that appears to be the case according to Figure A.1. Second, I also estimate results using multiple imputation in Table A.5. While estimates are again similar in magnitude, with the multiple imputation command I cannot compute standard

errors using two-way clustering as above. This table uses standard errors that are only clustered in one dimension by instructor. With larger bandwidths there are two statistically significant estimates for men that letter grades affect STEM participation by 3 p.p., but results in Figure 5 do not look visually compelling. Additionally, a check of my main results with one-way clustering suggests that this can lead to standard errors that are roughly 17% larger than two-way clustering, so p-values in Table A.5 may be too small and potentially misleading. Results for women are consistently close to zero regardless of how I handle the missing control variables.

4.6 Heterogeneity

While previous results show there is no effect on average for either men or women, it is possible effects could be nonzero for alternative cutoffs, subgroups, or outcomes. To address this, I first consider letter grade cutoffs separately rather than the stacked threshold. Results are shown in Figure A.3, with a wide range on the horizontal axis to show how estimates might look with different bandwidths. While most figures show no significant effect – including all of the graphs for women – there is some suggestive evidence that men are sensitive to letter grades at the C/D cutoff. Since this is the cutoff that represents the pass/fail mark for required-in-major courses at this university, it would make sense that this "hard" cutoff should matter more than others for mechanical reasons if not behavioral ones.

Corresponding estimates for each cutoff are shown in Table A.6. In column (8) of this table, using the same specification as the main results from above, I estimate that men earning a C versus D at the margin are 10.8 percentage points (20.2 percent) more likely to graduate with a STEM major within six years. I do also estimate an effect of 4.1 percentage points (4.7 percent) for men at the A/B cutoff, though the visual evidence appears to be much less compelling. Additionally, given the large number of estimates reported in Table A.6, I note that some will be statistically significant due to chance. Importantly, there is no evidence of an effect for women at any individual letter grade cutoff, as estimates range from

-3.5 to 4.2 p.p., none of which are significant. This suggests that women are not deterred from STEM by poor letter grades even at the crucial C/D threshold.

The null result also holds for a variety of subgroups where an effect seems most likely. I show results for these groups in Figure A.4 and Table 3, where the first column simply repeats the main specification from above for comparison. First, I look at courses taught by female and male instructors separately, since the interaction of instructor gender with student gender at letter grade thresholds might be more meaningful. I do not find compelling visual evidence of a discontinuity for men or women taught by either male or female instructors, although the estimate of 5.6 percentage points (8.5%) in column 2 is marginally significant for women in STEM courses taught by women. Estimates for men are 2.9 - 3.4 p.p. for both instructor genders, and -0.6 p.p. for women in classes taught by men, though none of these are significant.

Next, I consider only individuals who are first generation college students in their family, since this is typically a more vulnerable group which may be more sensitive to letter grade signals. Again, there is no compelling visual evidence that grades matter for this group. However, for male first generation students, there is marginal significance for the estimate that a higher grade makes them 4.5 percentage points (8.2%) more likely to major in a STEM field. The estimate for women here is -2.7 p.p but not significant. Again, I interpret even the marginally significant coefficient cautiously given the number of tests shown for different cutoffs and different groups.

I also consider freshman students who should have a lower cost to switch majors, and may therefore be more sensitive to letter grades. This subsample consists of a greater proportion of observations from biology and calculus classes,[9] with less representation from traditionally second-year courses such as organic chemistry or computer programming. The figure shows no evidence of an effect for either men or women, and the estimated results are not significant at 2.9 and -1.2 p.p. for men and women, respectively.

[9]About 90% of the observations in this subsample come from biology, calculus I, and calculus II.

The gender composition of economics majors and the factors affecting this is perhaps of particular interest to economists. While my previous analysis has focused on STEM courses and majors as defined by the Department of Homeland Security, I also collected data from both micro principles and intermediate microeconomics. For these courses, I show the data around the letter grade threshold in Figure A.5, with the outcome being graduating with an economics major (subfigure a) or business major (subfigure b). Business majors are interesting because many of them take economics courses but also because anecdotally it is fairly common for students to transfer into the business program if they achieve a high GPA i.e. good letter grades. This implies that higher letter grades in economics courses might actually lead to a lower probability of majoring in economics. However, I do not find visually compelling evidence that letter grades matter for either economics or business majors from these figures. Corresponding estimates are shown in Table A.7, most of which are not statistically different from zero. While one result is marginally significant for men having a business major in column 4 (6.8 percentage points or 22 percent), I do not consider this a meaningful result given the visual evidence and the fact that some estimates may appear significant by chance given the large number of estimates in this paper overall. Overall, the evidence indicates that letter grades in economics classes do not impact women's decisions to major in economics or business.

Finally, I consider the impact of letter grades for other subgroups on other majors as well as estimated annual earnings in Figure A.6 and Table A.8. In subfigure (a) and column (2), I expand the course pool to include economics classes and look at the effect of letter grade on the probability of graduating with a STEM or economics degree within six years. In subfigure (b)/column (3), I use only science courses including biology, calculus, and chemistry, and look at the impact on majoring in a science field. Then I use only engineering courses including calculus, computer programming, and statics and look at the impact on majoring in engineering in subfigure (c)/column (4).

Lastly, it is possible that students transfer to different majors that are less competitive

but still considered STEM fields, in which case I would estimate a zero with my main specification. In an attempt to address this concern, I match average annual earnings by major to my data (Carnevale et al., 2015).[10] In subfigure (d)/column (5), I use the STEM courses from the main sample (without economics classes) and look at the impact of letter grade thresholds on national median earnings by major (in 2013 dollars). Estimates in the table range from -3.3 to 2.9 percentage points and $1,156 for men and 0.4 to 4.1 p.p. and $500 for women but are largely not statistically significant. The only estimates significantly different from zero are for women in science courses majoring in science (4.1 p.p.), and for median earnings for men ($1,156). Visually, there may be a discontinuity for women in science, but I interpret this cautiously given the number of different results by subgroup shown in this section. There is no compelling visual evidence of an effect for earnings for men, nor for any of the other cases described in this exercise.

In summary, I show that by and large letter grades do not affect participation in STEM, even among most subgroups and at most grade thresholds. While there are some exceptions – for example, men at the A/B and C/D thresholds, women in courses taught by women, male first generation college students, or women in science courses majoring in science – most of these results are not visually compelling and need to be interpreted cautiously given multiple inference concerns. Thus, my overall conclusion is there is little compelling evidence that letter grade signals deter either men or women from STEM majors across a wide range of different grade thresholds, subgroups, and outcomes.

5 Discussion and Conclusion

In this paper, I provide the first causal evidence of the impact of letter grades on STEM major. My data contains 21,000 observations from seven distinct, core STEM courses at a large university and includes final numerical as well as letter grade. This allows me to

[10] It is difficult to match less common majors that may not exist at this university or do not exist in national average income estimates.

utilize a regression discontinuity at letter grade thresholds. The identifying assumption of this design is that all other determinants of college major besides letter grade vary smoothly through grade cutoffs. With this in mind while collecting data, instructors were interviewed individually and each stated that they did not allow manipulation across grade thresholds. Additionally, a rich set of covariates including those representing prior academic ability allows me to test for and find evidence in support of the identifying assumption.

Results indicate that negative letter grade signals do not contribute to the gender gap in STEM. Neither men nor women are deterred by poor letter grades relative to their peers on the other side of the cutoff who are otherwise similar. These results are robust to a variety of alternative specifications. Furthermore, these results are relatively precise. Using baseline means and point estimates from column (2) in Table 2, I can rule out that a higher letter grade affects STEM participation for women by more than 6.0 percent (3.2 percentage points).

To provide some context for the estimate magnitude my results enable me to rule out, consider related research. However, many of these papers have very low baseline means which can lead to dramatic percent effects with the small denominator. For this reason, I present both the point estimate and percent effect. Owen (2010) estimates that women who received an A at the margin in principles of economics courses are 15 percentage points ($\sim$167%) more likely to major in economics. This is nearly five times higher than the maximum effect I can rule out for generalized STEM courses in percentage point terms, and even higher as a percent. Even when I restrict my sample to only consider economics courses and majors, I rule out an effect bigger than 6.2 percentage points (9.8%) for women, which is still significantly smaller than the effect Owen finds.

In comparison to my main result again, Carrell et al. (2010) look at women with above-average SAT math scores in introductory STEM courses with a female instructor. They find these women are 15.5 percentage points ($\sim$40%) more likely to graduate with a STEM degree when their math and science courses are taught by all female faculty. Therefore I can reject

the null hypothesis that the impact of grade signals on women in STEM is equal to the effect of having all female instructors on women with above average math scores. I can also rule out that letter grades matter on the same magnitude as the impact of exposing students to a female role model. Porter and Serra (2020) study what happens when a "charismatic career woman who major in economics at the same university" talks to principles of economics classes and found that the fifteen minute visits increased the likelihood of women majoring in economics by 8 percentage points or a roughly 100 percent increase from the baseline mean.

In this paper, I also look at effects across various types of subgroups. These include female or male instructors separately, first generation college students, and freshman. I then consider the impact of grades in economics courses on economics or business majors, science courses on science majors, engineering courses on majoring in engineering, as well as the effect of letter grade thresholds on estimated annual earnings. I find no compelling evidence that letter grades matter in these cases, except perhaps for suggestive evidence for women in science courses. I also look at each letter grade threshold alone, and find no effect for women even at the C/D cutoff which represents the difference between passing and failing within a STEM major at this university.

Overall, these results suggest that negative grade signals - including those at the most important grade thresholds and within the most susceptible groups - are not responsible for the large gender gap in STEM. This highlights the importance of understanding other factors that are important in the gender gap in STEM majors and the corresponding gender gap in earnings.

References

Bureau of Labor Statistics. Women in architecture and engineering occupations in 2016, 2017. URL https://www.bls.gov/opub/ted/2017/women-in-architecture-and-engineering-occupations-in-2016.htm.

Bureau of Labor Statistics. Occupational employment statistics, 2019. URL https://www.bls.gov/oes/current/oes_stru.htm.

Sebastian Calonico, Matias D Cattaneo, and Rocio Titiunik. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326, 2014.

A Colin Cameron, Jonah B Gelbach, and Douglas L Miller. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427, 2008.

David Card and A Abigail Payne. High school choices and the gender gap in stem. *Economic Inquiry*, 59(1):9–28, 2021.

Anthony P Carnevale, Ban Cheah, and Andrew R Hanson. The economic value of college majors. 2015.

Scott E. Carrell, Marianne E. Page, and James E. West. Sex and science: How professor gender perpetuates the gender gap. *Quarterly Journal of Economics*, 125(3):1101–1144, 2010. ISSN 00335533. doi: 10.1162/qjec.2010.125.3.1101.

Kayal R Fontenot, Jessica L Semega, and Melissa A Kollar. US Census Bureau, Current Population Reports, P60-259, Income and Poverty in the United States: 2016. *US Government Printing Office, Washington, DC*, (September 2018), 2017.

Claudia Goldin. Gender and the undergraduate economics major: Notes on the undergraduate economics major at a highly selective liberal arts college. *manuscript, April*, 12, 2015.

Claudia Goldin, Sari Pekkala Kerr, Claudia Olivetti, and Erling Barth. The expanding gender earnings gap: Evidence from the LEHD-2000 census. *American Economic Review*, 107(5):110–114, 2017. ISSN 00028282. doi: 10.1257/aer.p20171065.

Maria Johnson and Vicki S. Helgeson. Sex differences in response to evaluative feedback: A field study. *Psychology of Women Quarterly*, 26(3):242–251, 2002. ISSN 03616843. doi: 10.1111/1471-6402.00063.

Adriana D. Kugler, Catherine H. Tinsley, and Olga Ukhaneva. Choice of Majors: Are Women Really Different from Men? *NBER Working Paper No. 23735*, page 38, 2017. URL http://www.nber.org/papers/w23735.

Joyce B. Main and Ben Ost. 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, 2014. ISSN 00220485. doi: 10.1080/00220485.2014.859953.

Margarita Mayo, Maria Kakarika, Juan Carlos Pastor, and Stéphane Brutus. Aligning or inflating your leadership self-image? A longitudinal study of responses to peer feedback in MBA teams. *Academy of Management Learning and Education*, 11(4):631–652, 2012. ISSN 1537260X. doi: 10.5465/amle.2010.0069.

National Center for Education Statistics. Status and trends in the education of racial and ethnic groups, indicator 26: Stem degrees, 2019. URL `https://nces.ed.gov/programs/raceindicators/indicator_reg.asp`.

Muriel Niederle and Lise Vesterlund. Explaining the gender gap in math test scores: The role of competition. *Journal of Economic Perspectives*, 24(2):129–44, 2010.

Rose E O'Dea, Malgorzata Lagisz, Michael D Jennions, and Shinichi Nakagawa. Gender differences in individual variation in academic grades fail to fit expected patterns for stem. *Nature communications*, 9(1):1–8, 2018.

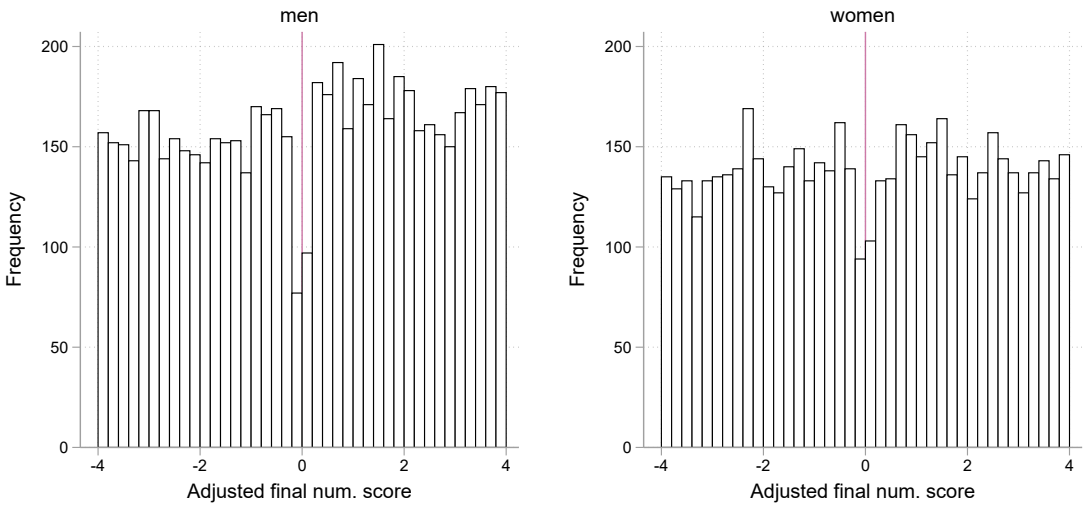
Ann L. Owen. Grades, gender, and encouragement: A regression discontinuity analysis. *Journal of Economic Education*, 41(3):217–234, 2010. ISSN 00220485. doi: 10.1080/00220485.2010.486718.

Cristian Pop-Eleches and Miguel Urquiola. Going to a better school: Effects and behavioral responses. *American Economic Review*, 103(4):1289–1324, 2013.

Catherine Porter and Danila Serra. Gender differences in the choice of major: The importance of female role models. *American Economic Journal: Applied Economics*, 12(3):226–54, 2020.

Jamin Speer. Bye bye ms. american sci: Women and the leaky stem pipeline. 2021.

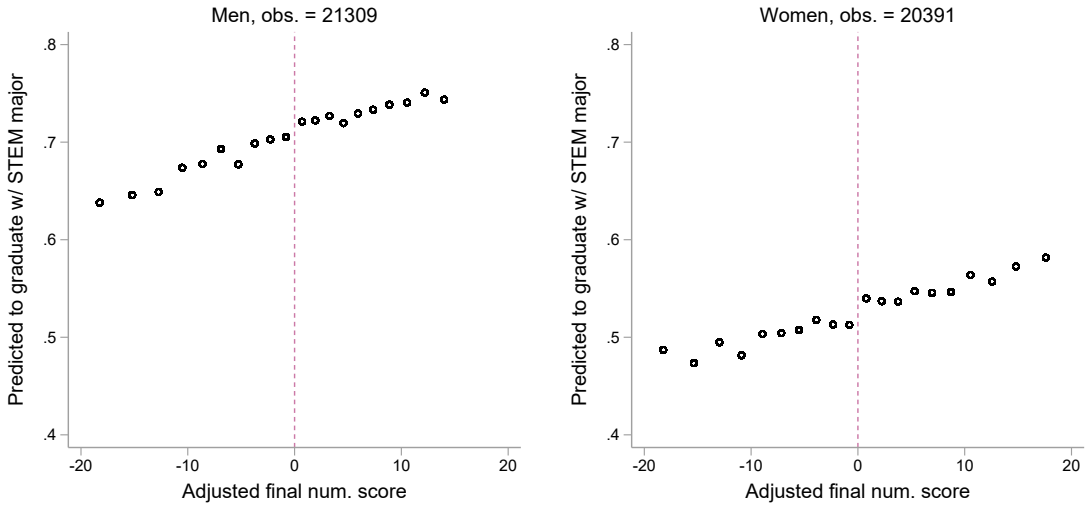
Figure 1: Observations Histogram



This figure shows the frequency (count) of observations close to the stacked regression discontinuity threshold for men and women in the sample STEM courses. The figure shows no unusual heaping of observations just above the threshold for either men or women, which if present could be a concern. There is a dip in observations very close to the threshold, which is consistent with what we know about instructors looking for a natural gap in the grading distribution at which to draw letter grade cutoffs.

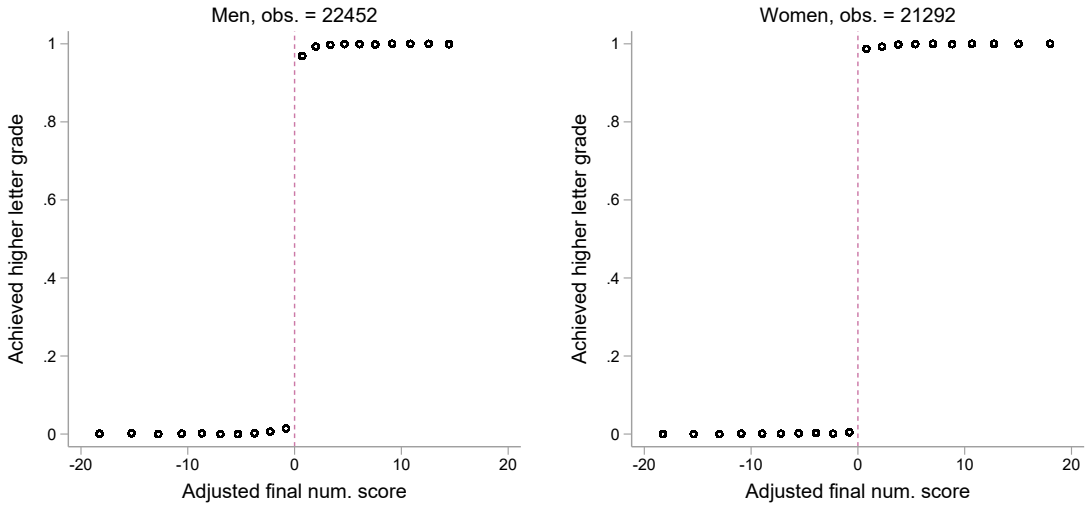
Figure 2: Test of Identifying Assumption: Predicted Outcome

Predicted to graduate with a STEM major within six years



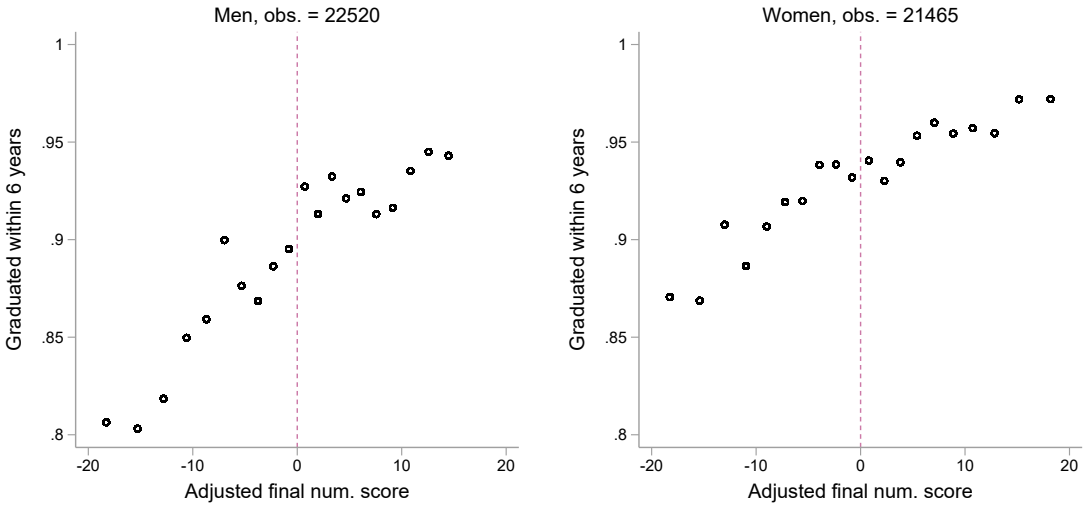
This figure shows predicted outcome for the main outcome of interest, graduated with a STEM major within six years. As a test of the identifying assumption, I predict the main outcome using only pre-determined characteristics. This predicted outcome varies smoothly through the stacked letter grade cutoff for both men and women, in support of the identifying assumption. Each point represents an equal number of observations.

Figure 3: Effect of Number Grade Crossing Threshold on Receiving Higher Letter Grade



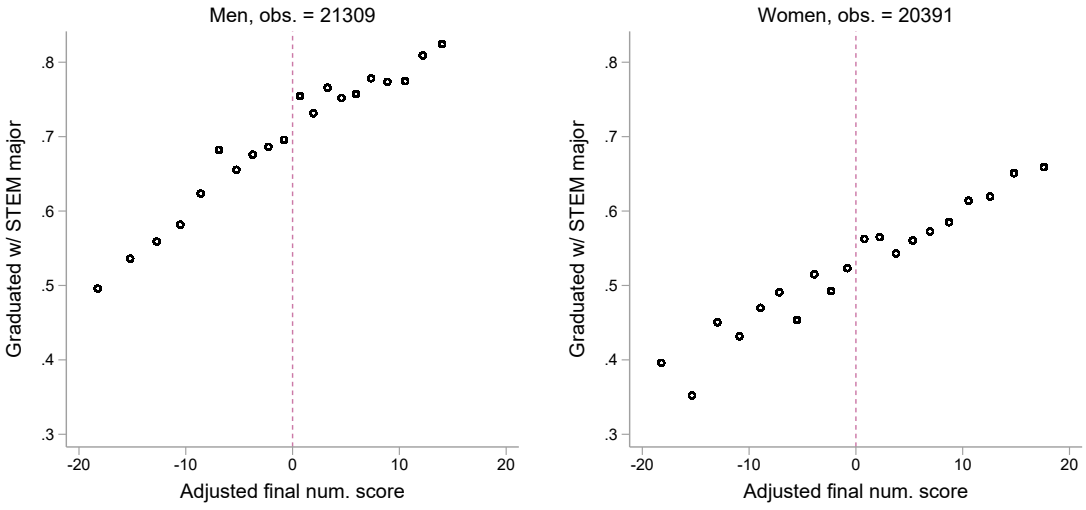
This figure shows effect of final numerical grade threshold crossing on final letter grade. Each point represents an equal number of observations.

Figure 4: Effect of Letter Grade Thresholds on Six-year Graduation Rate



This figure shows six-year graduation rate for students in the main sample. Graduation is an important outcome by itself, but especially so with these data because I only observe major for students who graduated. Visually there does not appear to be any significant effect of letter grade thresholds on the probability of graduation for either gender. Each point represents an equal number of observations.

Figure 5: Effect of Letter Grade Thresholds on Graduating With a STEM Major



This figure shows the main outcome of interest, graduated with a STEM major within six years. The point of this figure is to show the data and what treatment effect might exist, if any, for different bandwidths or functional forms. It is clear that for any reasonable bandwidth, there is no effect for either men or women. I estimate effects formally in Table 2. Each point represents an equal number of observations.

Table 1: Summary Statistics

	All students	Men	Women
female	0.46	0.00	1.00
HS stu. rank	56.39	64.64	48.19
max(SAT score, ACT conc. score)	1223.05	1249.63	1198.56
prior gpa	2.37	2.43	2.34
transfer hours	24.80	25.04	24.71
appl. fin. aid	0.58	0.58	0.57
1st gen. stu.	0.22	0.21	0.22
freshman	0.24	0.22	0.25
sophomore	0.44	0.44	0.44
transfer stu.	0.07	0.09	0.06
intl. stu.	0.02	0.02	0.01
grad. in 6 yrs	0.88	0.86	0.90
grad. w/ STEM major	0.58	0.66	0.50
Observations	21,533	10,790	9,188

This table shows summary statistics for all students in the main sample, and also broken down for men and women separately. Because many students do not take both the SAT and ACT, I use official concordance tables to convert ACT scores to comparable SAT scores, then take the max of SAT score and ACT concordance score where both exist. (See Table A.1 for more information on missing variables.)

Table 2: Effect of Letter Grade Thresholds on Graduating With a STEM Major

Panel A: Men	(1)	(2)	(3)	(4)	(5)	(6)
Above letter grade cutoff	0.0270	0.0304	0.0260	0.0300	0.0296	0.0177
	(0.0298)	(0.0224)	(0.0272)	(0.0221)	(0.0171)	(0.0144)
Observations	8,654	8,654	8,654	10,698	12,586	15,965
Outcome mean	.717	.717	.717	.719	.718	.713
Wild bootstrap p-value	.417	.232	.284	.122	.0627	.125
Instructor & term FEs		Y	Y	Y	Y	Y
Other control vars.			Y	Y	Y	Y
Opt. Bandwidth = 5.72	1x	1x	1x	1.25x	1.5x	2x
Panel B: Women						
Above letter grade cutoff	0.0154	0.0045	-0.0001	0.0032	0.0020	-0.0078
	(0.0174)	(0.0146)	(0.0158)	(0.0141)	(0.0104)	(0.0097)
Observations	10,078	10,078	10,078	12,246	14,267	17,778
Outcome mean	.528	.528	.528	.528	.529	.529
Wild bootstrap p-value	.433	.771	.998	.809	.823	.404
Instructor & term FEs		Y	Y	Y	Y	Y
Other control vars.			Y	Y	Y	Y
Opt. Bandwidth = 7.80	1x	1x	1x	1.25x	1.5x	2x

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

This table shows estimates of the effect of letter grade on the main outcome of interest, graduation with a STEM major within six years. Column (1) is the base specification using 1x the optimal bandwidth. Column (2) includes instructor and term fixed effects. Column (3) further includes available student-specific control variables. In columns (4-6), I increase the bandwidth. Estimated with a local linear regression and uniform kernel. Standard errors are in parenthesis and clustered at the instructor and term level. I also include wild bootstrap p-values in the footer of each panel since the number of instructors is low (sixteen), but results are similar.

Table 3: Results for Subgroups Where Effect Seems Most Likely

Outcome: Graduated with a STEM major

Panel A: Men	(1)	(2)	(3)	(4)	(5)
	Main result	Female instr	Male instr	First gen	Freshmen
Above letter grade cutoff	0.0304	0.0344	0.0292	0.0454*	0.0292
	(0.0222)	(0.0237)	(0.0228)	(0.0254)	(0.0270)
Observations	8,654	4,289	10,904	3,775	3,185
Outcome mean	.717	.813	.498	.556	.567
Instructor & term FEs	Y	Y	Y	Y	Y
Opt. Bandwidth =	5.72	9.35	6.97	9.58	10.6
Panel B: Women					
Above letter grade cutoff	0.0045	0.0556*	-0.0064	-0.0270	-0.0124
	(0.0139)	(0.0273)	(0.0138)	(0.0170)	(0.0173)
Observations	10,078	1,834	8,568	2,227	3,239
Outcome mean	.528	.653	.393	.405	.401
Instructor & term FEs	Y	Y	Y	Y	Y
Opt. Bandwidth =	7.8	7.49	5.79	5.97	10.74

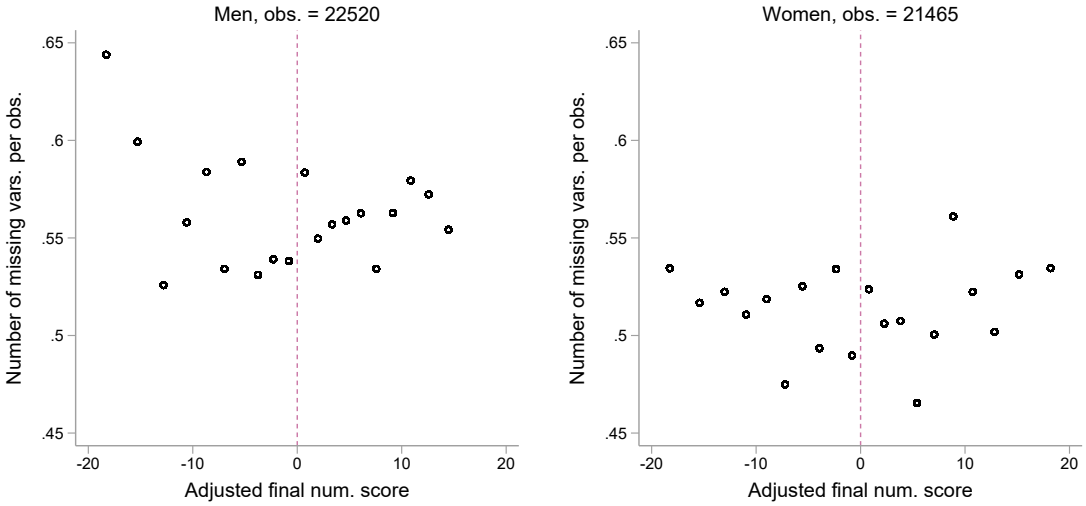
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

This table shows estimates for alternative samples where an effect seems most likely. Namely I look at only courses taught by female and female instructors (columns 2 and 3, respectively), and only students that are first generation or freshmen (columns 4 and 5). I plot these data in Figure A.4. All models include instructor and term fixed effects. Estimated with a local linear regression and uniform kernel. Standard errors are in parenthesis and clustered at the instructor and term level.

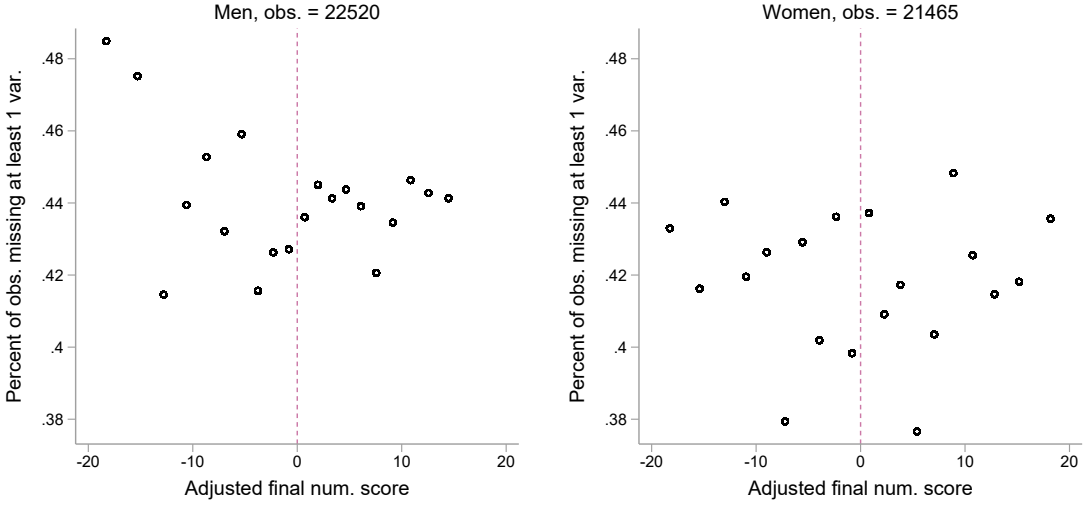
A Appendix

Figure A.1: Missing Variables Analysis

(a) Total number of missing variables for each observation



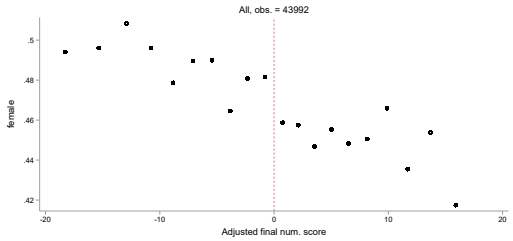
(b) Percent of observations missing at least one variable



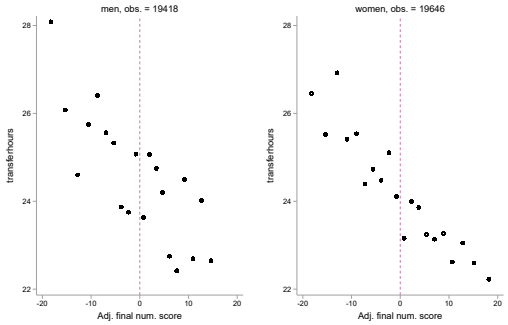
In these data, some observations are missing some variables. To address concerns that these missing characteristics are non-random and could bias those parts of my analysis that include these control variables, I present these two figures. Figure A.1a shows the total number of missing variables per observation and Figure A.1b shows the number of observations missing at least one variable. Both figures show that these measures of missing variables appear smooth through the threshold. Each point represents an equal number of observations.

Figure A.2: Verifying Covariates Are Smooth Through the Letter Grade Threshold

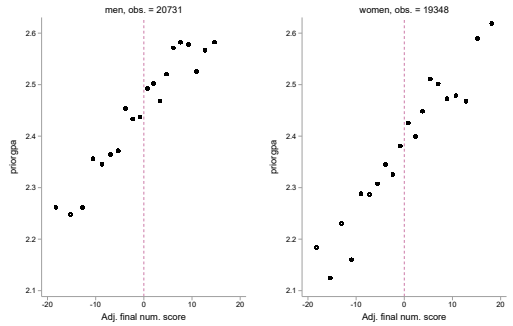
(a) Student Gender



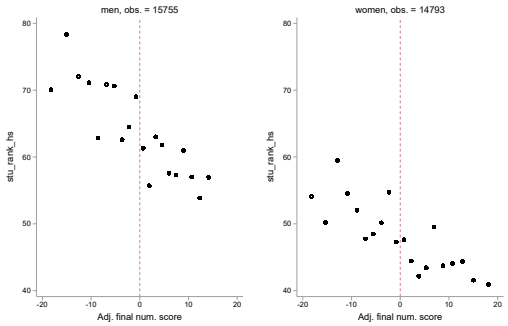
(b) Transfer Hours



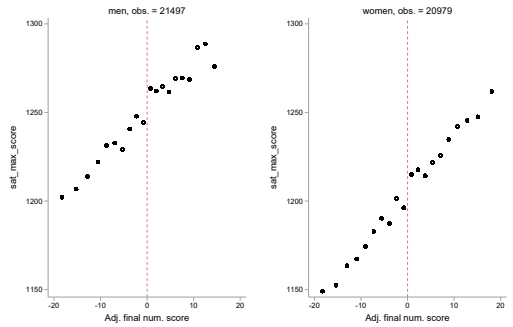
(c) Prior College GPA



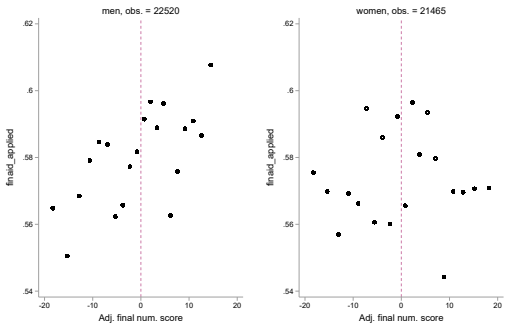
(d) High School Rank



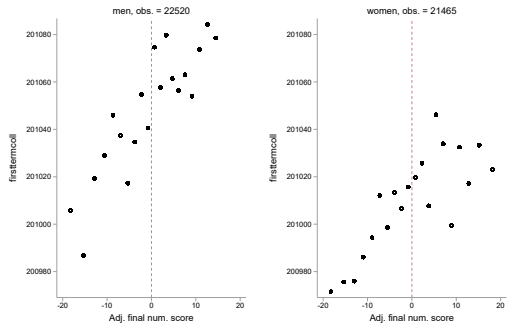
(e) max{SAT, ACT concordance}



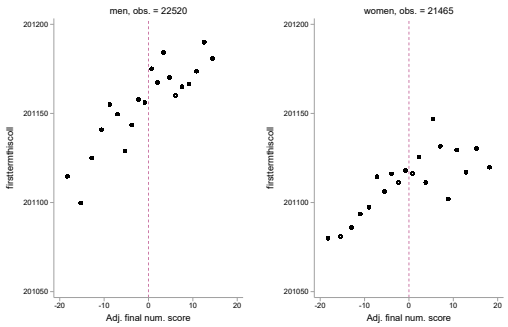
(f) Applied for Financial Aid



(g) First Term Any College

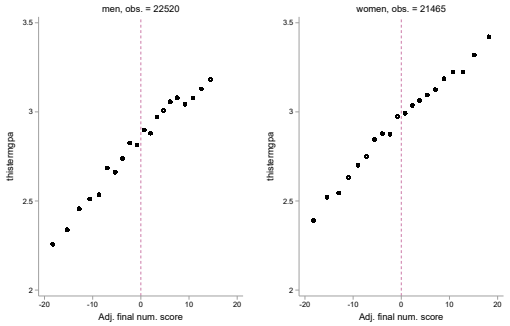


(h) First Term This College

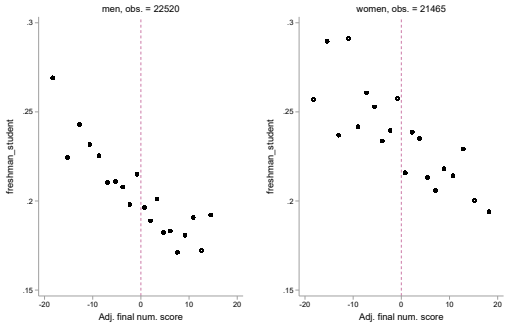


Continued on next page.

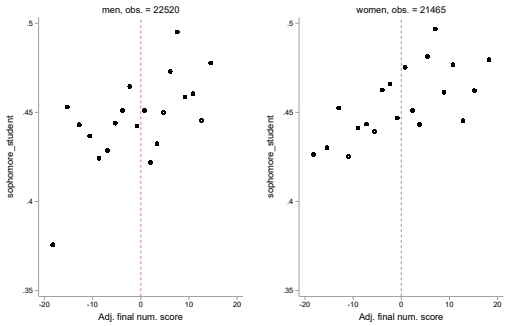
(i) Semester GPA Excl. Sample Courses



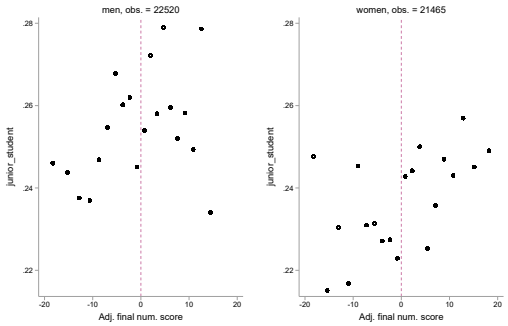
(j) Freshman



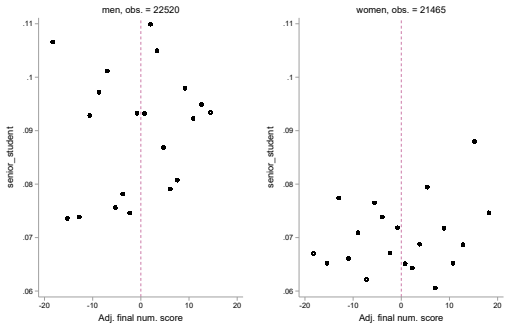
(k) Sophomore



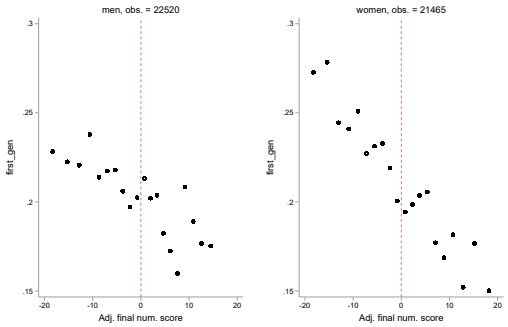
(l) Junior



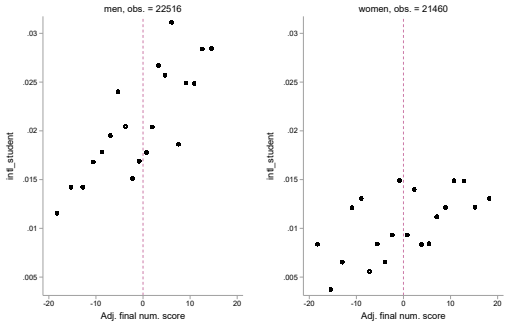
(m) Senior



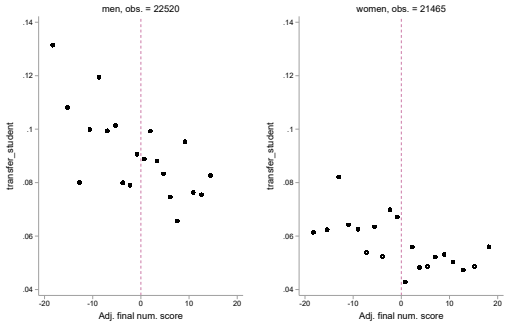
(n) First Generation Student



(o) International Student



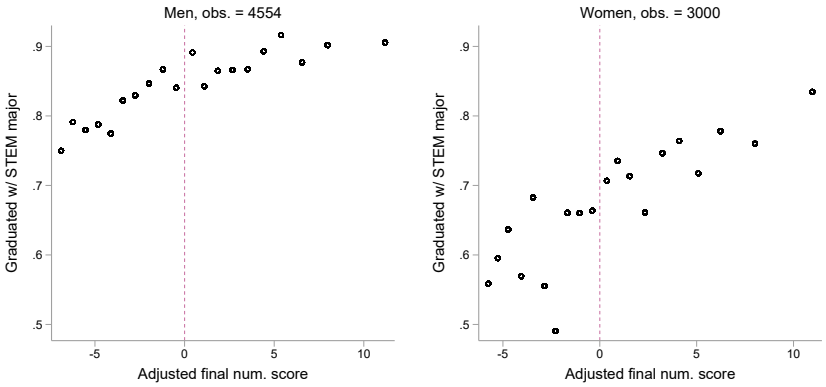
(p) Transfer Student



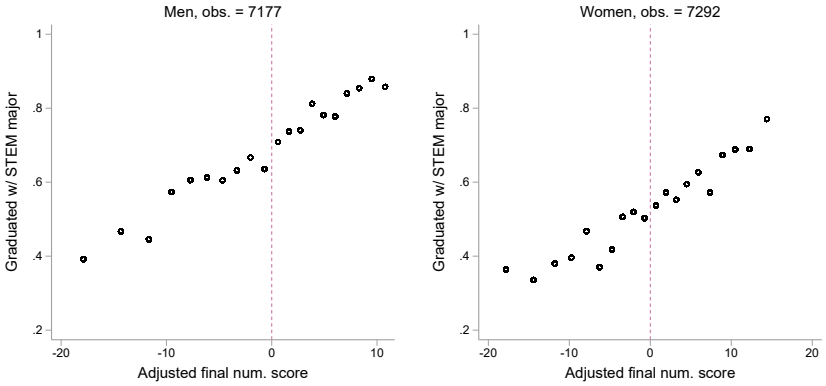
These figures show that all other variables appear smooth through the threshold, which supports the identifying assumption of the regression discontunity. I estimate each variable in Table A.2. Each point represents an equal number of observations.

Figure A.3: Effect at each letter grade cutoff

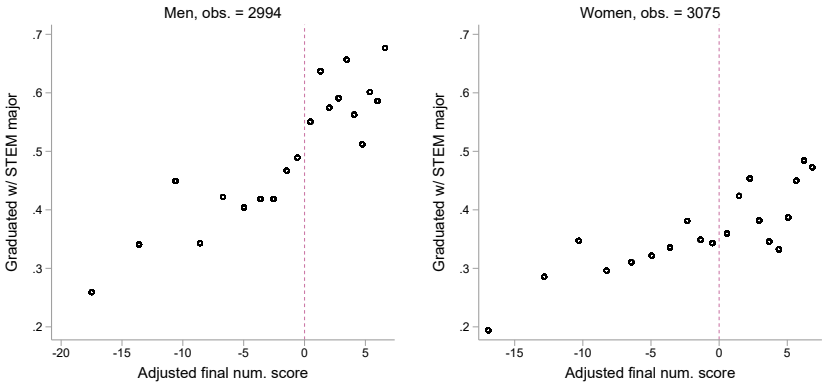
(a) Effect of A/B cutoff on graduation with a STEM major



(b) Effect of B/C cutoff on graduation with a STEM major



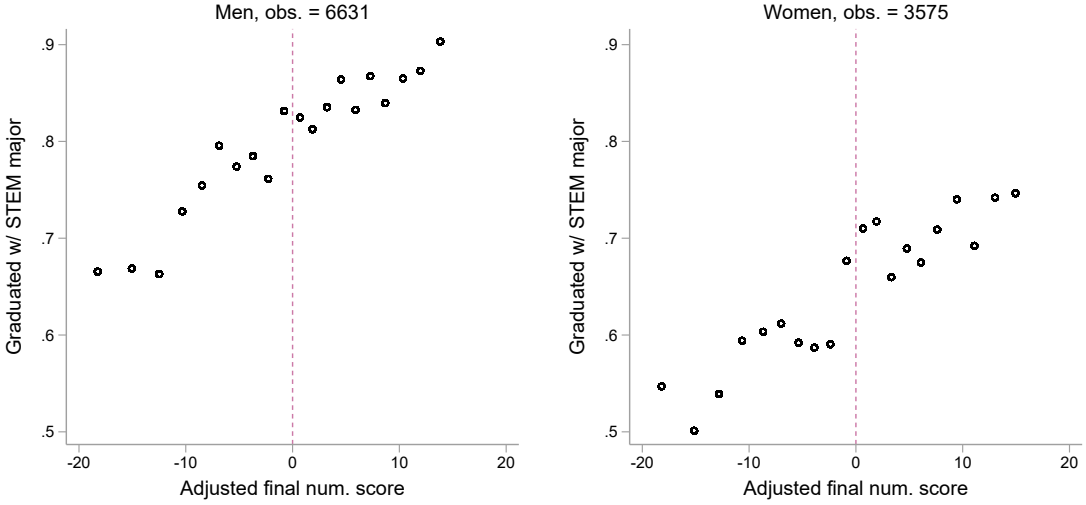
(c) Effect of C/D cutoff on graduation with a STEM major



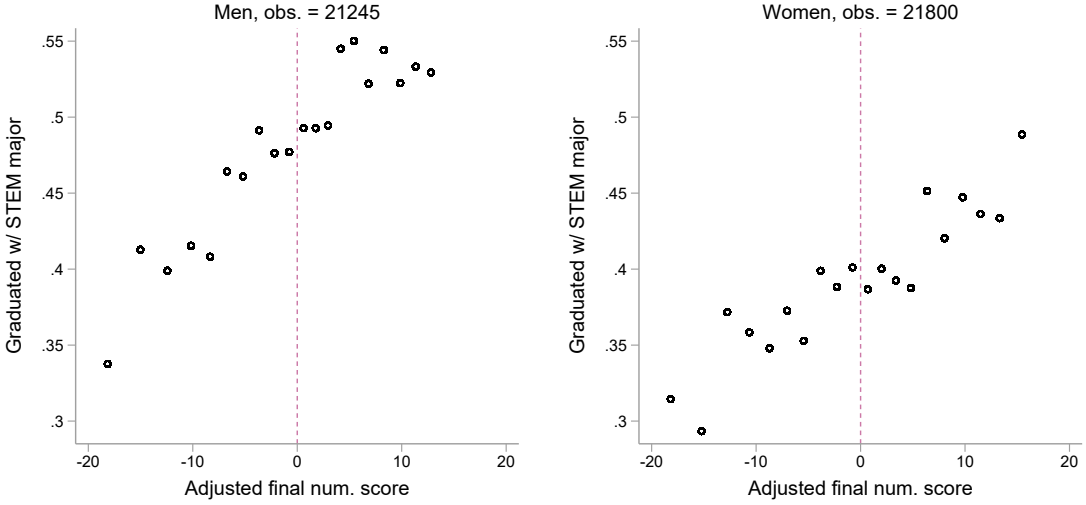
These figures show the main outcome, graduated with a STEM major within six years, for each letter grade cutoff separately. While most figures show no significant effect, there is some suggestive evidence of a discontunity for men at the C/D cutoff. Each point represents an equal number of observations.

Figure A.4: Subgroups Where Effect Seems Most Likely

(a) Only courses taught by female instructor

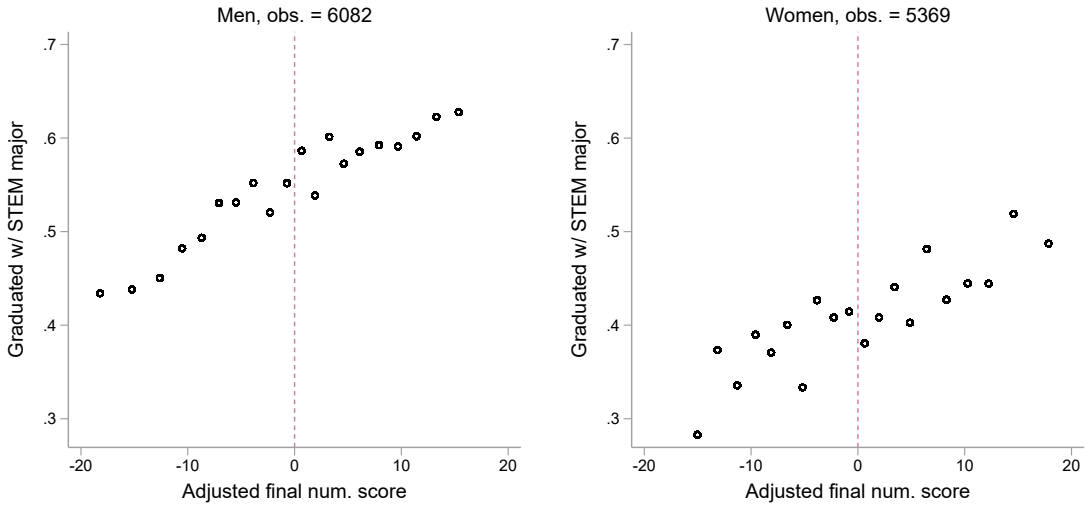


(b) Only courses taught by male instructor

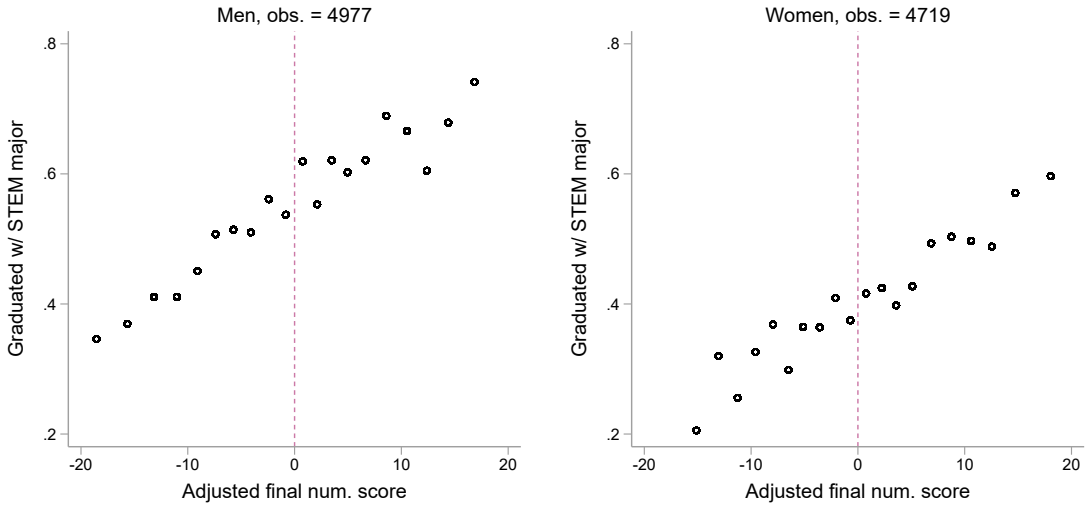


Continued on next page.

(c) Only first generation students



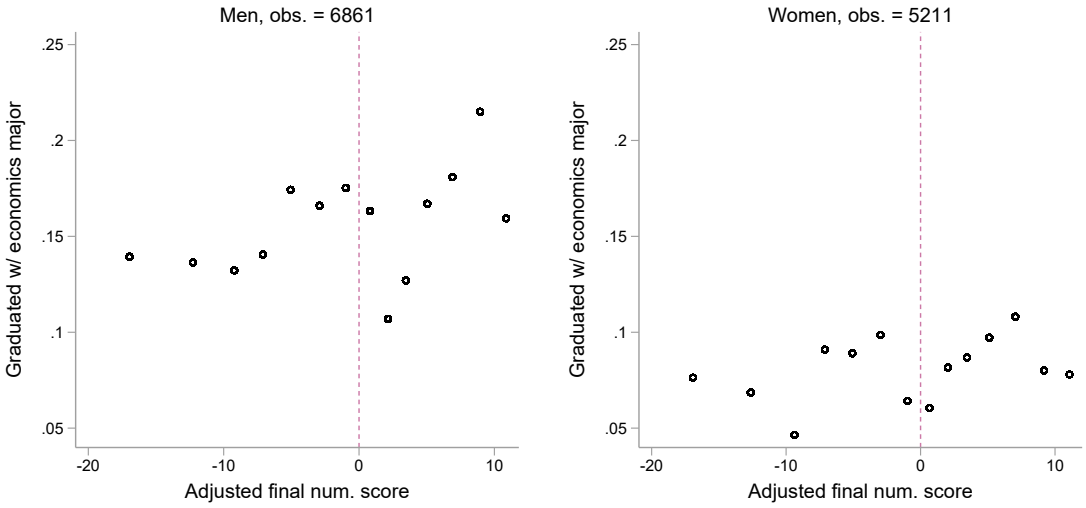
(d) Only freshmen students



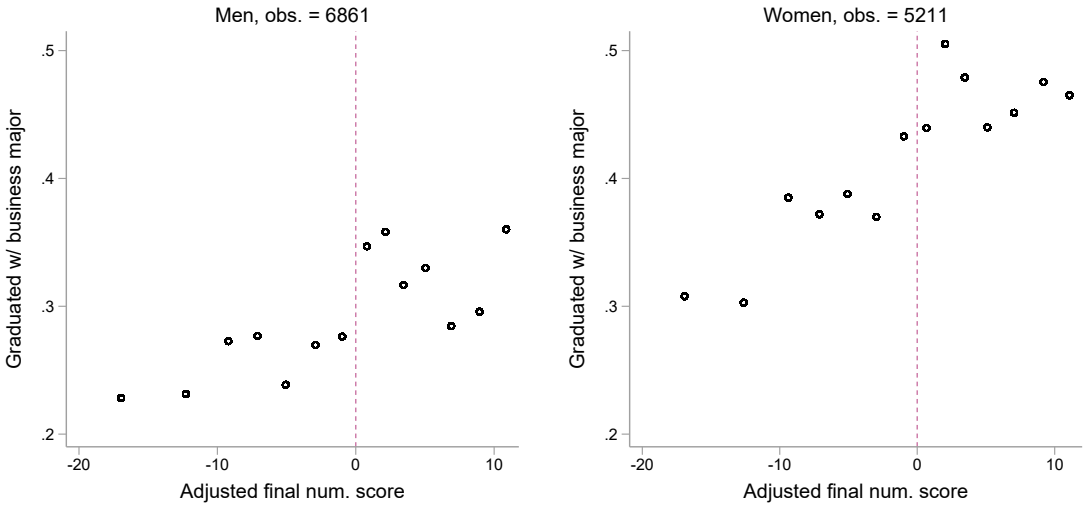
These figures show data around the threshold for subgroups where an effect seems likely. I estimate these regressions in Table 3. Each point represents an equal number of observations.

Figure A.5: Effect of letter grades in economics courses

(a) Effect on graduation with an econonmics major



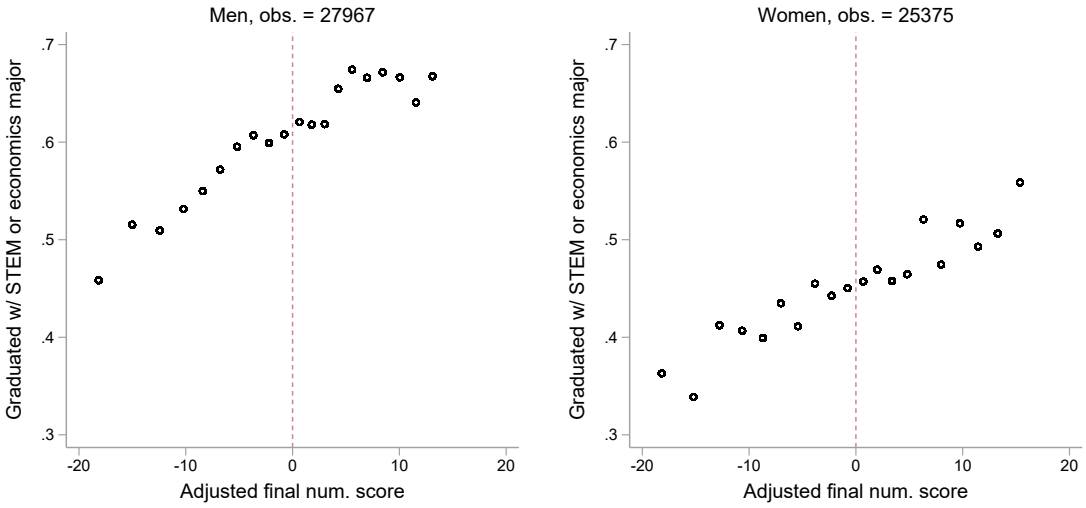
(b) Effect on graduation with a business major



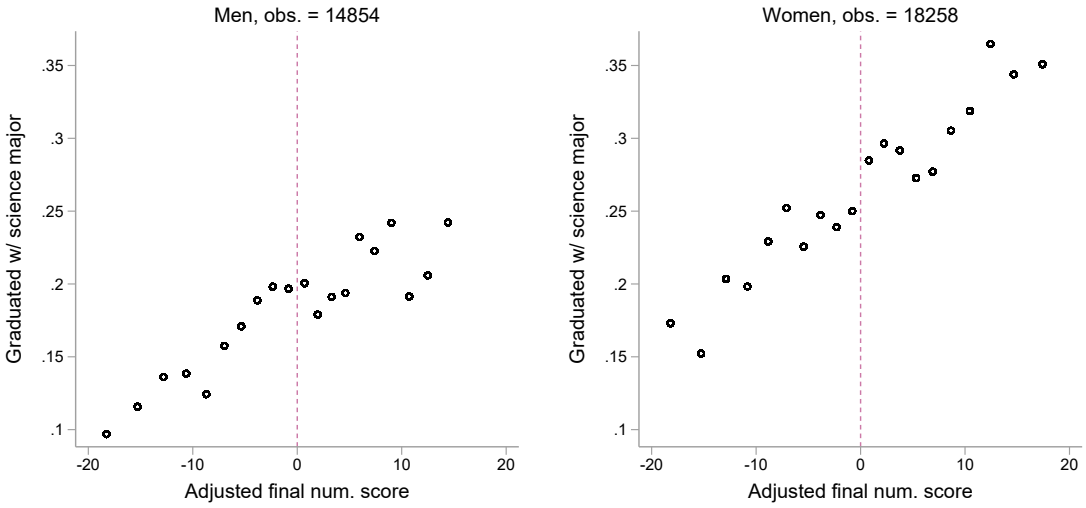
These figures show data using only economics courses, and look at the outcome of majorng in economics and business (subfigures (a) and (b), respectively). I estimate these effects in Table A.7. Each point represents an equal number of observations.

Figure A.6: Alternate subgroups of courses, and alternate outcomes

(a) All courses including economics and effect on STEM or economics major

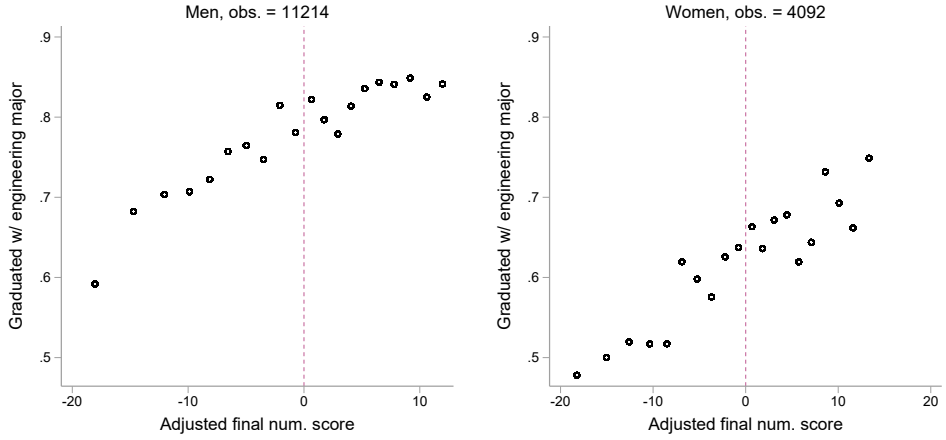


(b) Only science courses (biology, calculus, organic chemistry) and effect on science major

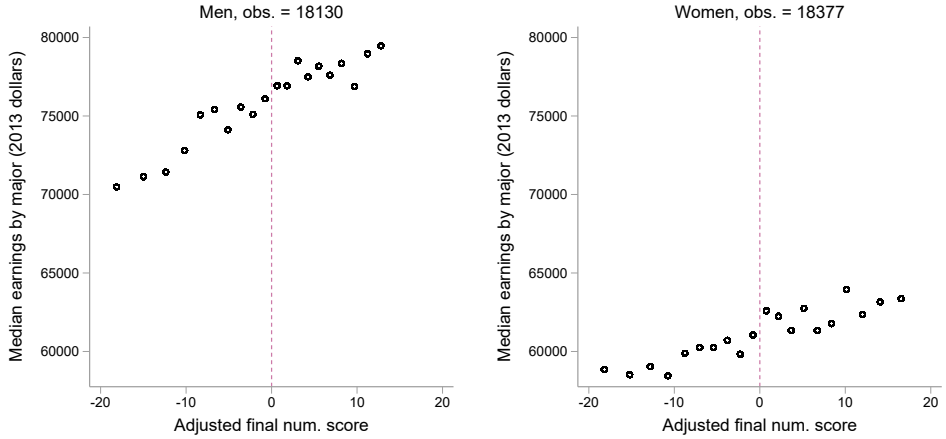


Continued on next page.

(c) Only engineering courses (calculus, computer programming, and statics) and effect on engr. major



(d) Main sample STEM courses and effect on estimated median earnings



These figures show data around the letter grade threshold for alternative subgroups and outcomes. I estimate these regressions in Table A.8. Each point represents an equal number of observations.

Table A.1: Missing variable statistics

	All students	Men	Women
avg. num. missing vars	0.63	0.58	0.54
missing >= 1 var.	0.49	0.45	0.44
female	0.07	0.00	0.00
HS stu. rank	0.30	0.30	0.32
SAT score	0.14	0.15	0.13
ACT score	0.46	0.48	0.45
max(SAT score, ACT conc. score)	0.03	0.04	0.02
prior gpa	0.10	0.08	0.10
transfer hours	0.12	0.15	0.09
intl. stu.	0.00	0.00	0.00
Observations	21,533	10,790	9,188

This table shows the average number of missing variables per observation (first row), the percent of observations missing at least one variable (second row), and the percent of observations missing each variable (remaining rows). Because many students do not take both the SAT and ACT, I use official concordance tables to convert ACT scores to comparable SAT scores, then take the max of SAT score and ACT concordance score where both exist. The resulting variable is only missing for 3% of the observations in the sample.

Table A.2: Test of identifying assumption: regression discontinuity estimates for covariates

Panel A: All students	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
	female	transfer hours	prior gpa	HS stu. rank	max(SAT score, ACT conc. score)	appl. fin. aid	1st term any coll.	1st term this coll.	GPA excl. sample courses	freshman	sophomore	junior	senior	1st gen. stu.	intl. stu.	transfer stu.
Above letter grade cutoff	-0.0288*	-0.3731	0.0184	-3.4359	9.1586**	0.0005	14.9416	9.4970	-0.0356	-0.0272*	0.0093*	0.0140	0.0029	0.0096	-0.0007	-0.0018
	(0.0152)	(0.4318)	(0.0566)	(3.8152)	(3.9876)	(0.0187)	(13.4645)	(14.1628)	(0.0245)	(0.0129)	(0.0045)	(0.0110)	(0.0090)	(0.0125)	(0.0047)	(0.0096)
Observations	14,997	14,444	14,739	11,575	15,757	16,273	16,273	16,273	16,273	16,273	16,273	16,273	16,273	16,273	16,270	16,273
Variable mean	.464	24.1	2.42	55	1,228	.587	201,046	201,150	2.91	.222	.457	.241	.0779	.207	.0151	.0691
Bandwidth = 5.08	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x
Panel B: Men																
		transfer hours	prior gpa	HS stu. rank	max(SAT score, ACT conc. score)	appl. fin. aid	1st term any coll.	1st term this coll.	GPA excl. sample courses	freshman	sophomore	junior	senior	1st gen. stu.	intl. stu.	transfer stu.
Above letter grade cutoff		0.0433	0.0032	-6.8439	9.1764	0.0194	16.3395	7.8851	-0.0331	-0.0084	-0.0276	0.0166	0.0174	0.0227	0.0028	0.0190
		(0.6826)	(0.0294)	(4.9951)	(8.0086)	(0.0210)	(12.6329)	(14.7383)	(0.0261)	(0.0178)	(0.0178)	(0.0103)	(0.0124)	(0.0149)	(0.0066)	(0.0138)
Observations		7,716	8,223	6,378	8,577	8,970	8,970	8,970	8,970	8,970	8,970	8,970	8,970	8,970	8,967	8,970
Variable mean		24.3	2.46	63.5	1,253	.583	201,053	201,160	2.86	.2	.448	.259	.0882	.202	.0212	.0878
Bandwidth = 5.70		1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x
Panel C: Women																
		transfer hours	prior gpa	HS stu. rank	max(SAT score, ACT conc. score)	appl. fin. aid	1st term any coll.	1st term this coll.	GPA excl. sample courses	freshman	sophomore	junior	senior	1st gen. stu.	intl. stu.	transfer stu.
Above letter grade cutoff		-0.5637	-0.0123	-4.5592	4.1412	0.0024	0.1671	-3.0615	-0.0313	-0.0089	-0.0140	0.0270***	-0.0052	0.0028	-0.0044	-0.0221**
		(0.4406)	(0.0671)	(3.0534)	(3.5846)	(0.0098)	(11.6287)	(9.6781)	(0.0449)	(0.0131)	(0.0114)	(0.0059)	(0.0081)	(0.0163)	(0.0035)	(0.0081)
Observations		9,635	9,505	7,341	10,286	10,510	10,510	10,510	10,510	10,510	10,510	10,510	10,510	10,510	10,508	10,510
Variable mean		24	2.39	47.5	1,206	.581	201,018	201,120	2.97	.234	.462	.233	.069	.209	.00971	.0559
Bandwidth = 7.82		1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

This table shows estimated coefficients for each observable student characteristic in the data. As a test of the identifying assumption, we expect these estimate to be close to zero to show that student characteristics vary smoothly across the letter grade threshold. I plot these data in Figure A.2. Out of all variables and panels, six estmates are significant, which is roughly what we'd expect to see due to random chance. Estimated with a local linear regression and uniform kernel, with bandwidths determing using the same method as for the main results, so that the bandwidths match for men and women. Standard errors are in parenthesis and clustered at the instructor and term level.

Table A.3: "Donut" RD: Graduated with a STEM major

Omitting X percentage points of number grade on either side of threshold

Panel A: Men	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
	(0.4)	(0.4)	(0.4)	(0.8)	(0.8)	(0.8)	(1.2)	(1.2)	(1.2)	(1.6)	(1.6)	(1.6)	(2.0)	(2.0)	(2.0)
Above letter grade cutoff	0.0430	0.0355	0.0329	0.0158	0.0143	0.0132	-0.0224***	-0.0163	-0.0184	0.0050	0.0095	0.0087	-0.0028	-0.0036	-0.0002
	(0.0261)	(0.0221)	(0.0250)	(0.0109)	(0.0139)	(0.0205)	(0.0045)	(0.0172)	(0.0211)	(0.0278)	(0.0263)	(0.0304)	(0.0241)	(0.0219)	(0.0312)
Observations	8,167	8,167	8,167	7,491	7,491	7,491	6,867	6,867	6,867	6,212	6,212	6,212	5,606	5,606	5,606
Outcome mean	.717	.717	.717	.718	.718	.718	.717	.717	.717	.714	.714	.714	.709	.709	.709
Instructor & term FEs		Y	Y		Y	Y		Y	Y		Y	Y		Y	Y
Other control vars.			Y			Y			Y			Y			Y
Opt. Bandwidth = 5.72	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x
Panel B: Women															
Above letter grade cutoff	0.0122	0.0018	-0.0019	0.0164	0.0068	0.0030	0.0095	0.0075	0.0026	0.0168	0.0147	0.0107	0.0108	0.0095	0.0044
	(0.0198)	(0.0160)	(0.0182)	(0.0209)	(0.0164)	(0.0196)	(0.0189)	(0.0191)	(0.0226)	(0.0304)	(0.0281)	(0.0316)	(0.0324)	(0.0330)	(0.0359)
Observations	9,627	9,627	9,627	9,049	9,049	9,049	8,505	8,505	8,505	7,924	7,924	7,924	7,407	7,407	7,407
Outcome mean	.528	.528	.528	.527	.527	.527	.525	.525	.525	.523	.523	.523	.522	.522	.522
Instructor & term FEs		Y	Y		Y	Y		Y	Y		Y	Y		Y	Y
Other control vars.			Y			Y			Y			Y			Y
Opt. Bandwidth = 7.80	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

This table shows estimates from regressions where observations are omitted from the middle on both sides of the threshold, for some x percentage points of number grade. The first three columns show results omitting 0.4 percentage points on either side of the cutoff, the next three columns omit 0.8 p.p., and so on up to 2.0 p.p. This is to address the potential concern that instructors draw letter grade cutoffs in part because of student characteristics that are unobservable to the researcher. This table shows that this is not a concern since the result is still a zero and is remains relatively precise even when excluding those students closest to the threshold. Bandwidths are fixed to match the main results. Estimated with a local linear regression and uniform kernel. Standard errors are in parenthesis and clustered at the instructor and term level.

Table A.4: Graduated with a STEM major

Subsample of obs. without missing variables

Panel A: Men	(1)	(2)	(3)	(4)	(5)	(6)
Above letter grade cutoff	0.0348	0.0507	0.0427	0.0455	0.0470	0.0316
	(0.0465)	(0.0316)	(0.0453)	(0.0325)	(0.0472)	(0.0305)
Observations	4,846	4,846	4,846	5,986	7,048	8,928
Outcome mean	.732	.732	.732	.735	.734	.728
Instructor & term FEs		Y	Y	Y	Y	Y
Other control vars.			Y	Y	Y	Y
Bandwidth = 5.70	1x	1x	1x	1.25x	1.5x	2x
Panel B: Women						
Above letter grade cutoff	0.0258	0.0020	-0.0010	0.0036	0.0031	-0.0089
	(0.0239)	(0.0173)	(0.0408)	(0.0216)	(0.0291)	(0.0871)
Observations	5,925	5,925	5,925	7,140	8,297	10,291
Outcome mean	.532	.532	.532	.535	.538	.537
Instructor & term FEs		Y	Y	Y	Y	Y
Other control vars.			Y	Y	Y	Y
Bandwidth = 7.82	1x	1x	1x	1.25x	1.5x	2x

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

This table repeats the estimates from Table 2 but with a subsample of the data containing only those observations for which no variable is missing. This subsample may not be a representative sample if variables are missing nonrandomly, and the smaller sample has less statistical power than the main specification. The point of this table is to show that the main results are robust to alternative methods of addressing the issue of missing variables. Estimated with a local linear regression and uniform kernel. Standard errors are in parenthesis and clustered at the instructor and term level.

Table A.5: Graduated with a STEM major

Results using multiple imputation for missing variables in columns 3-6

Panel A: Men	(1)	(2)	(3)	(4)	(5)	(6)
Above letter grade cutoff	0.0270	0.0304	0.0266	0.0302	0.0300	0.0178
	[0.0273]	[0.0206]	[0.0206]	[0.0158]*	[0.0136]**	[0.0106]
Observations	8,654	8,654	8,654	10,698	12,586	15,965
Outcome mean	.697	.697	.697	.698	.699	.695
Instructor & term FEs		Y	Y	Y	Y	Y
Other control vars.			Y	Y	Y	Y
Opt. Bandwidth = 5.72	1x	1x	1x	1.25x	1.5x	2x
Panel B: Women						
Above letter grade cutoff	0.0154	0.0045	0.0003	0.0034	0.0022	-0.0076
	[0.0157]	[0.0145]	[0.0146]	[0.0135]	[0.0094]	[0.0084]
Observations	10,078	10,078	10,078	12,246	14,267	17,778
Outcome mean	.522	.522	.522	.519	.516	.518
Instructor & term FEs		Y	Y	Y	Y	Y
Other control vars.			Y	Y	Y	Y
Opt. Bandwidth = 7.80	1x	1x	1x	1.25x	1.5x	2x

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

This table shows estimates of the effect of letter grade on the main outcome of interest, graduation with a STEM major within six years. For this table, I use multiple imputation to account for variables that are missing in some observations. The point estimates here are directly comparable to the main results in Table 2, although the standard errors here are clustered at the instructor level because I cannot compute two-way clustered errors with the multiple imputation command. Estimated with a local linear regression and uniform kernel.

Table A.6: Graduated with a STEM major

Results by individual letter grade thresholds

Panel A: Men	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	(A)	(A)	(A)	(B)	(B)	(B)	(C)	(C)	(C)
Above letter grade cutoff	0.0281	0.0406***	0.0381*	0.0510	0.0345	0.0353	0.1009*	0.1077**	0.0978**
	(0.0204)	(0.0122)	(0.0199)	(0.0471)	(0.0356)	(0.0344)	(0.0484)	(0.0407)	(0.0439)
Observations	1,989	1,989	1,989	3,811	3,811	3,811	2,197	2,197	2,197
Outcome mean	.855	.855	.855	.702	.702	.702	.532	.532	.532
Instructor & term FEs		Y	Y		Y	Y		Y	Y
Other control vars.			Y			Y			Y
Opt. Bandwidth	3.38			6.48			6.48		
Panel B: Women									
Above letter grade cutoff	0.0416	0.0183	0.0077	-0.0293	-0.0250	-0.0347	0.0010	-0.0207	-0.0228
	(0.0387)	(0.0323)	(0.0359)	(0.0205)	(0.0301)	(0.0329)	(0.0308)	(0.0251)	(0.0273)
Observations	1,827	1,827	1,827	3,619	3,619	3,619	3,716	3,716	3,716
Outcome mean	.665	.665	.665	.525	.525	.525	.408	.408	.408
Instructor & term FEs		Y	Y		Y	Y		Y	Y
Other control vars.			Y			Y			Y
Opt. Bandwidth	4.13			6.80			10.83		

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

This table shows estimates of the effect of letter grade on persistance in STEM major for each letter grade threshold separately. Estimated with a local linear regression and uniform kernel. Standard errors are in parenthesis and clustered at the instructor and term level.

Table A.7: Graduated with an economics or business major

Data from econonomics courses only

Panel A: Men	(1)	(2)	(3)	(4)	(5)	(6)
	Grad. with economics major			Grad. with business major		
Above letter grade cutoff	-0.0485	-0.0218	-0.0207	0.0683*	0.0418	0.0281
	(0.0238)	(0.0165)	(0.0117)	(0.0280)	(0.0318)	(0.0304)
Observations	3,555	3,555	3,555	3,322	3,322	3,322
Outcome mean	.154	.154	.154	.306	.306	.306
Instructor & term FEs		Y	Y		Y	Y
Other control vars.			Y			Y
Opt. Bandwidth =	6.28	6.28	6.28	5.82	5.82	5.82
Panel B: Women						
Above letter grade cutoff	-0.0059	0.0035	0.0078	0.0373	0.0224	0.0190
	(.)	(0.0059)	(0.0135)	(0.0356)	(0.0476)	(0.0526)
Observations	3,524	3,524	3,524	3,279	3,279	3,279
Outcome mean	.0857	.0857	.0857	.433	.433	.433
Instructor & term FEs		Y	Y		Y	Y
Other control vars.			Y			Y
Opt. Bandwidth =	8.52	8.52	8.52	7.86	7.86	7.86

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

This table shows results for only economics courses, looking at the effect of letter grade on graduation with an economics or business major within six years (columns 1-3 and 4-6, respectively). Estimated with a local linear regression and uniform kernel. Standard errors are in parenthesis and clustered at the instructor and term level. Note that the number of instructor clusters is quite low, with only four economics instructors in these regressions. This is why I cannot compute the standard error for women in column (1).

Table A.8: Results for alternative sample groups and outcomes

Impact on different majors and estimated earnings

Panel A: Men	(1)	(2)	(3)	(4)	(5)
	Main result	STEM w/ econ	Science mjrs	Engr mjrs	Median earnings
Above letter grade cutoff	0.0293	0.0272	-0.0326	-0.0146	1156.2427**
	(0.0222)	(0.0182)	(0.0209)	(0.0228)	(407.7100)
Observations	8,636	13,474	7,735	7,354	8,540
Outcome mean	.717	.62	.194	.797	76,514
Instructor & term FEs	Y	Y	Y	Y	Y
Opt. Bandwidth =	5.70	6.58	7.53	8.76	6.34
Panel B: Women					
Above letter grade cutoff	0.0051	0.0037	0.0414**	0.0118	500.0893
	(0.0142)	(0.0128)	(0.0160)	(0.0256)	(627.4462)
Observations	10,098	14,385	10,077	2,412	8,841
Outcome mean	.528	.456	.263	.631	61,248
Instructor & term FEs	Y	Y	Y	Y	Y
Opt. Bandwidth =	7.82	8.64	8.78	8.41	7.37

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

This table shows, for different subsets of the courses in the data, the impact of letter grades in those course on different outcomes. Column (1) repeats the main result from above using instructor and term fixed effects. Column (2) expands the course pool to include economics classes and looks at the effect of letter grade on the probability of graduating with a STEM or economics degree within six years. Column (3) uses only science courses including biology, calculus, and organic chemistry, and looks at the impact on majoring in a science field. Column (4) uses only engineering courses including calculus, computer programming, and statics and looks at the impact on majoring in engineering. Column (5) uses the STEM courses from the main sample (without economics classes) and looks at the impact of letter grade thresholds on national median earnings by major, in 2013 dollars. Estimated with a local linear regression and uniform kernel. Standard errors are in parenthesis and clustered at the instructor and term level.

B Appendix - probably not in main paper but included for reference

Table B.1: Effect of letter grades on graduating with a STEM major - 4 year graduation

Panel A: Men	(1)	(2)	(3)	(4)	(5)	(6)
Above letter grade cutoff	0.0208	0.0294	0.0227	0.0194	0.0107	0.0013
	(0.0214)	(0.0171)	(0.0250)	(0.0224)	(0.0171)	(0.0138)
Observations	7,197	7,197	7,197	8,814	10,229	12,821
Outcome mean	.576	.576	.576	.572	.569	.565
Instructor & term FEs		Y	Y	Y	Y	Y
Other control vars.			Y	Y	Y	Y
Opt. Bandwidth = 7.43	1x	1x	1x	1.25x	1.5x	2x
Panel B: Women						
Above letter grade cutoff	0.0192	0.0103	0.0083	0.0205	0.0169	0.0049
	(0.0254)	(0.0210)	(0.0264)	(0.0232)	(0.0162)	(0.0145)
Observations	7,327	7,327	7,327	8,953	10,467	13,187
Outcome mean	.475	.475	.475	.474	.475	.474
Instructor & term FEs		Y	Y	Y	Y	Y
Other control vars.			Y	Y	Y	Y
Opt. Bandwidth = 7.19	1x	1x	1x	1.25x	1.5x	2x

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Estimated with a local linear regression and uniform kernel. Standard errors are in parenthesis and clustered at the instructor and term level.

Table B.2: Effect of letter grades on graduating with a STEM major - 8 year graduation

Panel A: Men	(1)	(2)	(3)	(4)	(5)	(6)
Above letter grade cutoff	0.0272	0.0284	0.0251	0.0307	0.0278	0.0164
	(0.0305)	(0.0215)	(0.0245)	(0.0207)	(0.0161)	(0.0130)
Observations	8,293	8,293	8,293	10,243	12,056	15,253
Outcome mean	.748	.748	.748	.75	.75	.746
Instructor & term FEs		Y	Y	Y	Y	Y
Other control vars.			Y	Y	Y	Y
Opt. Bandwidth = 5.61	1x	1x	1x	1.25x	1.5x	2x
Panel B: Women						
Above letter grade cutoff	0.0131	0.0021	-0.0026	0.0029	0.0028	-0.0067
	(0.0188)	(0.0148)	(0.0174)	(0.0160)	(0.0112)	(0.0101)
Observations	9,843	9,843	9,843	11,986	13,970	17,403
Outcome mean	.536	.536	.536	.537	.538	.538
Instructor & term FEs		Y	Y	Y	Y	Y
Other control vars.			Y	Y	Y	Y
Opt. Bandwidth = 7.70	1x	1x	1x	1.25x	1.5x	2x

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Estimated with a local linear regression and uniform kernel. Standard errors are in parenthesis and clustered at the instructor and term level.