Reacting to Early Failure in University: Evidence from a Regression Discontinuity Design

Clemens Mueller[1]

University of Mannheim

Abstract

This paper investigates gender differences in persistence in educational attainment. I look at whether students complete university when they fail their very first university exam. I identify causal effects using university administration data and a sharp discontinuity at the passing threshold of the first university exam of 8,500 undergraduate students. Female students are completely unaffected and more resilient towards early failure in university. Male students who marginally fail their very first university exam are 15% less likely to successfully obtain a university degree. I add survey evidence to show that overconfidence and competitiveness explain the negative reaction of male students. I provide causal evidence of one reason why male students are less resilient in educational attainment. They have a strong negative reaction to early failure in university.

[This Version: June, 2023]

Keywords: educational attainment, gender, failure

[1]Comments appreciated. Please contact Clemens Mueller, University of Mannheim, clemens.mueller@uni-mannheim.de, +49 621 181 1362. I would like to thank Vicki L. Bogan, Alexandra Niessen-Ruenzi, and Alison Schultz as well as seminar participants at the University of Mannheim for helpful comments.

1. Introduction

Women are more persistent and resilient when it comes to educational attainment. Female students are for example more likely to complete high school (Murnane 2013) as well as university (Bailey and Dynarski 2011, Goldin et al. 2006). Conventional economic models are unable to explain these differences between male and female students in educational attainment. Factors such as socioeconomic status and ethnicity do not vary as male and female students grow up in the same families and attend the same schools. This paper aims to provide causal evidence on one reason why male students are less likely to complete university: Male students have a strong negative reaction to failure.

In this paper, I ask the question whether students drop out of university when they fail their very first university exam. To look at this, I use a sharp regression discontinuity design and compare students who marginally fail to those who marginally pass their very first university exam. The analysis is based on administrative data and detailed records of the first university exam of around 8,500 students of a mid sized German university. The exam I analyze is an introduction to financial mathematics course, mandatory in the first semester for all economics, business, and law majors. The final grade is fully determined by an exam 6 weeks after semester start, which means that the course is generally students' very first university exam.

Female students marginally above and marginally below the passing threshold successfully obtain a university degree with a probability of 89%. Male students marginally above the passing threshold complete university with a probability of 80%. Male students marginally below the passing threshold only complete university with a probability of 65%. Thus, there is an economically large and statistically significant discontinuity of 15%. These results are robust to various functional form specifications, bandwidth selection, and other econometric choices.

The empirical results hold under the assumption that students in a local area around the passing threshold are similar to each other, except for assignment into treatment (failing their very first university exam). I argue that marginally passing, or failing the first university exam is to some extent randomly determined. Precise manipulation of selection into treatment is an unlikely explanation. Students might very well be able to aim for a

certain number of points achieved in the exam. However, the precise passing threshold was unknown to students and course administrators alike before the exam. Grading was cross-sectional in nature such that 15% of the course failed the exam. Thus, the passing threshold was determined by the exam difficulty and competition in each respective cohort. The passing threshold varied across the years 2008-2018 as follows: {22.5, 17, 19, 20, 16.5, 16.5, 18.5, 18, 15, 16.5, 20.5}.

Reassuringly, there is no discontinuity of the distribution of male students around the passing threshold. Students on either side of the threshold are also similar on observable characteristics such as high school GPA, as a measure for student ability, and student age. Covariate continuity furthermore is balanced *within* gender, which means that, e.g. female students around the passing threshold have very similar high school grades and age. Also among students who failed their first university exam, female and male students look indistinguishable.

I next analyze heterogeneous effects and show that only German students react to failing their first exam by dropping out of university. There is no effect for Non-German students. This is consistent with an opportunity cost based explanation. Non-German students might face immigration or other restrictions and do not react to failing an exam. Next, only relatively older students react to failing their first exam by dropping out of university. Relatively older students are those who have worked or were involved in other activities before university, and as such likely face higher opportunity costs.

It is puzzling that male students show a strong negative reaction to early failure in university whereas female students do not. To analyze potential channels, I administer a survey among 927 students in the same course out-of-sample, in the first week of the fall semester 2022. I elicit expectations and attitudes towards failure and competition and, since the course is an introductory math class, some financial mathematics specific questions. There are several benefits of this exercise. First, this allows to measure attitudes and opinions that are not available in archival data. Second, I can link survey responses to students' realized performance in the exam. This allows to focus on students close to the passing threshold. These students are most important as they are closest to the regression discontinuity sub-population.

Expectations and overconfidence are likely a channel why male students drop out after failing an exam. There is literature that both men and women are overconfident, however men are more overconfident than women (Barber and Odean 2001, Niederle and Vesterlund 2007). In the survey, I asked students directly about their expected grade and on average I measure substantial overconfidence. Consistent with the literature, male students are more overconfident than female students. I next link survey responses to realized exam performance to show that male students are also more overconfident when *conditioning* on the realized performance in the exam. Male students around the passing threshold are in fact most overconfident.

Male students also self-assess as being less afraid and more prepared compared to female students. They are more likely to agree that they would be surprised to fail the course and less likely to agree that it would be a burden to fail the exam. The survey evidence indicates that male students are likely less emotionally prepared for early failure in university. Thus, failing the very first exam is likely to be a much more surprising event for male students compared to female students. They might react to this sudden shock of new information by dropping out of university.

Lastly, I look at competitiveness as a possible channel. We know from the literature that women shy away from competition, while men embrace it (Niederle and Vesterlund 2007, Buser et al. 2014, Flory et al. 2015, and Reuben et al. 2015). I confirm this empirically in the survey. Male students are more competitive compared to female students and are more likely to compare their performance with peers. It is also more important for male students to be better than their peers. The observed gender gap in attitudes towards competitiveness might have some explanatory power for why male students drop out when faced with failure, whereas female students do not. Consistent with this, it is precisely those students who face the strongest competition, relatively bad students, who react by dropping out of university.

I also analyze students' retake behavior. 81.7% of students attempt the retake exam in the future. Male students are 5% less likely to do so. This explains a third of the observed reaction and indicates that the reaction by male students is quick. Conditional on attempting the retake exam, male students are also 5% less likely to pass the retake

exam. They also perform worse, which indicates that they might exert less effort in the retake compared to female students.

By and large, students who fail their very first university exam pass the exam at the second attempt. Among those who marginally failed, only 5.6% fail the retake exam. This puts an upper bound on the mechanical component of the causal effect of failing the university exam on university completion. However, this mechanical component should be the same for both female as well as male students and thus cannot explain the baseline results.

I contribute to several strands of literature. First I add to the literature on educational attainment (Denning et al. 2022). I provide one causal channel on why male students have lower educational attainment. Male students, but not female students, seem to react negatively to early failure in university.

Second, I contribute to how men and women react differently to feedback (Azmat et al. 2019, Möbius et al. 2022). Previous research has shown that female students are more responsive than male students to positive incentives in the form of scholarships (Dynarski 2008) or negative incentives in the form of probation (Lindo et al. 2010). The results are in contrast to Wasserman (2021) and Wasserman (2022). In these papers female politicians are less (or equally) likely to persist after facing an electoral defeat.

The paper highlights the role expectations and belief updating can play in educational attainment and in general (Möbius et al. 2022, Thaler 2021, and Giustinelli 2022). In this paper, failing an exam leads to a sudden informational shock and this in turn leads to belief updating among students who fail the exam. Surprisingly, only male students react to this informational shock by dropping out of university, female students persist and do not react.

The remainder of this paper is structured as follows. Chapter 2 provides the institutional details and the data. Chapter 3 provides visual and reduced form regression discontinuity results. Chapter 4 discusses potential mechanisms. Chapter 5 concludes.

2. Institutional background and data

2.1. Institutional background

The data comes from the administration of a mid-sized German business school. All undergraduate students who major in economics, business, or law are mandated to take a course called "Financial Mathematics" in their first semester. The course covers basic concepts such as compound interest, net present value, annuity calculation and the rate of return of assets. 34% of students are majoring in business, 19% law, 16% in business and culture, 15% economics, and 14% in business education.

An important and distinct institutional feature is that the grade is 100% determined by a final exam and this exam already takes place after six weeks. This means that the course is the very first university exam for around 1,200 undergraduate students every year. Students have more than one opportunity to pass the exam. There are 3 credits, so-called European Credit Transfer and Accumulation System (ECTS) awarded upon successful completion. As a comparison, students need 180 ECTS to successfully complete a three-year undergraduate degree. Given the low number of credits awarded and the fact that students have several opportunities to pass the exam, the exam is low stakes. Students however perceive the exam as high stakes and we measure substantial fear and uncertainty ahead of students very first university exam.

Of particular relevance for the methodology of this paper is the structure and grading of the exam. Grading of students was historically done in a cross-sectional fashion. Every year, the passing threshold was determined ex-post, such that around 15% of students fail the exam. After the exam is written, the course administrators look at the point distribution and then determine the point threshold such that 15% of the cohort does not pass the exam. E.g. in the year 2011, the passing threshold was set to 20 out of 45 points. This means that a student who obtains precisely 20 points just marginally passes the course and receives the lowest passing grade. A student with 19 points, in contrast, just marginally fails and receives a failing grade. The points needed to successfully pass the exam is not constant over the years. The passing threshold varied across the years 2008-2018 as follows: {22.5, 17, 19, 20, 16.5, 16.5, 18.5, 18, 15, 16.5, 20.5}. Since passing/failing is cross-sectional in nature, the points needed to pass the exam were therefore

not known ahead of time to students and course administrators alike. The passing threshold is thus determined by 1) the difficulty of the exam and 2) the performance of the cohort. Given the uncertainty about the points needed to pass the exam, it is unlikely that students can precisely determine whether they pass or not. Later on in the paper, I revisit the ability of students to manipulate selection into treatment (failing the exam).

2.2. Data

I obtain the dependent variable, *Degree*, from the university administration. The variable is a dummy variable equal to one if a student has successfully obtained her degree. The variable is equal to zero if she has not successfully obtained her degree and is not currently enrolled anymore. Due to data protection reasons, students who are still enrolled cannot be considered.

I calculate the independent variable, *Points*, from historical grading data, collected from past course administrators. The data includes the total points, as well as the grade obtained in the financial mathematics exams of all students. The data also includes the passing threshold for each exam. The variable I compute is defined as the total points of the students minus the passing threshold. I refer to the variable as the point distance to the passing threshold. *Points* is equal to zero if the student just barely passes her exam with the exact points needed. It is equal to -1 for those who marginally fail the exam and equal to 1 for those who pass with a buffer of one point.

I obtain the gender of each student as recorded by the university administration. To control for student ability, I obtain the high school GPA. The high school GPA is by far the most important criteria for university admission in Germany. I also obtain a dummy variable equal to one if the nationality of the student is German, and zero otherwise. Lastly, to compute student age, I obtain the birth date.

The sample starts in the year 2010 as this marks the first year when grading data could be collected. I analyze all students who took the exam until the year 2017. The standard duration of an undergraduate degree is three years. I collect the information on university completion until the end of 2022, which means that students have at least four years to successfully complete their studies.

The research question looks at how male and female students react to failing their very first university examination. We therefore exclude students who might write other exams before the financial mathematics exam. This could contaminate the research design if students receive a signal on their quality beforehand. I apply the following filters to the dataset: I only keep the very first exam for each student in the data. This means that if a student failed her exam, but passes it at the second try, I only keep the first failed exam in the data. I drop students who do not write the exam at the semester start (2% of the sample), are sick, or do not show up on the exam day (2% of the sample). I also exclude students who deliberately cross out and thus fail the exam (1% of the sample). These filter steps guarantee that I look at each students' very first university exam and the observed effect is not mechanical, as students have subsequent tries to pass the exam. I end up with a sample of 8,588 students. The sample is purely cross-sectional, so every student only appears in the dataset once. Every student has a certain point distance to the passing threshold in their first university exam. The outcome variable then measures whether this student successfully obtained her degree or not.

2.3. Descriptive statistics

Descriptive statistics are shown in Table 1. Panel A shows the descriptive statistics for the full sample of students. 53% of students are female. 92% of undergraduate students are German. The remaining students are Chinese, Turkish, Bulgarian, among many other nationalities. The large fraction of German students is explained by the fact that for every undergraduate student, at least some courses are fully taught in German.

The average age of students at the time of the exam is equal to 19.9 at the mean and 20 at the median. The average German high school GPA is equal to 1.8 at the mean and 1.7 at the median. The high school GPA is by far the most important factor for university admission. The German grading scale ranges from 1.0 (best) to 5.0 (worst and a failing grade) and is inverted compared to an US-based GPA system. The grading is usually in increments of 0.3 as follows: 1.0, 1.3, 1.7, 2.0, 2.3, 2.7, 3.0, 3.3, 3.7, 4.0, 5.0. The best passing grade is a 4.0 and 5.0 is a failing grade.

The point distance to the passing threshold is 10.8 points at the mean and 12 at the

median. The average grade achieved in the financial mathematics exam is 2.7 at the mean and median. Based on descriptive statistics, students seem to perform much worse in university compared to high school. This is a typical feature of good students enrolled in a competitive university with cross-sectional grading.

Consistent with previous research, I observe a substantial female financial math gap in the data. Panel B shows only female and Panel C only male students. Female students achieve on average 9.2 points distance to the passing threshold. Male students achieve 12.6.

The average probability to successfully complete university is 92%. Female students are more persistent and graduate at higher levels compared to male students. 92% of female students and 91% of male students successfully obtain an undergraduate degree. This is consistent with previous literature such as Bailey and Dynarski (2011) and Goldin et al. (2006).

Table 1 – Summary statistics
The unit of observation is on a student level. Variable definitions are provided in the Appendix.

Panel A: Summary Statistics for All Students

Variable	N	Mean	SD	Min	25%	50%	75%	Max
Successful Degree	8,588	0.92	0.28	0	1	1	1	1
Point Distance to Passing Threshold	8,588	10.81	8.82	-10	5	12	18	25
Financial Math Grade	8,588	2.71	1.14	1.0	2.0	2.7	3.3	5.0
Male	8,588	0.47	0.50	0	0	0	1	1
School GPA	8,566	1.84	0.63	1	1.3	1.7	2.3	4
Age	8,588	19.93	1.77	16	19	20	20	42
German	8,588	0.92	0.27	0	1	1	1	1
Business Major	8,588	0.34	0.47	0	0	0	1	1
Law Major	8,588	0.19	0.39	0	0	0	0	1
Business and Culture Major	8,588	0.16	0.37	0	0	0	0	1
Economics Major	8,588	0.15	0.36	0	0	0	0	1
Business Education Major	8,588	0.14	0.35	0	0	0	0	1
Other Major	8,588	0.01	0.09	0	0	0	0	1

Panel B: Only Female Students

Variable	N	Mean	SD	Min	25%	50%	75%	Max
Successful Degree	4,539	0.92	0.27	0	1	1	1	1
Point Distance to Passing Threshold	4,539	9.19	8.70	-10	4	10	16	25
Financial Math Grade	4,539	2.92	1.13	1.0	2.0	2.7	3.7	5.0
School GPA	4,529	1.90	0.63	1	1.4	1.8	2.3	3.9
Age	4,539	19.91	1.78	17	19	20	20	40
German	4,539	0.90	0.30	0	1	1	1	1

Panel C: Only Male Students

Variable	N	Mean	SD	Min	25%	50%	75%	Max
Successful Degree	4,049	0.91	0.29	0	1	1	1	1
Point Distance to Passing Threshold	4,049	12.61	8.84	-10	7	14	20	25
Financial Math Grade	4,049	2.49	1.11	1.0	1.7	2.3	3.0	5.0
School GPA	4,037	1.78	1.13	1	1.7	1.6	2.2	4
Age	4,049	19.95	1.75	16	19	20	21	42
German	4,049	0.94	0.24	0	1	1	1	1

2.4. Survey data

I administered a survey out-of-sample, in the fall semester 2022. I asked 927 undergraduate students 16 questions related to students' attitudes towards competitiveness, failure, and expectations. I list all questions in the Appendix. The survey was administered in the first week of class before any contents were introduced. The goal of the survey is to shed light on potential channels for the baseline results.

I match survey responses to students' realized performance in the exam. This allows to compare male to female students conditional on realized grades. This allows to focus on

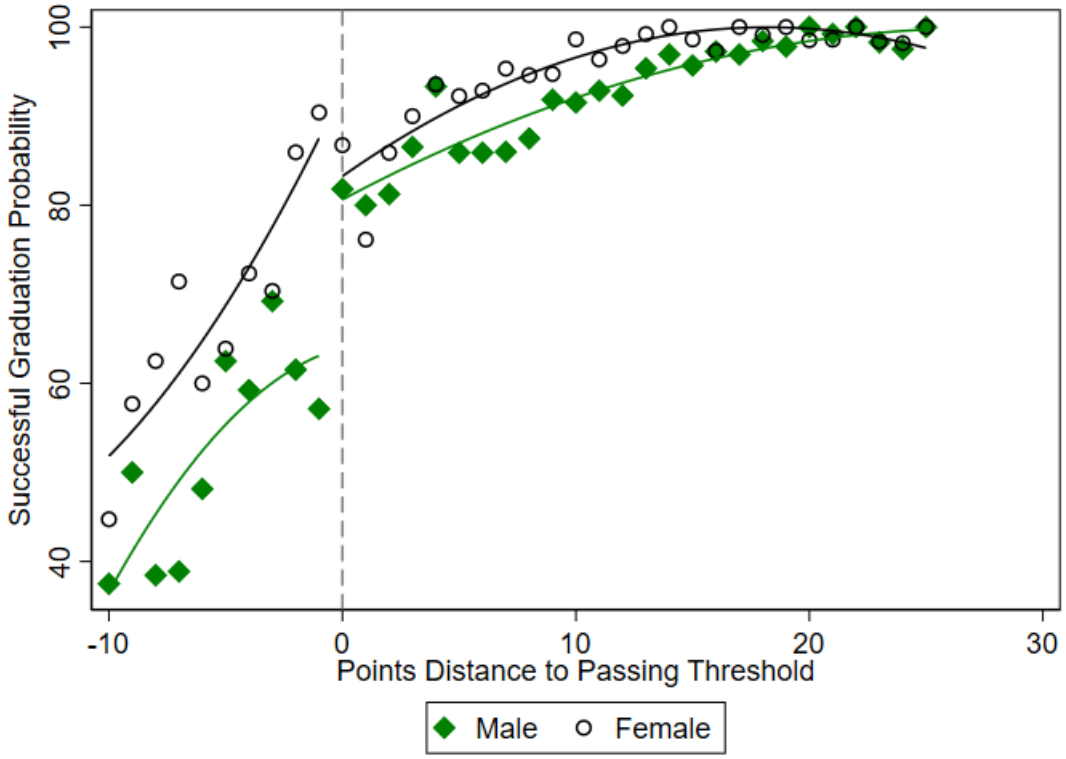
students close to the passing threshold.

3. Regression discontinuity evidence

3.1. Baseline regression discontinuity scatterplot

I visualize the average probability of obtaining a university degree, conditional on the point distance to the passing threshold in the financial mathematics exam, in a binned scatterplot and grouped by gender in Figure 1.

Figure 1 – Regression discontinuity: baseline results
This figure visualizes the raw data in a binned scatterplot. On the x-axis is the distance to the passing threshold in the financial mathematics exam. On the y-axis is the average university completion probability. Students are binned per point distance and capped at the extreme end at -10 and 25 points respectively. Male students are visualized in green squares and female students in white circles. The graph includes a polynomial of second order to both sides of the passing threshold and for each subgroup.



For female students there is no discontinuity around the passing threshold. On both sides around the threshold of zero, female students have a probability of successfully obtaining a degree of 89%.

Male students marginally above the passing threshold successfully obtain a degree with a probability of 80%. Male students marginally below the passing threshold obtain a degree with a probability of 65%. Based on the raw data alone, there seems to be a sharp drop of 15% in the probability of completing the university degree for male students. This is an economically sizable effect.

Female students are completing university at higher rates throughout the distribution, conditional on their results in a financial mathematics exam. Female students thus seem to be more resilient in educational attainment than male students. Additionally, failing their very first mathematics exam does not seem have any effect on the resilience of female students.

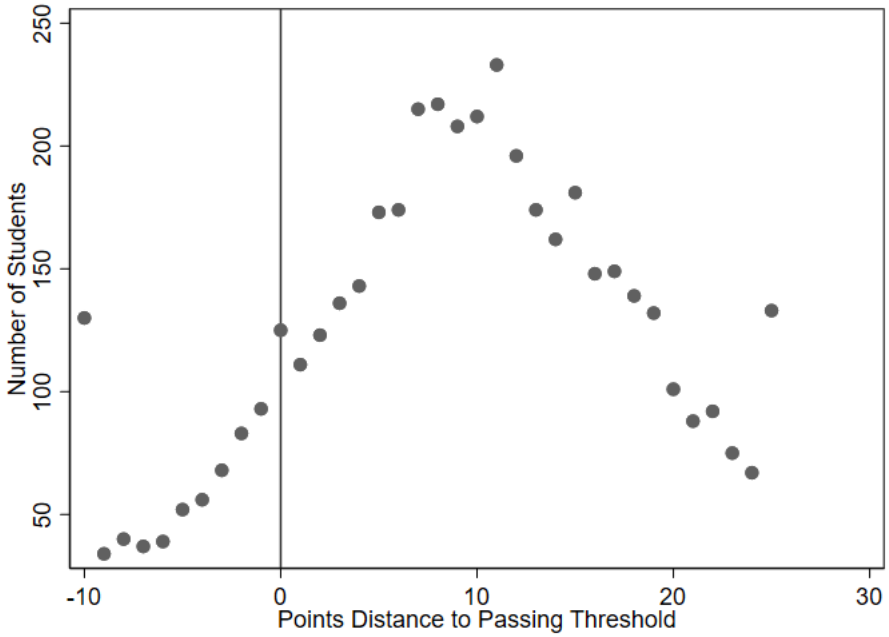
3.2. Are students able to manipulate the running variable?

The crucial assumption for analyzing causal effects is whether marginally passing, or marginally failing the first university exam is to some extent randomly allocated. A concern for a causal interpretation is whether students can precisely determine whether they pass the exam or not. This would be problematic if particularly skilled students manage to obtain just marginally enough points in order to pass the exam. If these students are also more likely to complete university, a causal interpretation is not valid. As a first test of this assumption, I visualize the distribution of students over all instances of the point distance to the passing threshold in Figure 2.

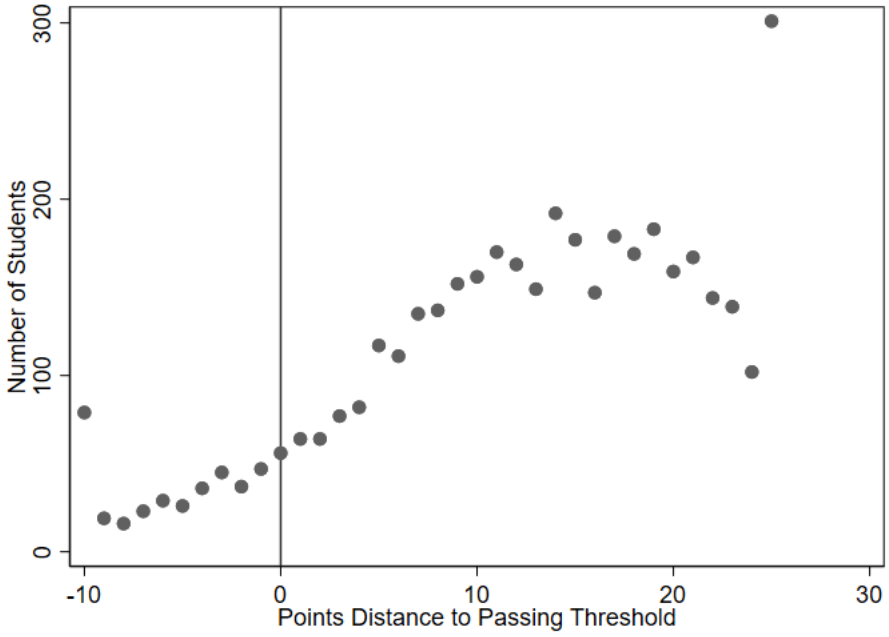
There is evidence for slight bunching above the passing threshold only for female students, however similar jumps appear throughout the distribution. Reassuringly, there is no discontinuity of the distribution of male students around the passing threshold.

Figure 2 – Threshold manipulation
This figure visualizes the distribution around the threshold. On the x-axis is the distance to the passing threshold in the financial mathematics exam. On the y-axis is the number of students. Students are binned per point distance and capped at the extreme end at -10 and 25 points respectively. Panel A shows the results only for female students. Panel B shows the results only for male students.

Panel A: Only Female Students



Panel B: Only Male Students



I argue that manipulation of the running variable, point distance to the passing threshold, by students is unlikely for several reasons: First, students might very well be able to determine how many points they achieve in the exam. However, it is probably not possible to do so with very high confidence, as grading by course instructors might be subjective. More importantly, the variable point distance to the passing threshold includes an additional component: the passing threshold. The passing threshold was not constant and unknown for students at the time of writing the exam. The precise points needed to pass the exam was also unknown to the instructors. The passing threshold varied across the years 2008-2018 as follows: {22.5, 17, 19, 20, 16.5, 16.5, 18.5, 18, 15, 16.5, 20.5} out of a maximum of 45 points. In every year, the passing threshold was set such that 15% of the students did not pass the course. The passing threshold was thus determined by outside factors such as the difficulty of the exam and the performance of each cohort of students. A student in the year 2010 needs to aim for precisely 19 points to pass the exam. If she would write the exam instead in the year 2008 or 2011, this would be a failing grade. The uncertainty involved in the passing threshold prohibits students to precisely manipulate the points they obtain in the exam to just marginally pass. To some degree, there is a random component in whether a student passes or fails her first university exam. Students are better off performing as good as they possibly can, and this is the most consistent explanation given the point distribution.

Second, no bunching occurs for male students around the passing threshold. There is limited bunching for female students, however this does not directly imply that students are able to precisely determine the point distance to the passing threshold. It might also be course administrators who push marginal students above the passing threshold. Indeed, some past course administrators corrected exams of students who marginally failed one additional time. Points were in some marginal cases adjusted upwards. This was only done for those who marginally failed the exam and never for those who marginally passed. Bunching above the threshold is thus not necessarily evidence in favor of running variable manipulation by students.

Nevertheless, the slight bunching of female students might be problematic. To mitigate this, I perform a robustness exercise. The discontinuity originates from two years in the

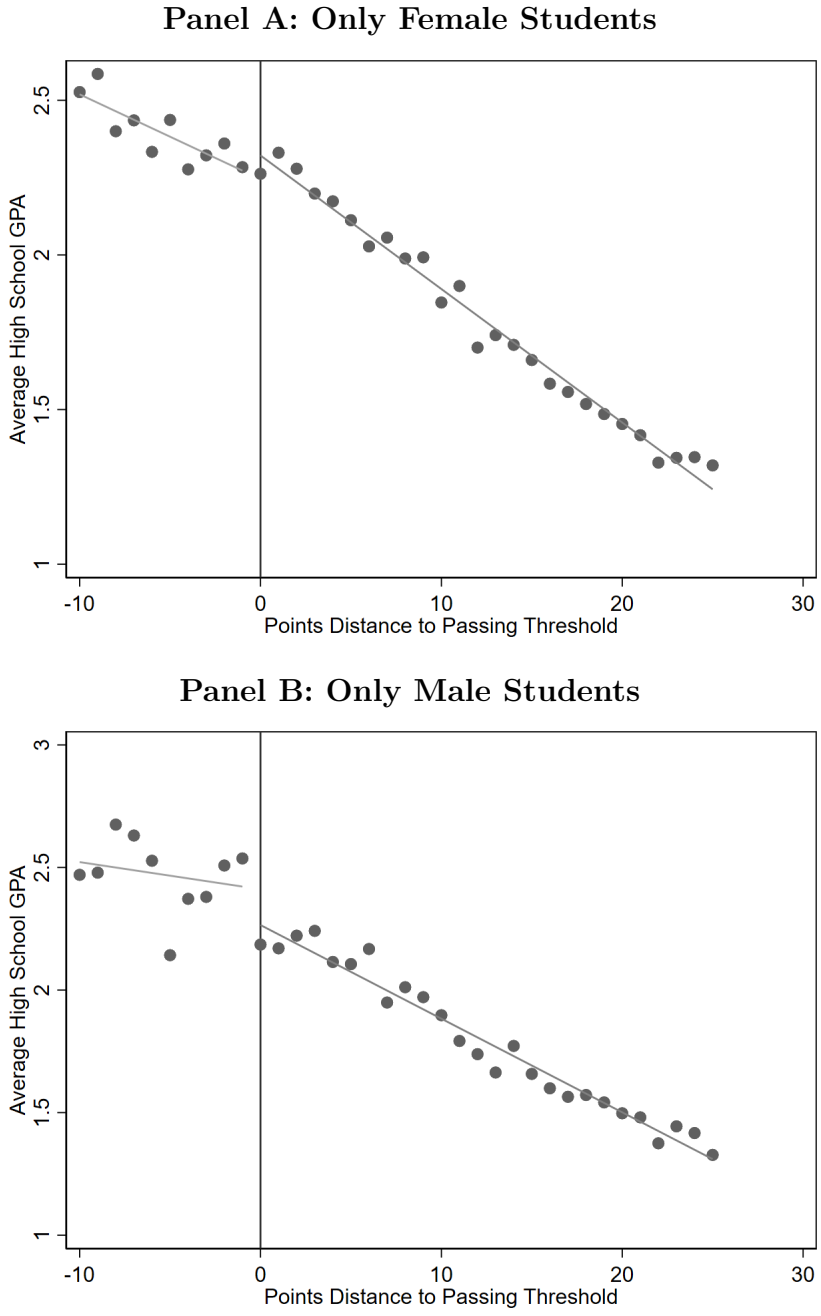
sample. Specifically, in the years 2011 and 2015, the course instructors regraded all exams marginally below the passing threshold. I repeat all analyses when excluding these two years. The results are shown in Figure A8. The distribution is smooth overall for male as well as female students and there is no visible bunching around the threshold. All results are unchanged when excluding these two years.

3.3. Are students on either side comparable?

Next, I analyze whether students on either side of the threshold are different when it comes to observable characteristics. As an imperfect proxy for student ability, the first characteristic I visualize is the high school GPA achieved. I perform a similar exercise as before, but instead of the number of students, I compute the average high school GPA for every point distance to the passing threshold.

The result is visualized in Figure 3. The average high school grade looks relatively smooth. For female students there is no visible discontinuity. For male students, there is a statistically insignificant jump of around 0.2 GPA around the threshold. However, similar jumps appear at other instances.

Figure 3 – High school GPA around threshold
This figure visualizes the high school average GPA around the threshold. On the x-axis is the distance to the passing threshold in the financial mathematics exam. On the y-axis is the average high school GPA, which is ranging from 1.0 (best) to 4.0 (worst). Students are binned per point distance and capped at the extreme end at -10 and 25 points respectively. Panel A shows the results only for female students. Panel B shows the results only for male students.

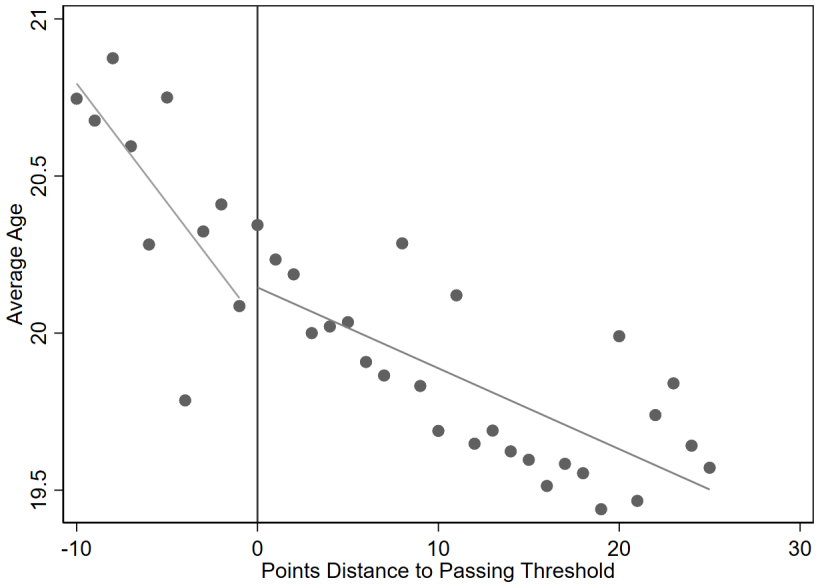


Lastly, I visualize the average age of students around the threshold in Figure 4.

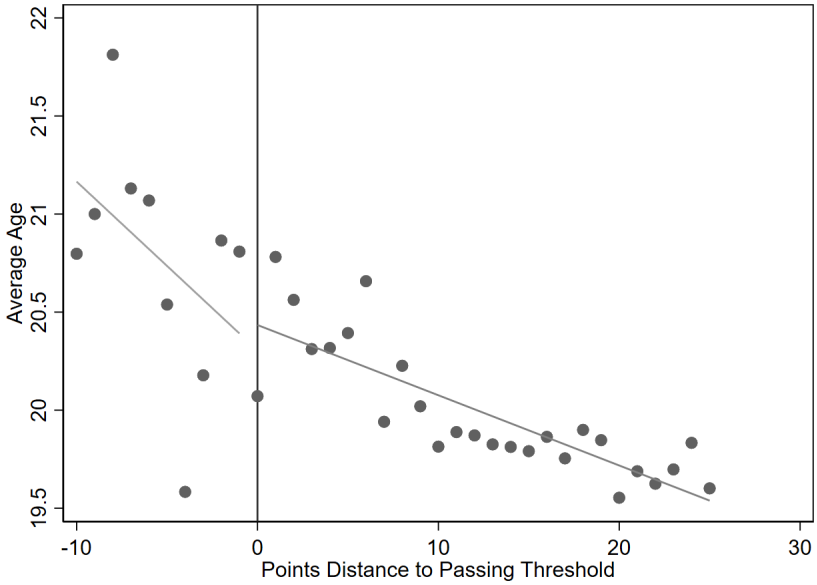
The average age looks continuous around the threshold, but the variable is noisy and I see frequent jumps in the distribution. For female students there is no visible discontinuity. For male students, those at the precise cutoff are somewhat younger compared to those below. However, widening the bandwidth by one point, students look very similar.

Figure 4 – Age around threshold
This figure visualizes the average age around the threshold. On the x-axis is the distance to the passing threshold in the financial mathematics exam. On the y-axis is the average age at the time of the exam. Students are binned per point distance and capped at the extreme end at -10 and 25 points respectively. Panel A shows the results only for female students. Panel B shows the results only for male students.

Panel A: Only Female Students



Panel B: Only Male Students



3.4. Baseline regression

I estimate the following baseline specification:

$$\begin{aligned} Degree_i = &\alpha Male_i + \beta Fail_i + \gamma(Male_i \times Fail_i) + f(P_i) + \theta(Fail_i \times f(P_i)) + \\ &\zeta(Male_i \times Fail_i \times f(P_i)) + \phi_t + \epsilon_i \end{aligned} \tag{1}$$

where *Degree* is defined as a dummy variable equal to one if student i successfully obtains an undergraduate degree, and zero if not. Year t is defined as the year of university entry which coincides with the year of the financial mathematics exam. The variable $Male$ is equal to one if the student is male, as indicated by university administration data. The variable $Fail$ is equal to one if the student failed her first university exam. The running variable, point distance to the passing threshold is included as a function either as a linear term or using higher order polynomials.

$Male_i$ captures a level shift between the average passing probability of male relative to female students. $Fail_i$ captures the intercept shift for female students who fail the exam. The variable of interest is thus the interaction term $Male_i \times Fail_i$ which picks up the effect of failing the first university exam for male students. The function $f(P_i)$ captures the effect of the point distance to the passing threshold for female students. The interaction with $Fail_i \times f(P_i)$ allows to include a different slope for female students who fail the exam. The interaction with $Male_i \times f(P_i)$ allows to include a different slope for male students who pass the exam. Lastly, the triple interaction $Male_i \times Fail_i \times f(P_i)$ captures a different slope estimate for male students who failed their first exam. I include Major $\times$ Year fixed effects and cluster standard errors on the running variable as suggested by Lee and Card (2008). The results are robust to using Eicker-Huber-White heteroskedasticity-robust standard errors following Kolesár and Rothe (2018) and Armstrong and Kolesár (2020).

To ease interpretation, I invert the running variable. I multiply the point distance to the passing threshold by -1 and subtract a constant of 0.000005 to the students with a value equal to zero. Treatment is defined as failing the very first university exam, so after this modification, I can interpret the treatment indicator $Male_i \times Fail_i$ as the causal effect of

failing the very first university exam for male students.

The results are shown in Table 2. When male students marginally fail their very first university exam, the probability of successfully obtaining a degree decreases by between 14% to 32% depending on the specification. Male students are on average less likely to obtain a degree compared to female students. When using local linear functions, women appear to be more likely to successfully finish university when they fail the exam, however this effect disappears when looking at either full sample linear, parametric regressions, or choosing a local linear non-parametric specification with a bandwidth of 2. Only male students seem to significantly react to failing their very first university exam. Overall the coefficients on $Male_i \times Fail_i$ is well aligned with the visual results presented earlier. Older students are less likely to successfully finish their undergraduate degree. I also see a strong relationship between high school performance and students' likelihood to complete university. Lastly, German students are much more likely to complete university compared to non-German students. These results are consistent with results in the literature.

The table also reports parametric regression discontinuity specifications using a second-order polynomial of the point distance to the passing threshold using the full sample of students. The optimal bandwidth is calculated as equal to 3 points around the threshold following Calonico et al. (2014). The regression output thus reports non-parametric local linear regression with the optimal bandwidth of 3 points, as well as using either 2 or 4 points around the threshold.

The results are robust to using local randomization regression discontinuity approaches. In the context of this research question, the running variable is not continuous, but can be seen as discrete. This leads to a moderate number of distinct masspoints. The number of discrete instances of the point distance to the passing threshold is equal to 36 unique values in the interval [-10,25]. Since the optimal bandwidth in this context is not necessarily appropriate, I refer to economic intuition. The most stringent bandwidth would be one point. This equates to comparing students just above to just below the threshold. The results are unchanged to using this most stringent comparison or widening the interval to either two or three points around the cutoff. This is essentially a

trade-off between sample size and the assumption of random assignment into treatment in a narrow window around the threshold.

Table 2 – Baseline regression discontinuity: reacting to early failure
This table reports the regression discontinuity of equation 1. The dependent variable is equal to one if the student successfully finished her undergraduate degree. Column (1) shows a linear regression using the full sample. Column (2) shows parametric regressions using a fully interacted model including a second order polynomial of the running variable. Columns (3) to (5) display non-parametric local linear regressions with a bandwidth of 2, 3, and 4 respectively. The optimal bandwidth is calculated as equal to 3 following Calonico et al. (2014). Variable definitions are provided in the Appendix. The regression includes Major × Year fixed effects. Standard errors are clustered on the level of the running variable: point distance to the passing threshold. ***, ** and * represents significance at the 1%, 5%, and 10% level, respectively. t-statistics are displayed in parenthesis.

	(1)	(2)	(3)	(4)	(5)
Sample:	Full Sample		Bandwidth: 2	Bandwidth: 3	Bandwidth: 4
Polynomial Order:	1	2	1	1	1
Male	-0.04**	-0.03	-0.07	-0.05	-0.06
	(-2.50)	(-1.37)	(-1.44)	(-0.76)	(-1.12)
Fail	0.02	0.06	0.06	0.17**	0.14**
	(0.47)	(1.31)	(1.67)	(2.55)	(2.57)
Male×*Fail*	-0.14**	-0.28***	-0.17**	-0.32***	-0.24***
	(-2.13)	(-5.54)	(-3.60)	(-4.05)	(-3.77)
SchoolGPA	-0.05***	-0.05***	-0.12**	-0.08**	-0.08**
	(-5.26)	(-5.20)	(-3.54)	(-3.09)	(-2.69)
Age	-0.01***	-0.01***	-0.02**	-0.02**	-0.02***
	(-4.26)	(-4.20)	(-3.88)	(-3.59)	(-4.57)
German	0.08***	0.08***	0.19**	0.14**	0.15***
	(5.59)	(5.45)	(3.55)	(3.16)	(4.04)
Observations	8,563	8,563	797	1,121	1,438
R-squared	0.24	0.24	0.15	0.13	0.14
Major × Year FE	YES	YES	YES	YES	YES

4. What explains the reaction of male students?

In the following, I will first analyze heterogeneity in the data and second why male students might react strongly to early failure in university, while female students do not. I do so using two complementary datasets. First, I rely on sources of heterogeneity in the data. Students differ along various characteristics, which might indicate why some drop out and others do not. The benefit of relying on the regression discontinuity sample is that it relies on a revealed choice: dropping out. The drawback is that I have little data

and imperfect proxies. The second dataset comes from a survey I administered out-of-sample among 927 students who took the course in the fall semester 2022. The benefit of the survey is that I could elicit expectations and self-assessments on potential channels that are unobservable in the archival data. I match survey responses to the realized exam performance, which allows to analyze gender differences in survey responses particularly for students around the passing threshold. This is the local student population most relevant for the research design and I particularly focus on gender differences in this local subset. The drawback is that since the survey includes the out-of-sample cohort of 2022, it is impossible to analyze who eventually drops out of university.

4.1. Heterogeneity

I explore two separate sources of heterogeneity in the data. First, I split the sample into German and Non-German students. Non-German students might have visa restrictions and face more legal and financial mobility restrictions compared to German students. Consistent with this, the effects are confined to male students who are German. Non-German students do not seem to react by dropping out of university.

The second source of heterogeneity I explore is student age. I split the sample at the median into relatively older and younger students. Only male students who are relatively old drop out of university. Relatively older students are more likely to be involved in some other activity before starting university. Such students might have worked, finished an apprentice program, etc. Older students might thus have higher opportunity costs of continuing education. Or to phrase differently, it might be easier for them to switch to another activity besides full time studying. Upon early failure in university, they might go back to their previous job or switch to another university.

Table 3 – Heterogeneity: Old and German students
This table reports heterogeneity regressions similar to equation 1. The dependent variable is equal to one if the student successfully finished her undergraduate degree. The sample is composed with a bandwidth of 4 points around the cutoff. The sample is split around the local median into two parts. In column (1) and (2), the students are split into those below the age of 20.2 (young students) and those above (old). In column (3) and (4), the students are split into German and Non-German. Variable definitions are provided in the Appendix. Standard errors are clustered on the level of the running variable: point distance to the passing threshold. ***, ** and * represents significance at the 1%, 5%, and 10% level, respectively. t-statistics are displayed in parenthesis.

	(1)	(2)	(3)	(4)
Sample:	Young students	Old students	German	Non-German
$Male$	-0.07	-0.10***	-0.07*	-0.17
	(-1.50)	(-4.16)	(-2.29)	(-0.85)
$Fail$	0.20**	0.13*	0.16**	0.14
	(2.97)	(2.16)	(3.20)	(1.13)
$Male \times Fail$	-0.22	-0.42***	-0.39***	0.38
	(-1.79)	(-12.94)	(-6.93)	(1.52)
Observations	677	675	1,221	130
R-squared	0.09	0.12	0.10	0.24
Controls	YES	YES	YES	YES
Year FE	YES	YES	YES	YES

4.2. Expectations, overconfidence, and failure

Expectations might play an important role and explain the differing response of male students compared to female students. To analyze this potential channel, I add evidence from the survey. The sample is composed of 927 students in the out-of-sample 2022 cohort. I specifically asked students at the beginning of the semester what grade they expect to earn in financial mathematics. Students could select every grade step from 1.0 (best) to 5.0 (worst, and a failing grade). Students expect significantly better grades (0.5 grade points on average) than they ended up achieving.

There is significant sorting of students into majors. Different majors differ on how competitive they are. By far the most important criteria to enter a certain major is the high school average grade. Because of this, all regressions include major fixed effects and thus for example compare male economics students to female economics students.

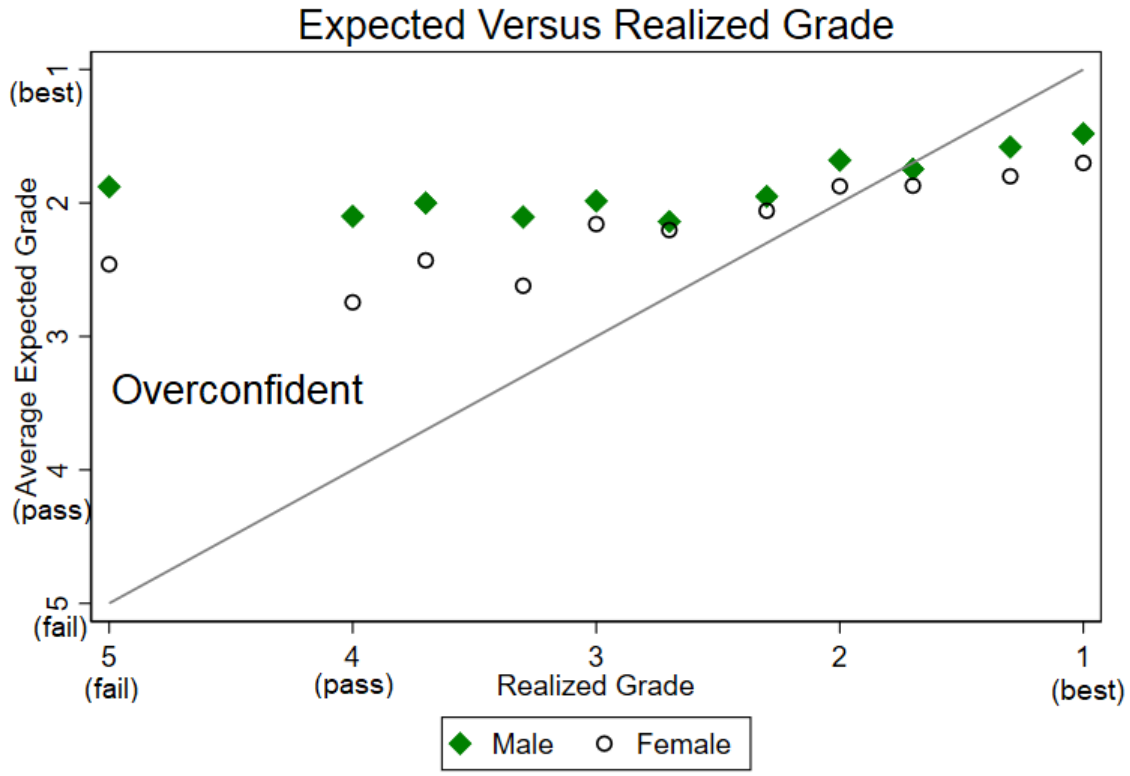
In Table 4 column (1), I see that male students on average expect 0.16 better grades compared to female students. The results are a first indication that male students are overconfident compared to female students. Next, I condition on the realized exam performance. I match the survey responses to realized grades to visualize the gender gap in

expectations. Figure 5 visualizes the results.

On the x-axis, we see students grouped by what grade they achieved in the course, separate by gender. On the y-axis, we see the average expected grade of these students. On average both female and male students expect worse grades only for the three best grades 1.0, 1.3, and 1.7. From grade 2.0 onwards, students expect better grades.

Along the complete realized grade distribution, male students expect better grades compared to female students. Thus, even conditioning on the performance in financial mathematics, male students seem to be overconfident. Overconfidence is small at the upper end of the distribution and smallest in the middle. It is largest at the tail end of the distribution, precisely in the area close to the passing threshold. Thus in the local area relevant for the baseline results, male students are most overconfident compared to female students. Overconfidence might thus explain some of the response to failing an exam for male students.

Figure 5 – Expected versus realized grade
This figure visualizes the average expected grade on the y-axis and the realized grade on the x-axis.



Next I look at the general attitude of male students towards failing. The goal is to elicit whether male compared to female students differ in their expectations towards failure specific to the financial mathematics exam. I ask students to what extent they agree to the following statements: "I would be surprised to fail the financial mathematics exam", "I am afraid of financial mathematics", "It is a burden to fail this course", and "I would consider dropping out of university if I fail this course".

In Table 4 columns (2) to (5) I show the results. Male students are significantly more likely to be surprised if they would fail the course. Male students are also much less likely to be afraid of financial mathematics. They are also slightly less likely to say failing the exam would be a burden. There is no difference in their personal perception of whether they would consider dropping out of university if they would fail the exam.

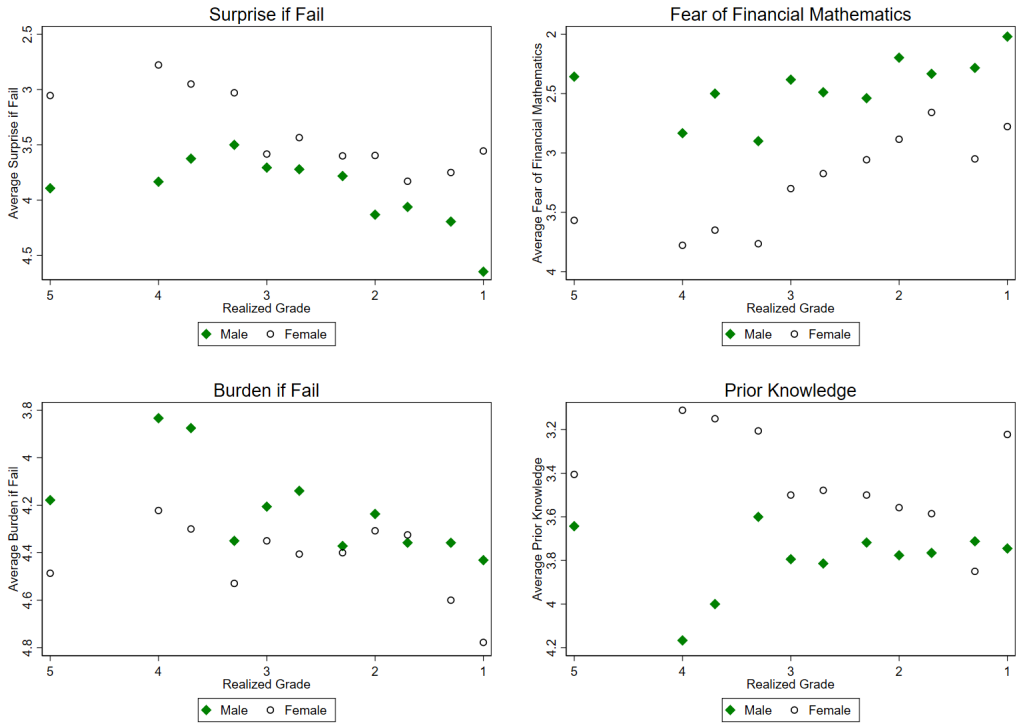
These results are consistent with the hypothesis that male students are less emotionally

prepared for failure. They tend to be more overconfident and self assess to be more surprised to fail the exam. They also state that they are not afraid of the exam. In Figure 6, I again see that the local student population towards the bottom of the performance distribution is the one where the gap between male and female students is highest.

Table 4 – Survey evidence: expectations and failure
This table shows survey results of the out-of-sample cohort of the fall semester 2022. In column (1) I ask respondents what grade they expect, ranging from 1.0 (best) to 4.0 (worst). In the following columns, I ask respondents on a 5 point Likert scale to what extent they agree with the following statements: (2) I will be surprised if I fail this exam. (3) I am afraid of this course. (4) It will be a burden for me if I fail this course. (5) I will consider dropping out if I fail this course. ***, ** and * represents significance at the 1%, 5%, and 10% level, respectively. t-statistics are displayed in parenthesis.

	(1)	(2)	(3)	(4)	(5)
	Expected Grade	Surprised if fail exam	Afraid of course	Burden if fail	Drop out if fail
Male	-0.16***	0.40***	-0.66***	-0.10*	0.08
	(-4.00)	(6.59)	(-8.94)	(-1.81)	(1.20)
Major FE	YES	YES	YES	YES	YES
Observations	927	927	927	926	926
R-squared	0.23	0.16	0.20	0.02	0.00

Figure 6 – Overconfidence and failure
This figure visualizes survey results of the out-of-sample cohort of the fall semester 2022. On the x-axis, students are grouped by their realized grade in the course. I asked respondents to what extent they agree to various statements. Top left: "I would be surprised to fail the financial mathematics exam", top right: "I am afraid of financial mathematics", bottom left: "It is a burden to fail this course", and bottom right: "prior knowledge will help me master this course". The responses were on a 5 point Likert scale.



4.3. Competitiveness

Another channel I investigate is the role of competition and attitudes towards competitiveness. In the context of the results, competition might be a shock and students might suddenly realize that they are facing strong competition when they fail their first university exam.

I first use evidence of the survey and see in Table 5 that male students are significantly more likely to compare their performance to peers. It is also more important for male students to be better than their peers. They are more competitive and self-assess as more likely to want to win a game. This is consistent with the literature where men are consistently seen as more competitive compared to female students.

Table 5 – Survey evidence: competition
This table shows survey results of the out-of-sample cohort of the fall semester 2022. I ask respondent on a 5 point Likert scale to what extent they agree to the following statements: (1) I often compare my results with my peers. (2) It is important for me to be better than my peers. (3) When I play a game I want to win. (4) My performance is important to my self-worth. (5) I often think about my own performance. ***, ** and * represents significance at the 1%, 5%, and 10% level, respectively. t-statistics are displayed in parenthesis.

	(1)	(2)	(3)	(4)	(5)
	Compare results with peers	Important to be better than peers	Want to win game	Performance important self-worth	Think often about performance
Male	0.17***	0.20***	0.41***	-0.18***	-0.18***
	(2.60)	(2.94)	(7.27)	(-3.17)	(-3.59)
Major FE	YES	YES	YES	YES	YES
Observations	927	927	927	926	927
R-squared	0.04	0.11	0.06	0.01	0.03

I add evidence from the archival data to this. Students in the sample were among the best in high school, but are suddenly compared to other equally high achieving students in university. I hypothesize that relatively worse students, those that suddenly face more fierce competition, are reacting more negatively to early failure in university.
To analyze this question in the regression discontinuity sample, I split students into two subgroups. Depending on their high school GPA, relatively worse students, who I argue face much more competition and relatively good students who face less competition.
Indeed, dropping out of university is strongly concentrated in the subgroup of male students who are relatively bad and face strong competition in university.

Table 6 – Regression discontinuity: competition
This table reports heterogeneity regressions similar to equation 1. The dependent variable is equal to one if the student successfully finished her undergraduate degree. The sample is defined with a bandwidth of 4 points around the cutoff. The sample is split around the local median into two parts. In column (1) and (2), the students are split into relatively good (above a GPA of 2.2.) and relative bad students, respectively. Variable definitions are provided in the Appendix. Standard errors are clustered on the level of the running variable: point distance to the passing threshold. ***, ** and * represents significance at the 1%, 5%, and 10% level, respectively. t-statistics are displayed in parenthesis.

	(1)	(2)
Sample:	Good students	Bad students
$Male$	-0.04	-0.11***
	(-0.76)	(-7.15)
$Fail$	0.16***	0.19**
	(4.04)	(2.71)
$Male \times Fail$	0.00	-0.59***
	(0.06)	(-8.23)
Observations	745	608
R-squared	0.09	0.12
Controls	YES	YES
Year FE	YES	YES

4.4. Retake behavior

Lastly, I analyze how male and female students differ in their retaking behavior after failing the exam. Do male and female students attempt the retake at similar rates? And conditional on retaking the exam, how do male and female students perform? To analyze these questions, I construct data which captures retake behavior of students who failed their first exam. I first calculate a dummy equal to one if the student attempts the retake exam, which 81.7% of students do. I then analyze gender differences in table 7.

Male students are 5% less likely to attempt the retake exam compared to female students, significant at a 10% level, and marginally insignificant at a 5% level. About one third of the baseline effect can thus be explained by the fact that male students do not attempt to retake the exam. Male students seem to react quickly and drop out of university.

Second, I analyze the performance in the retake conditional on retaking. Conditional on retaking, male students are 5% less likely to pass the retake. Looking at the point distance in the retake, there is no statistically significant difference between male and female students. Noteworthy is that while there is a gender gap overall in the financial mathematics exam, when looking at the subset of students who fail the exam, the gender

gap reverses and male students seem to perform worse than female students. This might indicate that female students exert more effort compared to male students in the retake exam.

Table 7 – Exam retake behavior
This table shows regressions on students retake behavior. In column (1) the dependent variable is a dummy equal to one if the student attempts a retake exam in the future. In column (2), the variable is a dummy whether the student passes the retake. In column (3) the dependent variable is the point distance to the passing threshold in the retake exam. The sample only includes students who failed their first attempt. ***, ** and * represents significance at the 1%, 5%, and 10% level, respectively. t-statistics are displayed in parenthesis.

	Attempt Retake	Pass Retake	Points Retake
Male	-0.05*	-0.05*	-0.81
	(-1.89)	(-1.81)	(-1.38)
SchoolGPA	-0.04	-0.09***	-3.09***
	(-1.48)	(-3.82)	(-5.47)
Age	-0.01	-0.00	0.03
	(-1.57)	(-0.80)	(0.22)
German	-0.02	0.08**	3.69***
	(-0.42)	(2.04)	(4.23)
Observations	984	805	805
R-squared	0.04	0.06	0.13
Major FE	YES	YES	YES
Year FE	YES	YES	YES

5. Conclusion

This paper analyzes the question whether failing the very first university exam causes students to drop out of university. I exploit university administration data of around 8,500 students and a sharp discontinuity at the passing threshold of the very first university exam. I show that male students who marginally fail their very first university exam are 15% less likely to successfully obtain a university degree. Female students on the other hand are much more resilient to failure in university. The channels are consistent with the explanation that overconfidence and attitudes to competitiveness explain the reaction of male students. The results provide causal evidence of one explanation on why male students are less likely to successfully obtain a university degree: male students react strongly negative to early failure in university.

References

Armstrong, T. B. and M. Kolesár (2020). Simple and honest confidence intervals in nonparametric regression. *Quantitative Economics 11*(1), 1–39.

Azmat, G., M. Bagues, A. Cabrales, and N. Iriberri (2019). What you don't know...can't hurt you? a natural field experiment on relative performance feedback in higher education. *Management Science 65*(8), 3714–3736.

Bailey, M. and S. Dynarski (2011). Gains and gaps: Changing inequality in u.s. college entry and completion.

Barber, B. M. and T. Odean (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics 116*(1), 261–292.

Buser, T., M. Niederle, and H. Oosterbeek (2014). Gender, competitiveness, and career choices *. *The Quarterly Journal of Economics 129*(3), 1409–1447.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica 82*(6), 2295–2326.

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

Dynarski, S. (2008). Building the stock of college-educated labor. *The Journal of Human Resources 43*(3), 576–610.

Flory, J. A., A. Leibbrandt, and J. A. List (2015). Do competitive workplaces deter female workers? a large-scale natural field experiment on job entry decisions. *The Review of Economic Studies 82*(1), 122–155.

Giustinelli, P. (2022). Expectations in education: Framework, elicitation, and evidence. *SSRN Electronic Journal*.

Goldin, C., L. F. Katz, and I. Kuziemko (2006). The homecoming of american college women: The reversal of the college gender gap. *Journal of Economic Perspectives 20*(4), 133–156.

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

Lee, D. S. and D. Card (2008). Regression discontinuity inference with specification error. *Journal of Econometrics 142*(2), 655–674.

Lindo, J. M., N. J. Sanders, and P. Oreopoulos (2010). Ability, gender, and performance standards: Evidence from academic probation. *American Economic Journal: Applied Economics 2*(2), 95–117.

Möbius, M. M., M. Niederle, P. Niehaus, and T. S. Rosenblat (2022). Managing self-confidence: Theory and experimental evidence. *Management Science*.

Murnane, R. J. (2013). U.s. high school graduation rates: Patterns and explanations. *Journal of Economic Literature 51*(2), 370–422.

Niederle, M. and L. Vesterlund (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics 122*(3), 1067–1101.

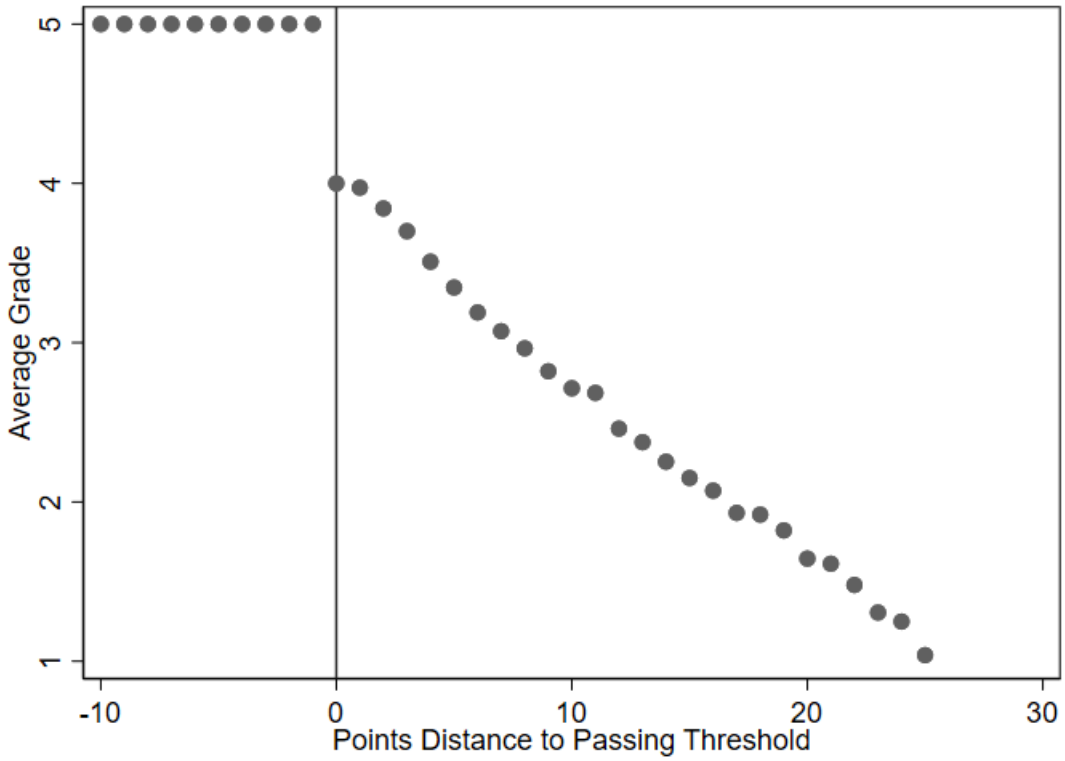
Reuben, E., P. Sapienza, and L. Zingales (2015). Taste for competition and the gender gap among young business professionals.

Thaler, M. (2021). Gender differences in motivated reasoning. *Journal of Economic Behavior & Organization 191*, 501–518.

Wasserman, M. (2021). Up the political ladder: Gender parity in the effects of electoral defeats. *AEA Papers and Proceedings 111*, 169–173.

Wasserman, M. (2022). Gender differences in politician persistence. *Review of Economics and Statistics, forthcoming*.

Figure 7 – Visualizing the Sharp Discontinuity
This figure visualizes the sharp discontinuity which is exploited in the analysis. On the x-axis is the distance to the passing threshold in the first university exam. On the y-axis is the average grade, which is a function of the point distance in the exam. Students are binned per point difference and capped at the extreme end at -10 and 25 points respectively. The German grading scale ranges from 1.0 (best) to 5.0 (worst and a failing grade), usually in increments of 0.3 as follows: 1.0, 1.3, 1.7, 2.0, 2.3, 2.7, 3.0, 3.3, 3.7, 4.0, 5.0. The best passing grade is a 4.0 and 5.0 is a failing grade. Marginally failing the exam results in a sharp drop from grade 4.0 to 5.0.



APPENDIX

A. Variable Definitions

This section provides the variable definitions. The data is on a pure cross-sectional student level.

1. *Points* – Number of points relative to the passing threshold. 0 indicates that the student has just passed the exam. -1 equals that one additional point was needed to pass the exam. +1 indicates that the students passed the exam with a buffer of one point. The variable points is binned at the two extremes at -10 and at +25.
2. *Grade* – Grade captures what grade the student achieved in her first university exam. The German grading scale ranges from 1.0 (best) to 5.0 (fail), usually in increments of 0.3 as follows: 1.0, 1.3, 1.7, 2.0, 2.3, 2.7, 3.0, 3.3, 3.7, 4.0, 5.0. The best passing grade is a 4.0 and 5.0 is a failing grade.
3. *Fail* – Dummy variable equal to one if the student did not pass her very first university exam: financial mathematics.
4. *Degree* – Dummy variable equal to one if the student has successfully completed her undergraduate university degree.
5. *SchoolGPA* – High school average grade which is used for university admission. The German educational system does not use standardized tests, thus high school GPA is by far the most important criteria for university admission. The German grading scale ranges from 1.0 (best) to 5.0 (worst and a failing grade), usually in increments of 0.3 as follows: 1.0, 1.3, 1.7, 2.0, 2.3, 2.7, 3.0, 3.3, 3.7, 4.0, 5.0. The best passing grade is a 4.0 and 5.0 is a failing grade.
6. *Age* – Age of the student at the time of the exam.
7. *German* – Dummy equal to one if the student is a German national.

Survey questions

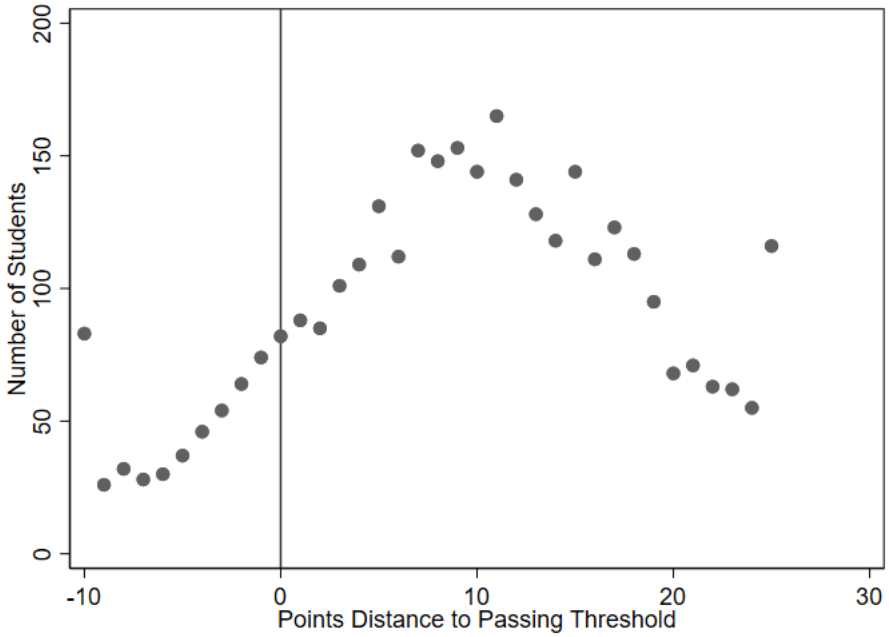
1. What grade are you expecting to earn in this course? The German grading scale ranges from 1.0 (best) to 5.0 (fail), usually in increments of 0.3 as follows: 1.0, 1.3, 1.7, 2.0, 2.3, 2.7, 3.0, 3.3, 3.7, 4.0, 5.0. The best passing grade is a 4.0 and 5.0 is a failing grade.
2. I often compare results with my peers. Likert scale ranging from 1 (I disagree strongly) to 5 (I agree strongly).
3. It is important for me to be better than my peers. Likert scale ranging from 1 (I disagree strongly) to 5 (I agree strongly).
4. If i play a game, I want to win. Likert scale ranging from 1 (I disagree strongly) to 5 (I agree strongly).
5. My university performance is important for my self worth. Likert scale ranging from 1 (I disagree strongly) to 5 (I agree strongly).
6. I often think about my university performance. Likert scale ranging from 1 (I disagree strongly) to 5 (I agree strongly).
7. It is important for me to be good in financial mathematics. Likert scale ranging from 1 (I disagree strongly) to 5 (I agree strongly).
8. Financial mathematics is of interest to me. Likert scale ranging from 1 (I disagree strongly) to 5 (I agree strongly).
9. Financial mathematics is an important subject for me. Likert scale ranging from 1 (I disagree strongly) to 5 (I agree strongly).
10. The contents in this course will be helpful for me later on. Likert scale ranging from 1 (I disagree strongly) to 5 (I agree strongly).
11. Prior knowledge will help me master the course. Likert scale ranging from 1 (I disagree strongly) to 5 (I agree strongly).
12. My peers think that financial mathematics is interesting. Likert scale ranging from 1 (I disagree strongly) to 5 (I agree strongly).
13. I would be surprised to fail the financial mathematics exam. Likert scale ranging from 1 (I disagree strongly) to 5 (I agree strongly).

14. I am afraid of financial mathematics. Likert scale ranging from 1 (I disagree strongly) to 5 (I agree strongly).
15. Men are better at solving mathematical problems compared to women. Likert scale ranging from 1 (I disagree strongly) to 5 (I agree strongly).
16. It is a burden to fail this course. Likert scale ranging from 1 (I disagree strongly) to 5 (I agree strongly).
17. I would consider dropping out of university if I fail this course. Likert scale ranging from 1 (I disagree strongly) to 5 (I agree strongly).

Figure A8 – Threshold Manipulation

This figure visualizes the distribution around the threshold. Two years were omitted from this graph, the years 2011 and 2015. Only in these two years, students who marginally failed were regraded by the course instructor. Students in those two years are therefore arguably more distant in terms of their performance compared to other years. On the x-axis is the distance to the passing threshold in the financial mathematics exam. On the y-axis is the number of students. Students are binned per point difference and capped at the extreme end at -10 and 25 points respectively. Panel A shows the results only for females. Panel B shows the results only for males.

Panel A: Only Female Students



Panel B: Only Male Students