The Effectiveness of Referral Services and Student Motivation in Higher Education: An Evaluation Using Regression Discontinuity

by

Jonathan Holmes

An essay submitted to the Department of Economics
in partial fulfillment of the requirements for
the degree of Master of Arts

Queen's University

Kingston, Ontario, Canada

August, 2012

Acknowledgments

Special thanks to Steven Lehrer, Jean-Luc Daoust and Ross Finnie for access to data and project guidance. I am also grateful to participants at the EPRI session during the Canadian Economics Association's 2012 annual conference and to participants of the student lecture series at Queen's University for their comments. Thanks to the Social Sciences and Humanities Research Council of Canada, the Education Policy Research Institute, the Canadian Economics Association, and the Queen's Economics Department for funding this project and conference attendance fees. All errors and omissions are my own. The views in this paper do not necessarily reflect the views of the Social Sciences and Humanities Research Council of Canada.

Contents

1 Introduction

Student dropout from university is a lost opportunity to invest in the future livelihood of youth and the strength of the economy. While stories do exist of students dropping out of school to form companies or to take advantage of better opportunities, many students drop out because they are overwhelmed by the post-secondary experience, and fail to successfully integrate into the university community (Rickinson & Rutherford, 1996). Were these students provided additional support in the form of personal or academic counseling, they may be more likely to persist and improve their academic performance. Since the long-run return to skills accumulation is large, the net benefit of these counseling programs could be huge.

For this reason, and because universities have an incentive to retain students in order to protect revenue streams, there exist many different programs designed to increase student retention. Both experimental and quasi-experimental studies show that some student support services succeed at better integrating students and increasing retention (Lotkowski, Robbins, & Noeth, 2004) (Bailey, Robbins, & Alfonso, 2005). Still, services remain chronically underused by students, even when they are freely available (MacDonald et al., 2009). This may be because students in a large campus are unaware of what is available, or are simply reticent to make use of services which are foreign and unfamiliar. As a consequence, it may be the case that reaching out directly to students at risk and referring them to existing services will increase their use of these services, thereby decreasing their probability of dropping out and improving their learning outcomes.

In this paper, I evaluate Tracking Students at Risk, which is a program put in the place at a science faculty in a large university in Ontario, whose goal is to increase retention and academic outcomes among struggling students. This study is novel, because the program's simple structure offers clearer conclusions about the mechanisms behind the success or failure of retention programs. In contrast to the majority of recent studies which give treated students access to a comprehensive set of academic and personal supports, Tracking Students at Risk only provides additional information about existing services to students. This study thus attempts to test whether a student's lack of knowledge about campus resources is a key contributor to poor grades and dropout. In this pa-

per, I exploit the discontinuity generated by the design of the program to estimate the causal effect of the provision of information to students on their academic success and probability of retention.

This paper adds to the literature in three distinct ways. First, it lends partial support to existing studies which show that women are much more likely to benefit from student support services. Women who participate in the program appear to increase their marks more, but increase retention rates less, than men who also participate. Second, it offers suggestive evidence that the psychological effect of failing an exam does not in itself strongly affect a student's overall performance. Finally, it suggests a new way of interpreting the selection bias inherent in nonrandom studies of retention programs. Unfortunately, because both the sample size and the likely effect size are small, the tests used simply do not have the statistical power to conclude whether Tracking Students at Risk was effective or not in increasing retention or other student outcomes. The frequent treatment of students who did not qualify for the program and the discrete nature of the dataset both also served to reduce the precision of estimates of the treatment effect. With additional data, it may be possible to confirm or deny the existence of the suggestive trends identified in this paper.

The rest of this paper will be structured as follows. I introduce the literature in section 2, and will discuss existing studies to evaluate programs to increase student retention. I present the data in section 3, and will perform a simple nonrandom comparison between treatment and control in section 4. I will then discuss the regression discontinuity methodology, introducing the econometric model in section 5, and then presenting my results in section 6. In section 7, I will describe threats to the validity of the methods, and will discuss how discouragement could have played a role. In section 8, I will conclude.

2 Literature Review

While there are many studies which attempt to analyze the effect of counseling and advising programs, a consensus has not yet emerged about what types of programs are successful. Because of the large variety of retention programs and the equally varied set of estimation strategies used to evaluate them, comparing studies is hard. For this and other reasons, studies of mentoring

and counseling programs which most closely resemble Tracking Students at Risk are quite mixed. Non-experimental studies are most likely to find large increases in retention for those enrolled in their programs. In contrast, randomized trials find that most programs have small and insignificant effects on both grades and retention, except for programs which incorporate monetary incentives of some kind. There is also some evidence that women and subgroups of the population at greater risk of dropout derive greater benefit from campus services compared to the rest of the student population.

Three recent literature reviews of counseling and advising programs reach different conclusions. Lotkowski et al. (2004) conclude that certain types of orientation programs and support services are successful in increasing retention among college students. In contrast, Patton, Morelon, Whitehead, and Hossler (2006), find that there is only weak support for the ability of mentoring and counseling programs to increase retention. Finally, Bailey et al. (2005) find that some types of student support programs can improve student outcomes, with comprehensive advising programs having the greatest effect. Differences in the conclusions among these literature reviews are due partly to differing units of study (orientation programs vs. counseling programs), and partly due to different interpretations of the available literature. So, while there have been studies showing that certain kinds of counseling and advising programs successfully reduce attrition and improve other academic results, there are also many studies which are incapable of finding a positive effect.

One factor which influences why these literature reviews reach different conclusions is the weight they give to a few non-experimental studies. These non-experimental studies give first-year students at high risk of attrition access to counseling, and have shown large and significant increases in retention for treated students (see for example Rickinson (1998); Wilson, Mason, and Ewing (1997); and Turner and Berry (2000)). These studies typically compare treated students to those who qualified for treatment but did not take it up. Clearly, these studies' major weakness is their inability to control for selection bias. However, as I will later argue, there is additional information that can be gleaned from non-experimental studies if one is willing to make assumptions about what is driving the selection bias.

Experimental and quasi-experimental studies tend to find that the effect of these programs is

much smaller. For example, Angrist, Lang, and Oreopoulos (2009) find that a comprehensive program including peer mentorship and group review sessions did not have any impact on student achievement, but that a program which combined these services with financial incentives for higher grades did. MacDonald et al. (2009) also find that a Canadian mentorship program similar to Tracking Students at Risk did not raise grades or lower elective dropout among treated college students, but mandatory dropout due to low grades or poor behaviour was lower in the treatment group. Using a quasi-experimental approach, Bettinger and Long (2005) find that remedial math and English programs helped increase student retention, but this study measured the effect of a mandatory remedial course for credit rather than elective tutoring outside of class time. While studies such as these are still relatively uncommon, evidence thus far suggests that optional interventions with no monetary incentives rarely increase grades or retention to a substantial degree.

Studies that report separate results by gender and other student subgroups also tend to find substantial heterogeneity between groups. Women tend to be more likely to take-up treatment where it is optional, and also tend to benefit more from these programs (Angrist et al., 2009) (MacDonald et al., 2009). There is also suggestive evidence that the sex pairings of mentor to mentee (or of counselor to student) make a difference in treatment take-up, with students more likely to meet with a mentor of the same sex (Angrist et al., 2009). In addition, students in certain groups at a high risk of attrition are more likely to benefit from these programs than other students. For example, Angrist et al. (2009) find suggestive evidence that students at the low-end of the grades distribution benefited more from a mentorship program than other students. MacDonald et al. (2009) also find that certain vulnerable groups (low-income families, children of parents with no post-secondary education, and foreign language students) have better outcomes.

Compared to the programs reviewed in all these studies, Tracking Students at Risk is much less comprehensive. It does not offer any monetary incentives or scholarships, none of its treatments are mandatory, and it does not offer additional services such as dedicated mentors or private study groups. Since most studies find that these small-scale programs rarely have an appreciable impact on retention or other student outcomes, it would be surprising to find that Tracking Students at Risk had substantial positive results. Still, by specifically targeting students who are struggling to

succeed, this program may in fact have a greater effect than expected.

3 Program Design and Data

Tracking Students at Risk was put in place in a science faculty at a large University in Ontario in the early 2000s and has now been expanded to cover all direct-entry programs at the school. The program is designed to identify struggling students at the beginning of their first year of university. Students write their first midterm exams in October, and these marks are submitted to the faculty. All students who fail at least two exams,[1] as well as a small number of students in special circumstances, are then invited to a meeting with the Coordinator of First Year Students. These meetings can happen as early as October, and continue throughout the first semester. There are additional meetings during the second semester for those students still deemed to be at risk, although this paper will not include this second round of meetings in its analysis. During the meetings, the Coordinator of First Year Students identifies areas of weakness in the student's record, and then refers them to campus resources such as the Academic Writing Help Center or Counseling Services. Other than the meeting with the Coordinator of First Year Students, treated students do not have access to any resources which are not available to other students on campus. As a consequence, the program's goal is to reduce informational barriers and motivate students to make use of the services which are already available to them.

Data used in this study comes from three cohorts of students who entered the science faculty in the years 2008, 2009, and 2010. Summary statistics for those students who qualified for the program by failing two or more midterms are listed in table 2. Midterm grades, which are all given a percentage mark, were provided by professors of core science courses taken by students in the first semester of their program.[2] Students final grades, which are scored on a ten point scale, as well as

[1] The passing grade for midterm exams in all classes is 50%. Students must get a mark of strictly less than 50% in order to fail.

[2] Some professors either failed to respond to a request for their midterm grades or refused to provide these grades. Midterm exams from these courses are treated as missing values. There are 159 students with missing biology midterms, 610 with missing chemistry midterms, 855 with missing math midterms, and 220 with missing physics midterms. Since the department also did not have access to these grades, they did not play a role in determining treatment status. As a result, the relevant variable to determine treatment status is the second lowest observed midterm mark, rather than a student's actual second lowest midterm mark.

Table 1: Summary Statistics

	2008/09			2009/10			2010/11		
	Mean	SD	OBS	Mean	SD	OBS	Mean	SD	OBS
Entry Age	18.2	1.80	855	18.1	1.81	910	18.3	2.89	934
Female	0.55	0.50	855	0.56	0.50	910	0.55	0.50	934
Local Student	0.53	0.50	855	0.59	0.49	908	0.59	0.49	934
English Speaker	0.65	0.48	855	0.62	0.48	910	0.60	0.49	934
French Speaker	0.18	0.38	855	0.19	0.39	910	0.21	0.40	934
Other Mother Tongue	0.18	0.38	855	0.18	0.39	910	0.20	0.40	934
Admission Average	83.5	7.19	850	83.5	6.98	903	83.0	6.99	931
Lowest Midterm Mark	53.5	22.9	808	53.2	17.4	647	52.8	19.6	872
Second Lowest Midterm Mark	67.8	18.1	724	66.8	16.6	549	68.5	17.0	759
Third Lowest Midterm Mark	79.2	16.3	471	75.4	14.8	312	77.3	15.4	471
Biology Midterm Grade	64.3	17.5	734	58.1	15.1	607	62.7	14.9	791
Chemistry Midterm Grade	63.8	21.0	603	70.5	19.4	412	57.4	23.0	560
Math Midterm Grade	70.5	26.6	541	64.1	20.3	390	77.2	17.9	545
Physics Midterm Grade	66.0	27.8	229	64.6	21.1	166	62.7	23.0	319
Biology Final GPA	5.66	2.59	600	5.27	2.58	747	5.24	2.60	796
Chemistry Final GPA	5.68	2.95	588	5.77	2.80	768	5.72	2.86	796
Math Final GPA	4.84	3.41	712	4.49	3.31	774	5.72	3.37	816
Physics Final GPA	5.40	2.92	268	5.27	2.95	304	5.13	2.86	322
Met With CFYS	0.092	0.29	855	0.095	0.29	910	0.086	0.28	934
Referred to Math Tutoring	0.041	0.20	855	0.063	0.24	910	0.037	0.19	934
Referred to Physics Tutoring	0.028	0.17	855	0.024	0.15	910	0.034	0.18	934
Referred to Chemistry Tutoring	0.043	0.20	855	0.024	0.15	910	0.020	0.14	934
Referred to Biology Tutoring	0	0	855	0	0	910	0	0	934
Referred to School Orientation	0.016	0.13	855	0.0077	0.087	910	0.019	0.14	934
Referred to Counselling	0.0035	0.059	855	0.011	0.10	910	0.012	0.11	934
Referred to Mentorship Program	0.044	0.21	855	0.040	0.20	910	0.046	0.21	934
SGPA	5.61	2.55	855	5.38	2.50	899	5.63	2.54	913
CGPA	5.64	2.52	855	5.39	2.49	899	5.52	2.44	896
Half Year Retention	0.97	0.17	855	0.96	0.21	910	0.96	0.20	934
1 Year Retention	0.87	0.34	855	0.86	0.34	910	0.88	0.33	934
2 Year Retention	0.80	0.40	855	0.79	0.41	910	.	.	0
3 Year Retention	0.74	0.44	855	.	.	0	.	.	0

all other demographic information, were obtained from the department. SGPA is the average GPA of all courses taken in first semester (Semester GPA), wheras CGPA is the average of all courses taken in the first year of university (Cumulative GPA). Summary statistics for the entire sample of students in the dataset are listed in table 1.

In order to increase precision, all estimates in this paper will be calculated using a pooled sample of all cohort years. Ex ante, I don't expect much heterogeneity in treatment effects across cohorts because the structure of the program stayed the same during the entire study period. In particular, the Coordinator of First Year Students was the same male individual across all three years of the study.[3] In addition, there were no major institutional changes to the programs which students were referred to, such as the Math Help Center, during the three years being studied. Summary statistics also do not appear to show any major changes in the observable characteristics of the sample in each year.

4 Non-Experimental Analysis

While all students failing at least two midterm exams were contacted and invited to a meeting with the Coordinator of First Year Students, many of these students did not accept the offer. As a consequence, it is possible to do an analysis similar to Rickinson (1998), Wilson et al. (1997), and Turner and Berry (2000). To do this, I will present results from a simple comparison of means between students who had a meeting and those who qualified for the program but did not take up treatment.

Tables 2 and 3 show that among those who qualified for the program, characteristics which predate students entry into university are not very different between those who took up treatment and those who did not. Table 2 shows that the difference of mean values of characteristics such as age or mother tongue are not significant at the 5% level or below. Moreover, among midterm grades, only biology is also significantly higher among treated students. This lack of significance is partly caused by the small sample size of only 292 students who qualified for the program, which

[3] At least one study has found that students are more likely to respond to requests to meet from a mentor when the mentor was the same sex (Bettinger & Long, 2005).

Table 2: Results from t-tests comparing mean characteristics of treated students to mean characteristics of students who qualified but were not treated

	No Meeting		Meeting		Difference
	mean	number	mean	number	
Age at Program Start	18.44	155	18.20	137	-0.242 (0.253)
Gender (Female == 1)	0.529	155	0.635	137	0.106 (0.0578)
Out-of-City Student	0.471	155	0.526	137	0.0546 (0.0587)
Admission Average	77.73	154	78.07	136	0.344 (0.653)
Biology Midterm Grade	40.76	150	45.26	134	4.503** (1.585)
Biology Final GPA	2.135	141	2.680	122	0.546* (0.213)
Chemistry Midterm Grade	36.71	125	38.99	103	2.274 (2.439)
Chemistry Final GPA	2.393	140	2.857	112	0.464 (0.253)
Math Midterm Grade	43.17	114	42.82	100	-0.351 (3.558)
Math Final GPA	1.442	138	1.400	110	-0.0420 (0.254)
Physics Midterm Grade	40.29	59	38.58	49	-1.706 (3.601)
Physics Final GPA	2.885	52	2.872	47	-0.0123 (0.408)
SGPA	2.501	155	2.896	135	0.396* (0.197)
CGPA	2.597	154	3.092	133	0.496* (0.195)
Half Year Retention	0.929	155	0.949	137	0.0199 (0.0283)
1 Year Retention	0.703	155	0.825	137	0.122* (0.0497)
2 Year Retention	0.495	105	0.782	87	0.286*** (0.0674)
3 Year Retention	0.411	56	0.711	45	0.300** (0.0961)

Standard Errors in Parenthesis

Note: Sample includes students who qualify for the program by failing at least two midterms in their first semester

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Linear regression of treatment status on covariates which predate university attendance

	(1) Met With CFYS	(2) Met With CFYS
Entry Age	-0.00700	-0.00424
	(0.0148)	(0.00261)
Female	0.112	0.0452***
	(0.0607)	(0.0110)
Out-of-City Student	0.0598	0.0222*
	(0.0602)	(0.0112)
English Speaker	0.0313	0.0110
	(0.0758)	(0.0147)
French Speaker	0.0830	-0.00815
	(0.114)	(0.0180)
Admission Average	0.00110	-0.00906***
	(0.00538)	(0.000789)
Constant	0.386	0.883***
	(0.509)	(0.0889)
R^2	0.0193	0.0524
N	290	2682

Standard errors in parentheses

Notes:

-Sample in (1) includes all students who qualify for the program by failing at least two midterms in first semester

-Sample in (2) includes all students who have a second lowest midterm mark in the first semester

-Standard errors are heteroskedasticity consistent.

-CFYS means Coordinator of First Year Students

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

reduces the precision of the estimates of the mean. In particular, sample size is the reason why the substantial ten point difference in the percentage of women in each group is not significant at the 5% level. Nevertheless, table 3 shows underlying characteristics have a very low explanatory power with respect to the probability of a student meeting with the Coordinator of First Year Students. In the linear probability model which is restricted to those students who technically qualify for the program (column 1 of table 3), a test of overall significance fails to reject the null hypothesis that the coefficients have no explanatory value with a p-value of 0.4737. So, while women appear to be somewhat more likely to take up treatment than men, pre-program observables are generally poor predictors of treatment take-up among students who qualify for the program.

While students who have a meeting have observable characteristics which are fairly similar to those who do not, treated students have outcomes which are substantially better than those who chose not to have a meeting. Treated students have a slightly higher GPA after one semester, and their GPA advantage rises to 0.5 points or half a letter grade after one year. While higher grades are partly explained by higher midterm marks, this does not explain the entire GPA gap between treatment and control. Treated students also have a retention rate which is a full 15 percentage points higher in the treatment group after one year, and which rises to 30 points after three years. The fact that the retention rate actually rises is striking, and suggests that some feature of treated students has a lasting impact on their behavior. Something is obviously helping these students achieve greater success than their peers.

Non-experimental studies could use the fact that students who take up the program are similar to those who do not as evidence that the groups are as good as randomized. They would see the substantial differences in outcomes as the causal effect of the program. Nevertheless, I believe there is good reason to disbelieve this logic because of selection bias. In my view, there are three possible reasons for selection bias to occur. First, students who have more knowledge and who are more talented at taking tests may be more likely to take up treatment. Second, students who take up treatment may be more motivated to improve their results, either by working harder or by seeking out additional help elsewhere. Finally, students who take up treatment may get better results for reasons unrelated to motivation or inherent skill.

These causes of selection bias are difficult to study because they cannot be measured directly, and because they are often strongly interrelated. Nevertheless, the idea that internal motivation is the most important component of selection bias has strong intuitive support. Whether a more gifted student would demand more or less additional help is ambiguous, because this would depend on the student's perception of the returns to having a meeting with the Coordinator of First Year Students. Moreover, it is likely that the portion of selection bias due to reasons unrelated to motivation and skill is small, because the effect of external factors on a student is often dwarfed by the student's response to these factors. For example, while attending a meeting may be related to having additional free time to study which would increase marks, less motivated students are less likely to make use of this free time for academic purposes. In contrast, a student who is motivated to improve his or her exam scores clearly has a much greater propensity both to respond to a request to meet with a school counselor and also to spend time to improve his or her grades through other means. The magnitude of the selection bias depends on the extent that students are able to increase their own marks.

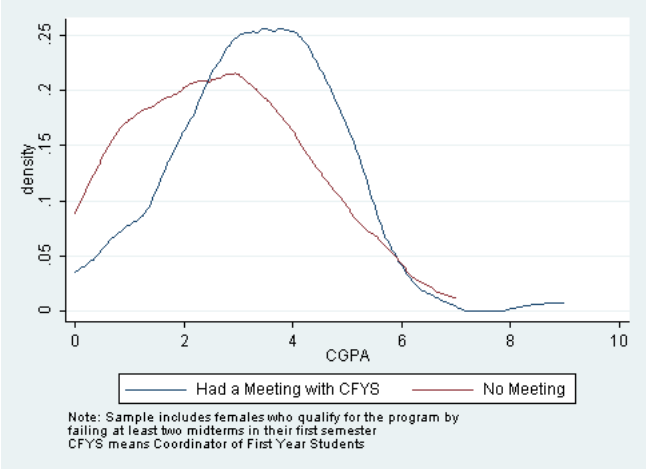
Empirical tests support the idea that inherent skill does not appear to fully explain the difference in GPA between treatment and control groups, which suggests that either the program itself or the motivation of students explains the rest. High school grades, which are a proxy for inherent talent, are not significantly different between those who enroll in the program and those who did not. While midterm grades among those who are treated are slightly higher, the difference in these grades is smaller than the overall difference in GPA at the end of the semester and year.[4] As a consequence, the total measured difference in GPA between those who had a meeting and those who did not cannot be fully explained by underlying ability, and thus must be explained by a combination of a student's underlying motivation to do better as well as the treatment effect of the program.

This characterization of selection bias has two consequences. First, assuming inherent skill

[4]Midterm grades in biology in particular are higher among students who took up treatment. This difference explains some, but not all, of the difference in final grades. To test this, I calculated projected final grades for each subject based on current midterm grades by course. The difference in final GPAs is larger than the difference in projected GPAs between treatment and control, which suggests that treated students had some additional advantage. Note that this analysis is not perfect, because I did not have access to midterm and final grades across all courses that students took (particularly courses taken outside the science faculty). If a student's biology midterm grade was a more accurate predictor of inherent ability in these other courses than their other science midterm marks, then the overall difference in GPAs between treatment and control may be completely due to inherent skill.

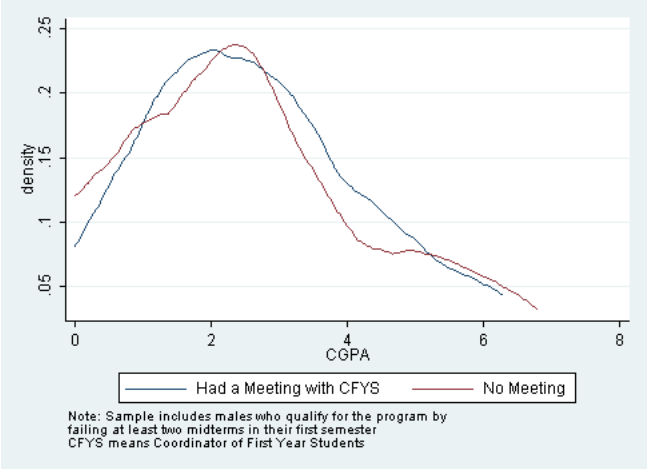
and motivation are the primary causes of selection bias, selection bias should be positive. This statement relies on the assumption that motivation cannot hurt a student's marks, and the fact that the estimated marks of students prior to the program tend to be slightly higher among treated students. As a consequence, non-experimental estimates of the effect of the program represent the maximum possible net benefit to students of Tracking Students at Risk. Second, assuming that factors unrelated to motivation and inherent skill play a negligible role in the overall selection bias, then once both the effect of the program and the effect of initial skill levels have been controlled for, non-experimental results are a good proxy for the effect that motivation plays on academic success.

Figure 1: Density of Cumulative GPA scores among females who qualified for the program



Understanding the sources of selection bias is important when contrasting the positive results of women to the lackluster results of men who had a meeting with the Coordinator of First Year Students. Figure 1 shows the distribution of year-end GPA results of females who attended and did not attend a meeting with the Coordinator of First Year Students (CFYS) among among those who qualified for the program. It shows that those who enrolled in the program had greater average grades, and this difference exists over the entire grades distribution. So, females appear to either have benefited from the program, or they were able to improve their academic scores to some extent through their additional motivation. In contrast, figure 2 shows that among males who qualified for the program, those who had a meeting had almost the same distribution of final grades as those

Figure 2: Density of Cumulative GPA scores among males who qualified for the program



who did not. This implies that neither the program nor the intrinsic motivation which pushed them to attend the meeting were successful in raising GPA results.

Parametric results of t-tests comparing treatment and control groups in tables 4 and 5 show that not only are men less likely to attend a meeting, but males who attended a meeting had statistically identical GPA results to males who did not. In contrast, having a meeting with the Coordinator of First Year Students was associated with higher retention levels among both males and females. This supports the idea that students attending meetings were more likely to be motivated and committed to succeeding at University. While having a meeting was associated with a higher increase in retention among males, this is partly due to the fact that males were increasing their retention results form a much lower base.

These statistics suggest, but do not prove, that males have a harder time transforming a motivation to succeed into improved GPAs. There are a number of possible problems with this analysis. First of all, extremely small sample sizes increase the probability that results are driven by a few outliers, and do not represent a general trend. In addition, the relationship between motivation and the probability of meeting with the Coordinator of First Year Students may be weaker than I implied. It is possible that the probability of attending a meeting for males was associated only with an individual's connection to the school, which would result in students who had a meeting

Table 4: Female treatment/control comparison

	No Meeting		Meeting		Difference
	mean	number	mean	number	
Entry Age	18.22	82	18.24	87	0.0219 (0.329)
Female	1	82	1	87	0 (0)
Out-of-City Student	0.415	82	0.494	87	0.0796 (0.0769)
Admission Average	78.67	82	78.48	87	-0.190 (0.848)
Biology Midterm Grade	39.28	78	45.53	86	6.259** (1.946)
Biology Final GPA	2.149	74	2.935	77	0.786** (0.283)
Chemistry Midterm Grade	38.33	67	41.00	66	2.670 (3.088)
Chemistry Final GPA	2.649	74	3.014	70	0.366 (0.327)
Math Midterm Grade	46.38	63	43.47	66	-2.909 (4.406)
Math Final GPA	1.377	77	1.357	70	-0.0195 (0.311)
Physics Midterm Grade	38.92	27	36.81	30	-2.112 (3.766)
Physics Final GPA	2.808	26	2.769	26	-0.0385 (0.485)
SGPA	2.515	82	3.109	86	0.595* (0.245)
CGPA	2.674	82	3.346	85	0.671** (0.242)
Half Year Retention	0.939	82	0.943	87	0.00350 (0.0365)
1 Year Retention	0.732	82	0.839	87	0.107 (0.0628)
2 Year Retention	0.585	53	0.768	56	0.183* (0.0886)
3 Year Retention	0.500	28	0.677	31	0.177 (0.128)

Standard Errors in Parenthesis

Note: Sample includes females who qualify for the program by failing at least two midterms in their first semester

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5: Male treatment/control comparison

	No Meeting		Meeting		Difference
	mean	number	mean	number	
Entry Age	18.68	73	18.12	50	-0.565 (0.401)
Female	0	73	0	50	0 (0)
Out-of-City Student	0.534	73	0.580	50	0.0458 (0.0919)
Admission Average	76.66	72	77.35	49	0.694 (1.019)
Biology Midterm Grade	42.36	72	44.77	48	2.408 (2.695)
Biology Final GPA	2.119	67	2.244	45	0.125 (0.324)
Chemistry Midterm Grade	34.84	58	35.39	37	0.549 (3.977)
Chemistry Final GPA	2.106	66	2.595	42	0.489 (0.400)
Math Midterm Grade	39.22	51	41.57	34	2.354 (6.051)
Math Final GPA	1.525	61	1.475	40	-0.0496 (0.435)
Physics Midterm Grade	41.44	32	41.38	19	-0.0603 (6.592)
Physics Final GPA	2.962	26	3	21	0.0385 (0.685)
SGPA	2.485	73	2.522	49	0.0375 (0.329)
CGPA	2.508	72	2.644	48	0.135 (0.323)
Half Year Retention	0.918	73	0.960	50	0.0422 (0.0455)
1 Year Retention	0.671	73	0.800	50	0.129 (0.0819)
2 Year Retention	0.404	52	0.806	31	0.403^{***} (0.105)
3 Year Retention	0.321	28	0.786	14	0.464^{**} (0.151)

Standard Errors in Parenthesis

Note: Sample includes males who qualify for the program by failing at least two midterms in their first semester

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

having higher retention rates while not necessarily having higher academic marks. Because of these caveats, and particularly because of the heavy reliance on the assumption that motivation drives selection bias, these results should merely be taken as suggestive.

This analysis both supports and refutes previous studies showing that males are much less likely to benefit from support programs than females. On the one hand, males in the program had a much higher retention rate than those who were not, indicating either the program helped to keep them in school or that they were able to translate a desire to persist into actual persistence. On the other hand, males were not able to improve their own GPA scores. Perhaps it is not only the case that these traditional types of programs are ineffective at raising male outcomes, but also that males are less likely to be able to use the tools available to them in order to raise their own outcomes. If this is indeed true, additional research should focus on what the barriers are which prevent male success.

5 Methodology

The structure of Tracking Students at Risk allows for the use of regression discontinuity because of the nature of the assignment rule. By ordering each student's midterm results from lowest to highest score, students who qualify for the program are those who score strictly less than 50% on their second lowest exam. As long as students cannot manipulate their exam marks perfectly, students who score barely below and barely above the 50% cutoff are fairly similar, and it is possible to treat this assignment rule as a form of randomized experiment.

5.1 Econometric Model

Suppose the constant effects model for treatment, with a very general functional form for all other possible regressors.[5] In equation 1, Y is the outcome variable (GPA or retention probability), X is the assignment variable (second lowest exam mark), $\mathbf{Z}$ is a matrix of all possible other variables

[5]The constant effects assumption is not necessary, but it helps to simplify the exposition. When the constant effects assumption is dropped, the estimator for treatment effect is actually a weighted average treatment effect over the entire population, where weights are determined by the probability that an individual's test score falls near the 50% threshold. For a full explanation, see Lee and Lemieux (2009, 19-21).

which could affect treatment, c is the treatment threshold (50%), and T is equal to one if the person was treated (had a meeting) and zero otherwise.

$$Y_i = \tau T_i + f(X_i) + g(\mathbf{Z}_i) + \epsilon_i \tag{1}$$

In this equation, both functions $f(X_i)$ and $g(\mathbf{Z}_i)$ are general functions with no restrictions except for what follows. Assume first that there exists a discontinuity in the assignment rule at the treatment threshold, which means $\lim_{X \to c^+} T_i \neq \lim_{X \to c^-} T_i$. Then, assume that the the expectation $E[Y_i|T_i = 0]$ is continuous in X_i at the point of discontinuity. Then, by the definition of continuity:

$$\lim_{X \to c^-} E(Y_i|T_i = 0) = \lim_{X \to c^+} E(Y_i|T_i = 0) \tag{2}$$

In the constant effects framework, this is all that is necessary to assume in order to identify the treatment effect. Under heterogeneous treatment effects, it is also necessary to assume the local randomization of test scores in the neighbourhood of the discontinuity (c=50%). This assumption implies that individuals cannot control the marks they get on tests exactly in the neighbourhood of the cut point, and is necessary to ensure that the treatment and control populations are comparable.

Under these assumptions, it is possible to identify and estimate the treatment effect using either equation 3 or equation 4 (Hahn, Todd, & Van der Klaauw, 2001). Equation 3 gives the case of a sharp regression discontinuity, where $\lim_{X \to c^+} T_i = 1$ and $\lim_{X \to c^-} T_i = 0$. Then, the treatment effect is:

$$\tau^{ITT} = \lim_{X \to c^+} Y_i - \lim_{X \to c^-} Y_i \tag{3}$$

The fuzzy regression discontinuity framework makes the more flexible assumption that $\lim_{X \to c^+} T_i \neq \lim_{X \to c^-} T_i$. Then, the treatment effect is:

$$\tau^{LATE} = \frac{\lim_{X \to c^+} Y_i - \lim_{X \to c^-} Y_i}{\lim_{X \to c^+} T_i - \lim_{X \to c^-} T_i} \tag{4}$$

When analyzing a treatment with imperfect compliance, the sharp regression discontinuity design performs a local intent to treat analysis. In this case, the local intent to treat effect is the effect of qualifying for Tracking Students at Risk, regardless of whether students actually attend the program or not. The fuzzy regression discontinuity design measures the local average treatment effect of the program. Under the constant effects framework, the local average treatment effect is equal to the average treatment effect. Under heterogeneous treatment effects, it measures the effect of the program on the students who had a meeting with the Coordinator of First Year Students because they barely fell below the threshold to qualify, but who would not have had a meeting otherwise. These students are called compliers, and student i is a complier if $\lim_{X \to c^-} T_i = 1$ but $\lim_{X \to c^+} T_i = 0$. The group of compliers is policy relevant, because it represents students who are struggling and are particularly at risk of dropping out.

5.2 Empirical Model

In practice, since all real-world datasets are finite, researchers cannot use observations which are arbitrarily close to a point of discontinuity in order to estimate a treatment effect, and it is always necessary to use observations which are further away in order to identify both the left side and right side limits. To include these observations, it is necessary to make additional functional form assumptions, which may bias results if untrue. The regression discontinuity literature presents two main methods for minimizing this possible bias: local linear regression, and a smooth polynomial regression.

Local linear regression estimates the size of the treatment effect by fitting a linear function using only observations within a certain bandwidth of the cut point. This sacrifices precision in order to reduce the bias due to misspecification of the functional form. With large datasets and small bandwidths, the bias due to misspecification is small regardless of the true functional form underlying the data.

Unfortunately, it is impractical to use local linear regression functions in this dataset for two reasons. First, the dataset is relatively small, with only 312 observations that have a second lowest midterm score between 40% and 60%. As a consequence, it is necessary to use a bandwidth that is

large enough so as to make the assumption of local linearity much less plausible. More importantly, the assignment variable in this case clearly does not have continuous support. When marking an exam, professors and TAs generally give scores at unit integers or at other convenient fractions. So, while there are 1146 unique students in the dataset who have at least two nonzero exam marks, the assignment variable (the second lowest midterm mark) only has 378 unique observations, which is just over 3 observations per heap. Within ten marks of the cut point, there are 312 observations clustered in only 100 unique heaps, which gives a slightly higher 3.12 observations per heap. As explained by Lee and Card (2008), discrete assignment variables undermine both the theoretical support for and the practical application of local linear regression.

There is one more problem with using local linear regression analysis. In practice, each class uses a different marking scheme, which implies that each class will tend to heap at different unit integers. For example, if the physics midterm is marked out of 5 and the biology midterm is marked out of 10, then physics midterm marks will be heaped at at either 0 %, 20%, 40%, 60%, 80%, or 100%, whereas the biology students will be heaped at unit multiples of 10%. Should students whose second worst midterm was biology be systematically different from students whose second worst midterm was physics, then the assumption that $\lim_{X \to c^-} E(Y_i|T_i = 0) = \lim_{X \to c^+} E(Y_i|T_i = 0)$ may be violated, and local linear estimates of the discontinuity will be biased (Barreca, Lindo, & Waddell, 2011).[6] There are two main reasons for this bias, which I solve by explicitly modeling a parametric function on either side of the cut point. First, using a sample of observations within a small bandwidth of the cut point can ignore heaps which fall just outside, putting too much weight on heaps that are included. By setting a bandwidth which uses all available data, this problem is avoided. The second problem is that functional form assumptions within the estimated bandwidth may be wrong. To solve this problem, I incorporate modeling error into the estimated standard errors of the model, a process which will be described later.[7]

[6] I am slightly abusing mathematical notation here. Technically, any discrete-valued function will always satisfy $\lim_{X \to c} f(x) = f(c)$ if c is in the domain of f, and the limit does not exist if c is not in the domain of f. This is because all Cauchy sequences in the discrete domain of f which converge to c cannot get arbitrarily close to c without hitting c exactly. The point I am trying to make is that different classes may have different outcomes, and this might bias the function if not controlled for in some way.

[7] Another option for dealing with this problem is to estimate the model with a separate intercept and slope coefficients for each subject to control for possible differences in functional form. So far, I have only seen this method recommended for local linear estimates of regression discontinuity, and only in cases where there is a mixture of both

Since a local linear function is both theoretically invalid and impractical given the dataset, I will instead attempt to model $f(X)$ from equation 1 explicitly using a flexible polynomial. The econometric model I will be estimating is listed in equation 5. In this equation, τ is the estimated treatment effect, D is a dummy variable such that $D = I(X < c)$,[8] n is the order of the polynomial, and both α and $\beta_{p,j}$ are regression coefficients.

$$Y_i = \alpha + \tau D_i + (1 - D_i) \sum_{p=1}^{n} (\beta_{p,1} X_i^p) + D_i \sum_{p=1}^{n} (\beta_{p,2} X_i^p) + \epsilon_i \tag{5}$$

The estimated intent to treat effect $\widehat{\tau^{ITT}}$, is equal to $\hat{\tau}$ from equation 5. The probability of receiving treatment is calculated in equation 6 using the same method. In this equation, δ is the estimated change in treatment status at the cut point, and both γ and $\omega_{p,j}$ are regression coefficients.

$$T_i = \gamma + \delta D_i + (1 - D_i) \sum_{p=1}^{n} (\omega_{p,1} X_i^p) + D_i \sum_{p=1}^{n} (\omega_{p,2} X_i^p) + \epsilon_i \tag{6}$$

Then, the estimated local average treatment effect is:

$$\widehat{\tau^{LATE}} = \frac{\hat{\tau}}{\hat{\delta}} \tag{7}$$

Note that equation 7 is equivalent to estimating an exactly identified instrumental variables regression, where D_i is an instrument for T_i. In this paper, I will use the generalized method of moments for all of my regression estimates.

The critical parametric assumption in these formulas is the choice of n, which is the number of polynomial terms in each regression. Choosing the wrong n will bias the results due to misspecification. In practice, there are a number of ways of validating the choice of n, all of which rely to some extent on measuring the overall goodness of fit of the regression. However, all of these methods are imperfect, because the only requirement of local polynomial regression is that the model is well

discrete and continuous data. Because my entire dataset is discrete, I have elected to use standard errors clustered on each discrete value of the running variable, which is a more general way of modeling specification error.

[8] I is in indicator function that is equal to one if X is less than c, and zero otherwise

fit at the cut point c. It may be the case that a model with a good overall fit would estimate the extreme value c with more error than a model with a poorer overall fit. As a consequence, all regression discontinuity estimates in this paper will be tested for robustness using functional forms with varying numbers of polynomial terms.

In this paper, I will validate my choice of n using a non-parametric approach suggested in Lee and Lemieux (2009, pp. 45-48). This method tests whether the chosen polynomial fits the data better than a non-parametric alternative. The test works by dividing the assignment variable into K bins of equal size, each of which is represented by a represented by a bin dummy B_k. These dummies are then added to the model in equation 8.

$$Y_i = \alpha + \tau D_i + (1 - D_i)\sum_{p=1}^{n}(\beta_{p,1}X_i^p) + D_i\sum_{p=1}^{n}(\beta_{p,2}X_i^p) + \sum_{k=2}^{K-1}\psi_k B_k + \epsilon_i \tag{8}$$

Note that two bin dummies are excluded because of collinearity with the constant and τ. To test whether or not a model that is strictly more flexible than the parametric model has more explanatory power, it is sufficient to test the joint hypothesis that $\psi_2 = \psi_3 = ... = \psi_{K-1} = 0$. Since Lee and Lemieux (2009) do not provide any guidance as to the number of bins to use, I have chosen to use 20 bins total, because gaps of 5% are the most economically meaningful.[9] In each of the regressions, I have included a row with the p-value of the test of the null hypothesis that the added coefficients do not add any explanatory power to the model.

For specification testing, standard errors must take into account both the stochastic error component and the potential for modeling error in the choice of $f(x)$. To simplify the analysis, I will assume that both components of the error term are random, and that specification error at the cut point is identical for both treatment and control. Under these assumptions, the specification error can be treated as a random effect, and standard errors clustered on each unique value of the forcing variable are consistent (Lee & Card, 2008). All standard errors reported in this paper for

[9] As mentioned before, raising one's mark by 5% represents an increase in letter grade at all points in the grade distribution except when a student scores either above 90%, which is an A+, or below 50%, which is a failing grade. I am assuming that students who have the same projected letter grade coming out of their first midterm are more likely to be subject to common shocks than students whose midterm exam marks are similar but who have a different projected final grade.

regression discontinuity are clustered.

In practice, the assumptions underlying clustered standard errors may be unrealistic. In particular, the assumption that the specification error at the cut point is identical for treatment and control is unlikely given I am using a flexible functional form on either side of the cut point. As a consequence, standard errors estimated in this paper are a lower bound for true standard errors which take into account all forms of modeling error.

6 Results

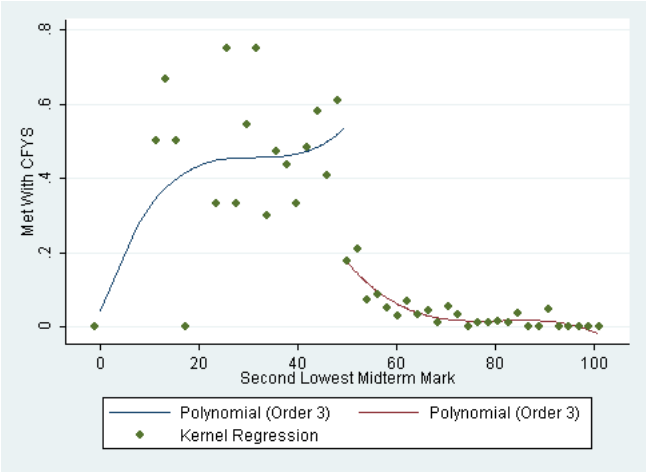
In this section, I will present results for regression discontinuity estimates of the effect of Tracking Students at Risk, in order to compare them to the simple averages found in section 4. I will begin with estimates of the size of the discontinuity of the treatment variable at the cut point. I will then show results for both GPA and retention. As recommended by Imbens and Lemieux (2008), I will present both graphical and empirical results for each section.

6.1 Treatment Variable Discontinuity

Figure 3 shows that the assignment rule was not perfectly followed, but that the treatment variables is discontinuous at the threshold value of 50%. The points on the graph are local averages of the proportion of students who met with the Coordinator of First Year Students calculated using a kernel regression with a rectangular kernel.[10] Plotted over these points is a third-order polynomial which is estimated using equation 6. Perfect compliance below the threshold was not achieved because, while all students who qualified for the program were invited to come meet with the Coordinator of First Year Students, many did not actually respond. In addition, some students who did not fail two midterms were also invited to a meeting at the discretion of the Coordinator of First Year Students. As a consequence, a fuzzy regression discontinuity design will be used for all future calculations of the treatment effect of the program.

[10] As recommended by (Lee & Lemieux, 2009, pp.31) the kernel bandwidth is set such that bins are both collectively exhaustive and non-overlapping.

Figure 3: Fraction meeting with Coordinator of First Year Students by their second lowest exam score



The existence of a discontinuity is confirmed in table 6, which shows parametric results from estimating equation 6. As a rule of thumb, the order of the polynomial to use is the lowest order polynomial to pass the test against a strictly more flexible functional form (Lee & Lemieux, 2009, pp. 45-48). In this case, this is the second-lowest order polynomial, which estimates that a student who barely fails their second worst midterm have a probability of seeing the Coordinator of First Year Students of about 36% higher than a student who barely passes their second worst midterm. This statistically significant result is robust to the inclusion of additional higher order polynomial terms.

In table 7, I have divided the sample into subgroups in order to see which students were most strongly induced to have a meeting with the Coordinator of First Year Students. I have omitted results for a linear functional form because it was rejected in the full sample. Just like in the aggregate statistics, women were much more likely to be induced to have a meeting than men were. The act of failing their second exam raised the probability of having a meeting for women by 50% but only 25% for men. Students whose primary language was neither English nor French were also much more likely to be have a jump in treatment.[11] No other differences by group were robust to

[11] However, this fact is hard to interpret, because those who list their primary language as bilingual (French/English) are a substantial group who are lumped together with foreign language speakers.

Table 6: The effect of qualifying for the program on the probability of meeting with the Coordinator of First Year Students

	(1) Met With CFYS	(2) Met With CFYS	(3) Met With CFYS	(4) Met With CFYS
Treatment Effect	0.439***	0.358***	0.369***	0.323***
	(0.0396)	(0.0514)	(0.0653)	(0.0775)
Constant	0.0968***	0.141***	0.174***	0.192***
	(0.0149)	(0.0217)	(0.0253)	(0.0269)
Bandwidth	50	50	50	50
Observations	2030	2030	2030	2030
Fuzzy	NO	NO	NO	NO
Polynomial Order	1	2	3	4
P-Value of Test†	0.0000722	0.507	0.521	0.552

Standard errors in parentheses

Treatment is defined as having failed the midterm with the second lowest mark

Standard errors are clustered on each discrete value of the forcing variable.

† Test is against H0: This functional form is correct. See equation 8.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

different functional forms.

Table 7 also shows that certain groups, due to small sample sizes and high variance, have a first-stage treatment effect which is very weak. This problem is different from the attenuation bias caused by weak instruments in small samples. Because all slope coefficients for the first-stage regression are also used as excluded instruments, the first stage regression in two-stage least squares is always strong. Instead, a small coefficient estimate of $\hat{\delta}$ will tend to increase the estimated local average treatment effect $\widehat{\tau^{LATE}} = \frac{\hat{\tau}}{\hat{\delta}}$ in absolute value. As a consequence, any error in the first-stage of the regression will be magnified. While all results for second-order polynomials of the first stage are significant, higher order polynomial results for many of the categories (particularly for French speakers and students over 19) are both small and also insignificant at the 5% level. This error magnification should not cause a problem for inference, because standard errors will also increase, but it reduces the precision of point estimates of the mean.

Table 7: The effect of qualifying for the program on probability of having a meeting - Subgroups

	Second Order Polynomial		Third Order Polynomial		Fourth Order Polynomial	
	Treatment Effect	OBS	Treatment Effect	OBS	Treatment Effect	OBS
Males	0.287***	883	0.248**	883	0.222*	883
	(0.0724)		(0.0857)		(0.110)	
Females	0.424***	1147	0.476***	1147	0.430**	1147
	(0.0805)		(0.106)		(0.132)	
Sudents Under 19	0.348***	1764	0.394***	1764	0.318***	1764
	(0.0586)		(0.0717)		(0.0853)	
Students 19 or Older	0.455***	266	0.333	266	0.242	266
	(0.129)		(0.179)		(0.241)	
Local Students	0.398***	1191	0.396***	1191	0.314**	1191
	(0.0789)		(0.0930)		(0.116)	
Out-of Town Students	0.331***	839	0.348**	839	0.353*	839
	(0.0820)		(0.115)		(0.150)	
English Speakers	0.326***	1385	0.353**	1385	0.233	1385
	(0.0827)		(0.110)		(0.137)	
French Speakers	0.341*	258	0.276	258	0.434	258
	(0.145)		(0.190)		(0.255)	
Other Language Speakers	0.503***	387	0.459**	387	0.551***	387
	(0.114)		(0.153)		(0.163)	

Treatment is defined as having qualified for the program by failing two exams.

Standard errors are clustered on each discrete value of the forcing variable.

Standard errors in parenthesis

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6.2 Effect on GPA

As shown in Section 4, students who had a meeting with the Coordinator of First Year Students had a GPA which is a half letter grade higher than students who did not have a meeting but who still qualified for the program. Because it measures the causal impact of the program, regression discontinuity provides a unique chance to quantify the extent that this difference is due to Tracking Students at Risk (at least for those students who fall near the treatment threshold), and by extension, the extent that this difference is due to substitution bias. Regression discontinuity estimates of the effect of the program were not significant under any of the alternate model specifications. While this gives suggestive evidence that the program was ineffective at raising student GPAs, definitive results remain elusive because of the small sample size combined with the low power of the regression discontinuity methodology,

Figure 4: Cumulative GPA after first year of studies as a function of second lowest exam score

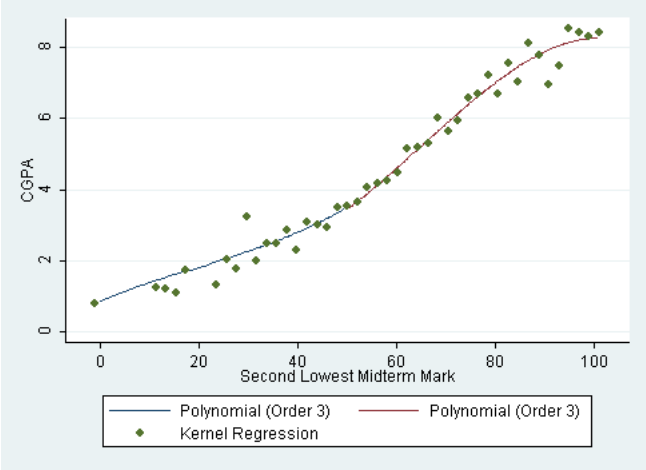


Figure 4 graphically shows the relationship between a student's second lowest midterm mark and their cumulative GPA.[12] The graph format is identical to figure 3. If it were the case that, at the margin, the program was very effective in raising student achievement, we would expect that students who scored just below 50% on their second lowest midterm to have a slightly higher GPA

[12]For this entire section, I show results only for cumulative, or end-of-year, GPA. Results for sessional GPA, which is the GPA at the end of the fall semester, are quite similar.

than those who scored just above 50% on their second lowest midterm. However, no such gap is present, to the extent that a flexible polynomial which is allowed to be different on both sides of the cut point appears to be continuous to the naked eye. As a consequence, this graph appears to support the null hypothesis that, at least for students close to the cut point, the program did not raise their academic outcomes.

Table 8: The Local Average Treatment Effect of the program on the cumulative GPA after one year

	(1) CGPA	(2) CGPA	(3) CGPA	(4) CGPA
Treatment Effect	-0.521	0.574	0.136	-0.126
	(0.386)	(0.642)	(0.747)	(0.983)
Constant	3.677***	3.188***	3.442***	3.612***
	(0.125)	(0.230)	(0.281)	(0.321)
Bandwidth	50	50	50	50
Observations	2015	2015	2015	2015
Fuzzy	YES	YES	YES	YES
Polynomial Order	1	2	3	4
P-Value of Test†	0.0627	0.533	0.701	0.236

Standard errors in parentheses

Treatment is defined as having taken up the program.

The instrument is a dummy variable equal to one if the individual qualified for treatment.

Standard errors are clustered on each discrete value of the forcing variable.

† Test is against H0: This functional form is correct. See equation 8.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Numerical estimates of the treatment effect also fail to reject the null hypothesis that the program had no effect on student GPA. Table 8 reports the local average treatment effect estimated using a fuzzy regression discontinuity design under a number of functional form specifications.[13] Out of the five functional form specifications listed, column 2 is best according to the test against a nonparametric functional form and column 1 is likely misspecified. After excluding column one, it is clear that no estimate of the treatment effect is greater than one standard deviation away from zero. In other words, none of the estimates of the local average treatment effect are at all close to being significantly different than zero at any reasonable significance level.

[13]Sharp results are available in table 15 in the appendix.

The problem with these tests is that, while I cannot reject the null hypothesis, it is also impossible for me to rule out the possibility that the program has an economically meaningful effect on GPA. Given that the program is both simple and inexpensive, a GPA increase among compliers of 0.5 points (half a letter grade) would be a substantial amount. But point estimates for the effect of the program range from -.12 to 0.57 points of GPA. The large standard errors for each estimate (which, because of functional form misspecification, may understate the true standard errors) include an even greater range of results for the treatment effect. Finally, the fact that estimates of the size of the discontinuity are strongly dependent on the choice of functional form casts further doubt on the reliability of the point estimates themselves. Clearly, regression discontinuity does not have enough statistical power to successfully identify or rule out an effect in a small sample with a program that has a low expected treatment effect.

I also estimated the same regression discontinuity problem for a number of population subgroups, and listed fuzzy regression discontinuity results in table 17 in the appendix.[14] The estimates in the table are suspect for four reasons. First, they suffer from the same problems present in the full sample, and no result is significant. Second, estimates of the effect of the program on GPA are highly dependent on functional form assumptions in most cases. For example, the point estimates of the GPA benefits of the program range between .07 to 1.38 points for men and -.771 to .877 for women, meaning it is ambiguous whether one sex had greater gains then the other (if either sex gained at all). Third, if the estimated change in treatment status at the cut point is low, the denominator in equation 7 is low, which increases both point estimates and standard errors of the local average treatment effect. Finally, outliers can have a huge effect in small samples, and this effect can be magnified by the use of polynomial functions.[15] For these reasons, it is dangerous to read deeply into large point estimates which are insignificant, particularly in small samples.

Because of these caveats, the subgroup analysis cannot definitively say anything about which

[14] I will not perform multiple hypothesis tests because all subgroup results are just as ambiguous as the full sample results.

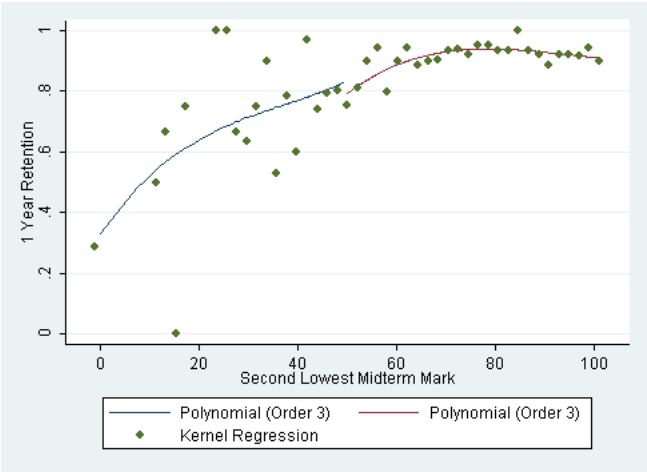
[15] A common problem in numerical analysis is that polynomial functions used for interpolation tend to oscillate at their extremes (this is sometimes called Runge's phenomenon), which magnifies error in problems of interpolation and extrapolation. This oscillation is less pronounced in regression analysis because the regression function does not have to pass through every point, but becomes a greater issue as the order of the polynomial rises and as the sample size falls. Given that estimates of the function at the cut point are extreme values by definition, they may suffer greatly from this problem.

groups had greater GPA results. While French students appear to consistently have had greater point estimate gains than their English student peers, and while older students consistently had lower GPA gains than their younger peers, both these groups had very small sample sizes and so suffer most from the problems listed above. The substantially negative point estimates for older students are particularly suspect. Further analysis with a much lager sample would be necessary to conclude whether these trends indicate heterogeneous treatment effects or bias due to misspecification.

6.3 Effect on Retention

Section 4 estimates of retention rates showed that students who met with the Coordinator of First Year Students had a one year retention rate which was 12% higher than their peers who also qualified for the program. While this represents a much greater percentage improvement than for GPA, regression discontinuity estimates of the effect of the program are even more ambiguous than they are for GPA.

Figure 5: Fraction of students retained after one year as a function of their second lowest exam score



The ambiguity of the results is clear from the picture in figure 5. This graph plots the probability of being retained after one year[16] in the program against their second lowest midterm mark. While

[16]All results in this section will refer to one-year retention. Results for half-year retention, two-year retention, and

the parametric functions on either side of the cut point appear to show a small discontinuity, this gap is dwarfed by the large variance in local average retention rates for students who score less than 60% on their second lowest midterm. For these students, the average retention rate has huge swings, which are substantially higher than the small estimated gap shown on the graph. There are two reasons for the data having these huge swings. First, a small sample size means that there are few observations per bin, and thus the average retention rate per bin will be measured with less precision. Second, the second lowest midterm mark appears to explain only a small fraction of the variation in the retention rate among students. While having an assignment variable which does not have a great degree of explanatory power does not invalidate the regression discontinuity method, it raises standard errors and thus requires a larger dataset in order to identify the same effect size.

Table 9: The Local Average Treatment Effect of the program on one-year retention

	(1) 1 Year Retention	(2) 1 Year Retention	(3) 1 Year Retention	(4) 1 Year Retention
Treatment Effect	-0.0370	0.0340	0.109	0.0623
	(0.0856)	(0.149)	(0.195)	(0.266)
Constant	0.868***	0.803***	0.776***	0.764***
	(0.0284)	(0.0535)	(0.0767)	(0.0952)
Bandwidth	50	50	50	50
Observations	2030	2030	2030	2030
Fuzzy	YES	YES	YES	YES
Polynomial Order	1	2	3	4
P-Value of Test†	0.000000830	0.113	0.00110	0.000000135

Standard errors in parentheses

Treatment is defined as having taken up the program.

The instrument is a dummy variable equal to one if the individual qualified for treatment.

Standard errors are clustered on each discrete value of the forcing variable.

† Test is against H0: This functional form is correct. See equation 8.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

three-year retention are also ambiguous. I have chosen retention after one year because, in practice, very few students drop out after only one semester, and I only had upper year retention rates for subsamples of the population. In section 4, I remarked that 2 year and 3 year retention is much higher for treated students than one-year retention is. Unfortunately, because students in more recent cohorts had not completed their second and third years of university at the time of data collection, sample sizes for these groups are much smaller. As a consequence, estimates for the treatment effect of the program are just as erratic for second and third-year retention results as they are for first-year retention.

The inability of regression discontinuity to provide a meaningful result in this case is further emphasized by table 9, which shows fuzzy estimates of the effect of the program on retention.[17] None of the point estimates lie outside of one standard deviation of zero, and so it is impossible to reject the null hypothesis. Point estimates of the treatment effect range from a decrease in the probability of being retained of 3.7% to an increase in the probability of being retention of 6.2%. As in the case of cumulative GPA, these point estimates clearly encompass values which are economically meaningful. Unlike GPA estimates though, there is also good evidence that the parametric estimate of the discontinuity are biased due to misspecification. With the exception of the second-order polynomial, specification tests reject the hypothesis that a semi-parametric function is not a better fit for the data. Even the second order polynomial barely passes the test with a p-value of 0.113. The reason for this is clearly evident in figure 5; particularly for midterm marks which are less than 60%, it is unclear what underlying functional form actually fits the data. As a consequence, the already high standard errors are likely to be underestimations of the true standard errors once modeling error is fully taken into account.

As with the GPA results, a subgroup analysis listed in table 18 of the appendix does not yield any statistically significant results, and suffers from the same list of problems. Nevertheless, there are a few interesting trends. First of all, if the program did have an effect on retention, it appears it was felt exclusively on male students, as males have point estimates which are substantially higher than females across all different specifications. French students also appear to have benefited more than students with other first languages. While these differences must be taken with a few grains of salt, this does signal some interesting trends for further study.

7 Threats to Identification

7.1 Effect of Failing an Exam

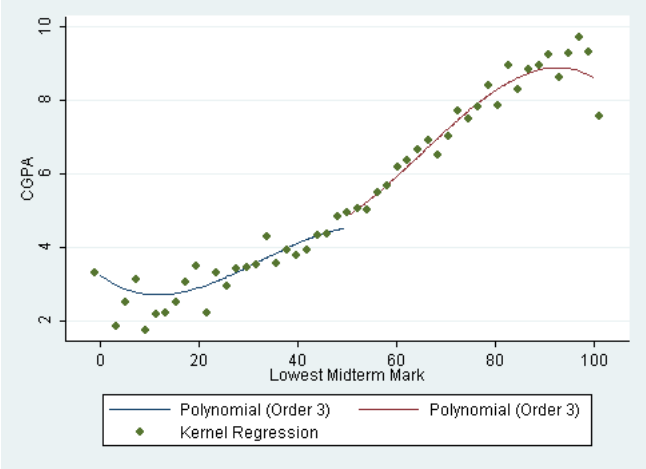
The key assumption of regression discontinuity is that the expected value of the dependent variable conditional on not receiving treatment is continuous at cut point ($\lim_{X \to c^-} E(Y_i|T_i = 0) =$

[17] Sharp results are listed in table 16, in the appendix

$\lim_{X \to c^+} E(Y_i|T_i = 0)$). This assumption would be violated if the act of failing an exam in itself provoked a behavioral response from the student, a phenomenon I will call the discouragement effect. For example, failing an exam could make a student lose hope and drop out, or it may give the impression (both real or imagined) that the student has to work harder in the second half of the course. If either of these cases is true, then the difference between students who fail and students who pass their exam with the second lowest mark is not completely attributable to the program being offered. Past studies which rely on a failed grade as an assignment rule have assumed that this discouragement effect does not exist, and so have ignored it in their analysis (see for example Matsudaira (2008) and Jacob and Lefgren (2004)). If the discouragement effect does exist, it would need to be corrected for.

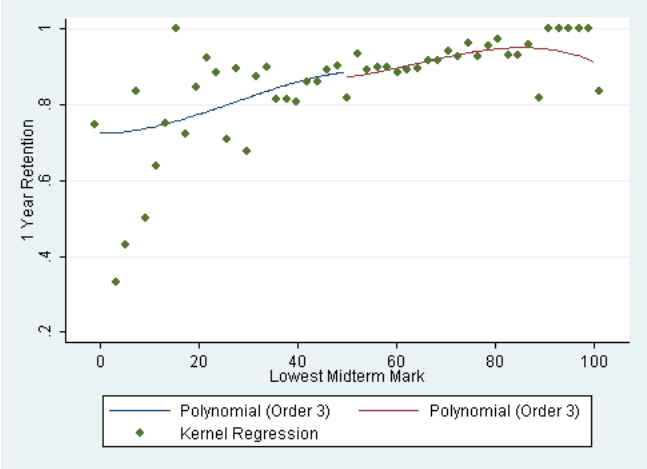
The dataset provides an easy way to test for the discouragement effect by using students' lowest exam score. To my knowledge, there are no policies or programs at the school in question which differentiate between students who pass all of their midterms, and students who fail just one midterm. If the discouragement effect exists for the second lowest midterm, it should be also be detectable from a regression discontinuity analysis based on students' lowest midterm mark.

Figure 6: Cumulative GPA after first year of studies as a function of the lowest exam score



Graphical analysis of the discouragement effect shows that it is either small or non-existent. Figures 6 and 7 plot GPA and retention outcomes with respect to midterm marks. On average,

Figure 7: Fraction of students retained after one year as a function of the lowest exam score



barely passing the midterm with the lowest mark is associated with a slightly higher GPA and a slightly lower retention according to these graphs. These statements are true both for the plotted third-order polynomial functions, as well for the data points immediately to the left and right of the cut point. Nevertheless, the size of both these gaps falls well within the natural variability of the function on either side of the cut point, and there are cases of much larger jumps in both GPA and retention at other values of the lowest midterm mark.

Table 10 shows parametric estimates of the discouragement effect on GPA. To be comparable with other results in this paper, this table defines the treatment effect to be the act of barely failing the exam with the lowest mark. With the exception of the first-order polynomial model which fits the data very badly, none of the alternate specifications show that failing a midterm has a significant effect on overall GPA. Point estimates from failing the exam with the lowest exam mark are still high, meaning that it is impossible to conclude with certainty whether discouragement exists or not. Moreover, none of the alternate specifications pass the specification test against a more flexible functional form, which may indicate that there is additional specification bias. Nevertheless, the fact that their magnitudes are very similar to the results from the sharp regression discontinuity results in table 15 is telling. Excluding the first-order polynomial, sharp point estimates of the intent to treat of the program vary between -.044 to 0.2 points of GPA. In contrast, failing the first

Table 10: The effect of barely failing the exam with the lowest mark on cumulative GPA

	(1) CGPA	(2) CGPA	(3) CGPA	(4) CGPA
Treatment Effect	-0.513**	0.106	-0.325	0.0141
	(0.187)	(0.261)	(0.340)	(0.458)
Constant	4.909***	4.666***	4.856***	4.802***
	(0.110)	(0.190)	(0.208)	(0.277)
Bandwidth	50	50	50	50
Observations	2307	2307	2307	2307
Fuzzy	NO	NO	NO	NO
Polynomial Order	1	2	3	4
P-Value of Test†	0.00000449	0.00467	0.0178	0.0174

Standard errors in parentheses

Treatment is defined as having failed the midterm with the lowest mark

Standard errors are clustered on each discrete value of the forcing variable.

† Test is against H0: This functional form is correct. See equation 8.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

midterm has an estimated effect of between -.33 and 0.1 points of GPA. Were Tracking Students at Risk having a positive effect on GPA, we would expect that point estimates of the intent to treat effect of the program to be consistently higher than point estimates of the discouragement effect. While the average point estimate of the intent to treat effect is higher than the average point estimate of the discouragement effect, the fact that there is significant overlap casts doubt on whether the program actually made a difference.

Table 11 shows parametric results of the discouragement effect on one-year retention. As with GPA, no coefficient result is significantly different from zero, and point estimates are all large enough to be economically meaningful, which means that the existence of a discouragement effect is ambiguous. Nevertheless, it is interesting that all point estimates are positive, meaning that this regression has estimated that the average retention of students who barely fail their exam with the lowest mark is actually higher than those students who barely pass this exam. This is consistent with the hypothesis that, if anything, the act of failing an exam motivates students to persevere.[18]

[18]Given my understanding of how undergraduate students work, this seems like a counter-intuitive result. One possibility is that there are some students who, upon receiving a failing grade in their first semester, immediately drop out and are dropped from the sample. If this were the case, then the remaining students in the sample who barely failed the exam would be slightly more likely to persevere from then on then students who barely passed. To

Table 11: The effect of barely failing the exam with the lowest mark on one-year retention

	(1) 1 Year Retention	(2) 1 Year Retention	(3) 1 Year Retention	(4) 1 Year Retention
Treatment Effect	0.0150	0.0307	0.0138	0.0701
	(0.0233)	(0.0311)	(0.0401)	(0.0488)
Constant	0.876***	0.865***	0.873***	0.867***
	(0.0140)	(0.0189)	(0.0254)	(0.0277)
Bandwidth	50	50	50	50
Observations	2327	2327	2327	2327
Fuzzy	NO	NO	NO	NO
Polynomial Order	1	2	3	4
P-Value of Test†	0.726	0.697	0.422	0.271

Standard errors in parentheses

Treatment is defined as having failed the midterm with the lowest mark

Standard errors are clustered on each discrete value of the forcing variable.

† Test is against H0: This functional form is correct. See equation 8.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Sharp regression discontinuity results for both the intent to treat effect of the program and the discouragement effect have very similar point estimates. Excluding the linear specification, point estimates of the intent to treat effect of the program range from a 1% increase in retention to a 4% increase in retention, whereas point estimates of the discouragement effect range from 1% to 7%. The fact that Tracking Students at Risk has not caused point estimates to increase by any appreciable margin lends support to the hypothesis that the program did not make a big difference to student retention.

7.2 Student Manipulation

Another possible threat to identification is whether students have the ability to control their exam marks in the neighborhood of the point of discontinuity. Because midterm exams normally either have many questions or long answer questions with uncertain marking schemes, it is very difficult for students at risk of failing to be able to accurately predict if they will just barely pass or just barely

my knowledge, students who drop out in this way should still be included in the sample and should still be counted as students who have dropped out, but since I was not directly involved in the data collection process I am not certain whether this was followed in all cases.

fail an exam. As a consequence, it is reasonable to assume that students initial exam scores are locally random close to the cut point. However, there are often students who attempt to have their marks changed after the exam, and these marks can be changed if the professor agrees they have been scored unfairly. If the process of negotiation were to cause students above and below the cut point to be different, this could bias the results from the regression discontinuity quasi-experiment.

Nevertheless, I would argue that the bias due to these post-exam negotiations of test marks is small for three reasons. First, there are a limited number of cases of changed marks every year, meaning that their impact on the results is small. Second, students who are aware of the Tracking Students at Risk program are likely to be more informed about other campus services and may be aware of the ability to negotiate a meeting with the Coordinator of First Year Students even without technically qualifying for the program. As a consequence, there is little incentive for these students to alter their exam scores in order to qualify for the program. Finally, all students have an incentive to attempt to negotiate a higher mark, regardless of what marks they got on their exams. Other than the existence of the Tracking Students at Risk program, there is no other mandated program to my knowledge at the school which applies only to students who fail two or more midterms. As a consequence, there is little reason to believe that there is an incentive for students to systematically sort around the 50% threshold on their second lowest exam.

Formal tests appear to support the claim that there is little nonrandom sorting across the threshold. The most common way of identifying sorting behavior is to test for discontinuities in variables which should not be affected by the program itself. In figures 8-10, I show that admission average, gender, and age do not appear to be discontinuous at the threshold, indicating that groups on either side of the threshold are comparable.

Regression estimates support these results. Tables 12-14 show parametric estimates of the effect of qualifying for the program on admission average, gender, and age. Underlying characteristics are not significantly different from each other under any of the specifications. Because standard errors are high, and because the fit of the functional form specifications may be bad at the cut point, I cannot say for sure that minor sorting has not had an influence on point estimates. Nevertheless, it appears at least that, if it exists, the problem is likely small.

Figure 8: Admission average as a function of the second lowest midterm mark

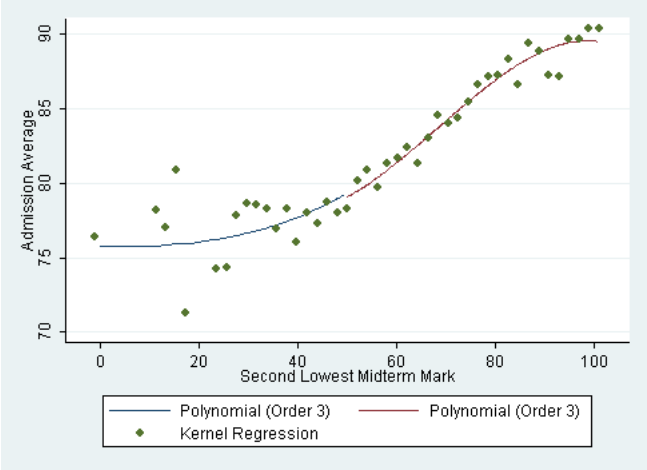


Figure 9: Gender as a function of the second lowest midterm mark

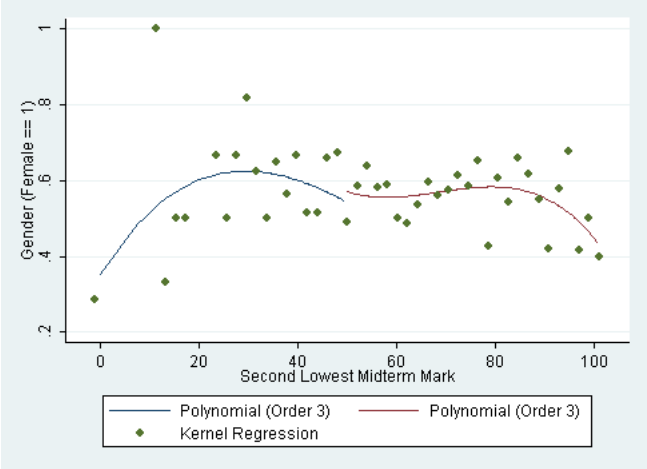


Figure 10: Age on the first day of the program as a function of the second lowest midterm mark

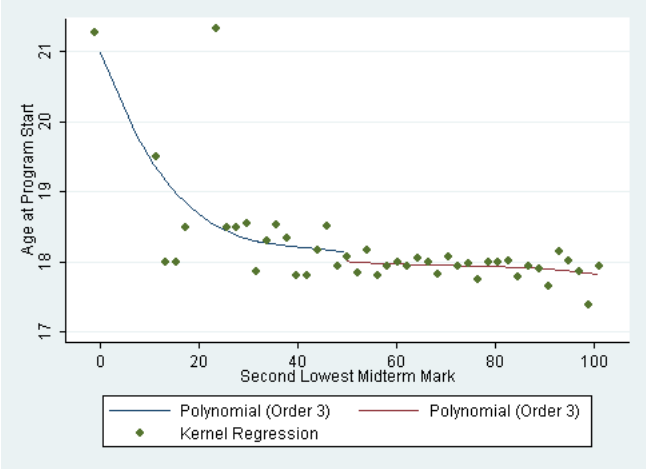


Table 12: The effect of qualifying for the program on admission average

	(1) Admission Average	(2) Admission Average	(3) Admission Average	(4) Admission Average
Treatment Effect	-0.389	0.833	0.255	1.033
	(0.665)	(0.906)	(1.203)	(1.416)
Constant	79.22***	78.45***	79.08***	79.35***
	(0.266)	(0.341)	(0.426)	(0.512)
Bandwidth	50	50	50	50
Observations	2024	2024	2024	2024
Fuzzy	NO	NO	NO	NO
Polynomial Order	1	2	3	4
P-Value of Test†	0.000956	0.000432	0.000624	0.00128

Standard errors in parentheses

Treatment is defined as having failed the midterm with the second lowest mark

Standard errors are clustered on each discrete value of the forcing variable.

† Test is against H0: This functional form is correct. See equation 8.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 13: The effect of qualifying for the program on gender

	(1) Female	(2) Female	(3) Female	(4) Female
Treatment Effect	0.0141	-0.00301	-0.0268	-0.0585
	(0.0553)	(0.0744)	(0.105)	(0.130)
Constant	0.576***	0.537***	0.569***	0.551***
	(0.0248)	(0.0361)	(0.0531)	(0.0612)
Bandwidth	50	50	50	50
Observations	2030	2030	2030	2030
Fuzzy	NO	NO	NO	NO
Polynomial Order	1	2	3	4
P-Value of Test†	0.0000779	0.0000900	0.00236	0.0356

Standard errors in parentheses

Treatment is defined as having failed the midterm with the second lowest mark

Standard errors are clustered on each discrete value of the forcing variable.

† Test is against H0: This functional form is correct. See equation 8.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 14: The effect of qualifying for the program on age on the first day of the program

	(1) Entry Age	(2) Entry Age	(3) Entry Age	(4) Entry Age
Treatment Effect	-0.0802	0.259	0.125	0.454
	(0.200)	(0.242)	(0.290)	(0.315)
Constant	18.01***	18.00***	18.01***	18.01***
	(0.0567)	(0.0741)	(0.0943)	(0.111)
Bandwidth	50	50	50	50
Observations	2030	2030	2030	2030
Fuzzy	NO	NO	NO	NO
Polynomial Order	1	2	3	4
P-Value of Test†	0.747	0.701	0.611	0.686

Standard errors in parentheses

Treatment is defined as having failed the midterm with the second lowest mark

Standard errors are clustered on each discrete value of the forcing variable.

† Test is against H0: This functional form is correct. See equation 8.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

8 Conclusion

Finding a suitable proxy for student motivation and engagement is very difficult. Effort, like intelligence, is a slippery concept that cannot be accurately derived from a questionnaire. In contrast, a student's willingness to participate in a voluntary program such as Tracking Students at Risk, which promises to improve his or her results in exchange for extra time investment, is a much better proxy for their willingness to make sacrifices in order to succeed. In this paper, I have argued it is possible to isolate the effect of motivation by making a few assumptions about the selection bias in a non-experimental evaluation of the program, once the effect of the program had been estimated using regression discontinuity.[19]

Unfortunately, in this case, a regression discontinuity analysis of the treatment effect of Tracking Students at Risk gives ambiguous results. The regression discontinuity method has very low power in relation to a random-control trial, and both a small sample size and the discrete nature of the assignment variable increased standard errors. Uncertainty over which parametric specification was correct also increased the likely degree of specification error in each regression. As a consequence, it is neither possible to confirm nor reject the hypothesis that Tracking Students at Risk had an appreciable effect on exam marks and persistence.

Nevertheless, this paper does yield a few conclusions. First, it is clear that Tracking Students at Risk did not yield huge and drastic results for students who took part. This is not surprising, because of the fact that the intervention was so simple. Assuming that selection bias is not negative, the program increased GPA on average by a maximum of 0.496 points after one year, and raised retention rates by a maximum of 12.2%. Quasi-experimental results suggest that benefits for retention are much lower, with point estimates ranging between a 1% and a 6% increase in retention. Still, the exact effect of the program, or indeed whether the program had an effect at all, cannot be determined by this study. This fact is unfortunate, because a clearer answer would have been very useful both in determining the role that information plays in raising student outcomes,

[19] In addition to the assumptions about the structure of the selection bias, it is also necessary to assume constant treatment effects. Whereas regression discontinuity weights individuals based on their propensity to fall near the cutoff, a simple difference in outcomes weights all individuals who qualify for the program equally. It is thus necessary to assume these two groups respond to the program in the same way (which is more plausible if the second group is restricted to only those who are close to the cut-off)

as well as the effect that motivation has on student results.

Second, this paper provides suggestive evidence that the psychological effect of failing an exam is not large. Unfortunately, because of the same problems which plagued the central regression discontinuity analysis, this cannot be said for certain. Nonetheless, point estimates appear to show that failing an exam increases the probability of retention at the end of the year, if anything.

Finally, this paper expands on the literature of gender differences. Like in previous studies, women were much more likely to take up treatment compared to men. Treated women had a higher GPA than those who were not, indicating either that they benefited from the program or that they were able to improve their marks through other means. In contrast, treated men did not increase their GPA at all, meaning that neither the program nor their desire to improve their marks succeeded in improving academic outcomes. On the one hand, this may be the case that the school was unable to attract motivated men into the program, which is consistent with the low percentage of men who took up treatment. On the other hand, it may be the case that motivated men are simply unable to make use of existing services in order to increase their GPA. In contrast, the difference in retention rates among treated and untreated men was much greater then the difference between treated and untreated women. Point estimates from regression discontinuity estimates suggest that this gap may be even larger than is suggested by the summary statistics. If the program truly did increase retention results for men more than women, this result is inconsistent with current literature which shows women benefiting more then men in similar programs. Clearly, more work remains to be done to understand the different relationship men and women have towards student support programs.

References

Angrist, J., Lang, D., & Oreopoulos, P. (2009). Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, *1*(1).

Bailey, T. R., Robbins, S. B., & Alfonso, M. (2005, January). Paths to persistence: An analysis of research on program effectiveness at community colleges. *Lumina Foundation for Education New Agenda Series*, *6*(1). Retrieved from `http://hdl.handle.net/10244/268`

Barreca, A. I., Lindo, J. M., & Waddell, G. R. (2011, September). *Heaping-induced bias in regression-discontinuity designs* (Working Paper No. 17408). National Bureau of Economic Research. Retrieved from `http://www.nber.org/papers/w17408`

Bettinger, E. P., & Long, B. T. (2005, May). *Addressing the needs of under-prepared students in higher education: Does college remediation work?* (Working Paper No. 11325). National Bureau of Economic Research. Retrieved from `http://www.nber.org/papers/w11325`

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*(1), 201-209.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615 - 635. doi: 10.1016/j.jeconom.2007.05.001

Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *The Review of Economics and Statistics*, *86*(1), 226-244.

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

Lee, D. S., & Lemieux, T. (2009, February). *Regression discontinuity designs in economics* (Working Paper No. 14723). National Bureau of Economic Research. Retrieved from `http://www.nber.org/papers/w14723`

Lotkowski, V. A., Robbins, S. B., & Noeth, R. J. (2004). *The role of academic and non-academic factors in improving college retention* (ACT Policy Report). Retrieved July 10, 2012, from `http://www.act.org/research/policymakers/pdf/college_retention.pdf`

MacDonald, I. H., Malatest, R., Assels, R., Baroud, R., Gong, L., Bernstein, L., ... Green-

wood, J. (2009, December). *Final impacts report: Foundations for success project* (Tech. Rep.). R.A. Malatest and Associates Ltd. Retrieved July 10, 2012, from http://malatest.com/CMSF%20FFS%20-%20FINAL%20Impacts%20Report.pdf

Matsudaira, J. D. (2008, February). Mandatory summer school and student achievement. *Journal of Econometrics*, *142*(2), 829-850.

Patton, L. D., Morelon, C., Whitehead, D. M., & Hossler, D. (2006). Campus-based retention initiatives: Does the emperor have clothes? *New Directions for Institutional Research*, *2006*(130), 9–24.

Rickinson, B. (1998). The relationship between undergraduate student counselling and successful degree completion. *Studies in Higher Education*, *23*(1), 95-102.

Rickinson, B., & Rutherford, D. (1996). Systematic monitoring of the adjustment to university of undergraduates: A strategy for reducing withdrawal rates. *British Journal of Guidance and Counselling*, *24*(2), 213-225.

Turner, A. L., & Berry, T. R. (2000). Counseling center contributions to student retention and graduation: A longitudinal assessment. *Journal of College Student Development*, *41*(6), 627 - 636.

Wilson, S. B., Mason, T. W., & Ewing, M. J. M. (1997). Evaluating the impact of receiving university-based counseling services on student retention. *Journal of Counseling Psychology*, *44*(3), 316-320.

9 APPENDIX: Supplementary Tables

Table 15: The effect of qualifying for the program on cumulative GPA after one year

	(1) CGPA	(2) CGPA	(3) CGPA	(4) CGPA
Treatment Effect	-0.227	0.207	0.0522	-0.0437
	(0.164)	(0.235)	(0.288)	(0.343)
Constant	3.628***	3.267***	3.465***	3.589***
	(0.1000)	(0.162)	(0.183)	(0.183)
Bandwidth	50	50	50	50
Observations	2015	2015	2015	2015
Fuzzy	NO	NO	NO	NO
Polynomial Order	1	2	3	4
P-Value of Test†	0.0665	0.418	0.693	0.267

Standard errors in parentheses

Treatment is defined as having failed the midterm with the second lowest mark

Standard errors are clustered on each discrete value of the forcing variable.

† Test is against H0: This functional form is correct. See equation 8.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 16: The effect of qualifying for the program on one-year retention

	(1) 1 Year Retention	(2) 1 Year Retention	(3) 1 Year Retention	(4) 1 Year Retention
Treatment Effect	-0.0162	0.0122	0.0401	0.0201
	(0.0375)	(0.0535)	(0.0720)	(0.0862)
Constant	0.864***	0.808***	0.794***	0.776***
	(0.0228)	(0.0367)	(0.0491)	(0.0545)
Bandwidth	50	50	50	50
Observations	2030	2030	2030	2030
Fuzzy	NO	NO	NO	NO
Polynomial Order	1	2	3	4
P-Value of Test†	0.00000860	0.00104	0.00189	0.000169

Standard errors in parentheses

Treatment is defined as having failed the midterm with the second lowest mark

Standard errors are clustered on each discrete value of the forcing variable.

† Test is against H0: This functional form is correct. See equation 8.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 17: The local average treatment effect of the program on cumulative GPA after one year - Subgroups

	Second Order Polynomial		Third Order Polynomial		Fourth Order Polynomial	
	Treatment Effect	OBS	Treatment Effect	OBS	Treatment Effect	OBS
Males	0.0708	876	0.0702	876	1.338	876
	(1.179)		(1.669)		(2.298)	
Females	0.877	1139	0.130	1139	-0.771	1139
	(0.597)		(0.661)		(1.026)	
Sudents Under 19	0.994	1750	0.849	1750	0.625	1750
	(0.673)		(0.710)		(1.086)	
Students 19 or Older	-1.112	265	-4.037	265	-7.559	265
	(1.339)		(3.178)		(8.190)	
Local Students	0.901	1185	0.270	1185	0.00966	1185
	(0.821)		(1.014)		(1.415)	
Out-of Town Students	0.213	830	-0.329	830	-0.212	830
	(1.047)		(1.143)		(1.398)	
English Speakers	0.809	1376	0.382	1376	-0.387	1376
	(0.916)		(0.991)		(1.759)	
French Speakers	1.946	254	3.342	254	2.727	254
	(2.209)		(3.715)		(2.460)	
Other Language Speakers	-0.482	385	-1.726	385	-1.588	385
	(1.062)		(1.679)		(1.479)	

Treatment is defined as having taken up the program.

The instrument is a dummy variable equal to one if the individual qualified for treatment.

Standard errors are clustered on each discrete value of the forcing variable.

Standard errors in parenthesis

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 18: The local average treatment effect of the program on one-year retention - Subgroups

	Second Order Polynomial		Third Order Polynomial		Fourth Order Polynomial	
	Treatment Effect	OBS	Treatment Effect	OBS	Treatment Effect	OBS
Males	0.377	883	0.448	883	0.797	883
	(0.319)		(0.480)		(0.706)	
Females	-0.0469	1147	-0.0242	1147	-0.0934	1147
	(0.0575)		(0.0791)		(0.105)	
Sudents Under 19	0.00885	1764	0.0656	1764	0.0381	1764
	(0.163)		(0.193)		(0.285)	
Students 19 or Older	0.0142	266	0.246	266	0.349	266
	(0.262)		(0.508)		(0.980)	
Local Students	0.112	1191	0.0453	1191	0.227	1191
	(0.179)		(0.250)		(0.376)	
Out-of Town Students	0.0136	839	0.157	839	-0.0289	839
	(0.248)		(0.300)		(0.378)	
English Speakers	-0.0697	1385	0.0668	1385	-0.0458	1385
	(0.182)		(0.208)		(0.396)	
French Speakers	0.224	258	0.0711	258	0.288	258
	(0.155)		(0.189)		(0.245)	
Other Language Speakers	-0.0664	387	0.0675	387	-0.0452	387
	(0.116)		(0.134)		(0.176)	

Treatment is defined as having taken up the program.

The instrument is a dummy variable equal to one if the individual qualified for treatment.

Standard errors are clustered on each discrete value of the forcing variable.

Standard errors in parenthesis

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$