Economics of Education Review 29 (2010) 187–199

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

The effects of quantitative skills training on college outcomes and peers☆

Kristin F. Butcher, Patrick J. McEwan*, Corrine H. Taylor

Wellesley College, United States

ARTICLE INFO

Article history:
Received 11 June 2009
Accepted 12 June 2009

Keywords:
College
Regression discontinuity
Quantitative skills
Remedial
Peer groups

ABSTRACT

This paper estimates the causal effect of taking a course in quantitative reasoning on students' academic performance and classroom peer-group composition at a liberal arts college. To identify effects, the paper compares the outcomes of otherwise similar students who barely passed a baseline quantitative skills assessment (not taking the course) with students who barely failed (taking the course). The regression-discontinuity estimates show little impact on academic outcomes for student close to the passing cutoff, including grades on subsequent courses with quantitative content, but we are unable to distinguish small from zero effects. Exogenous course assignment does affect the composition of students' classroom peer groups in subsequent years. The effects can only be generalized to students in the vicinity of the passing threshold (but not students with much worse quantitative skills at the baseline). We discuss implications for research and policy on remediation.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

Colleges and universities enroll students from diverse intellectual, economic, racial, and ethnic backgrounds. This diversity necessarily means that some students are better prepared in some academic subject areas than others. In recognition of this diversity of skills and past training, and in order to ensure that all students become proficient in mathematical, logical, and statistical tools required in the curriculum, Wellesley College implemented a quantitative reasoning (QR) requirement in 1997. Students must take a QR assessment administered during first-year orientation. If they fail that test, they must take and pass a one-semester quantitative skills course in their first year.[1] This course is designed to teach applied quantitative skills to students who were unable to pass the assessment exam such that they can effectively participate in any course in the curriculum.

Courses designed to deliver skills to allow students to effectively participate in the broader curriculum are often termed remedial, developmental, or gatekeeper[2] programs in American higher education. Such courses are common

☆ We are grateful to Larry Baldwin and Jessica Polito for their help in organizing and interpreting the administrative data, and to Yo-Jud Cheng and Sadia Raveendran for excellent research assistance. We thank Paco Martorell, Tom Dee, and the participants of the Mellon 23 Conference (September 2008) at Wellesley College for helpful comments. The Andrew W. Mellon Foundation and Wellesley College provided financial support. We are responsible for any errors. The views expressed here are the authors' and not necessarily those of Wellesley College, the Andrew W. Mellon Foundation, or any other entity.

* Corresponding author at: Wellesley College, Dept. of Economics, 106 Central St., Wellesley, MA 02481, United States. Tel.: +1 781 283 2987; fax: +1 781 283 2177.
E-mail address: pmcewan@wellesley.edu (P.J. McEwan).

0272-7757/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.econedurev.2009.06.001

[1] The QR requirement has two parts. After passing either the QR exam or the one-semester QR course, students must subsequently pass a QR "overlay" course—drawn from a menu across the curriculum—that applies quantitative skills in a disciplinary setting. There is a list of courses that fulfill this requirement, all of which require analysis of data.

[2] The QR course at Wellesley is best described as a "gatekeeper" course in this over-arching literature, since students earn credit toward a degree by taking the course. Developmental and remedial courses often are not credit-bearing.

and costly, despite weak evidence on their impact on student learning (Koski & Levin, 1998; Levin & Calcagno, 2008). Impact estimates are often based on simple comparisons of post-program outcomes between participating and non-participating students. Students are usually assigned to participate on the basis of attributes (like baseline test scores) correlated with lower performance. Thus, such comparisons are almost preordained to yield disappointing and biased assessments of program effects. Instead, this paper uses a regression-discontinuity design to obtain internally valid estimates of the causal effect of taking a QR skills course on subsequent student performance (Imbens & Lemieux, 2008; Lee, 2008).

Our empirical approach is facilitated by test-based assignment of students to the QR course. We compare students whose initial QR assessment scores fell just below the passing cutoff score (the treatment group) with students whose scores fell just above the cutoff (the control group). We confirm that students' probabilities of taking the course increase sharply below the cutoff, but that treatment and control students are otherwise similar in their baseline characteristics such as race and SAT scores. We examine the impact of taking the course on a number of academic outcomes, including overall grade point averages, grades in the first QR "overlay" course, the number of courses taken with quantitative content, and the composition of students' classroom peer groups throughout their college careers. We find no consistent and statistically significant differences in this limited set of academic outcomes, though we do find that the course causes a statistically significant difference in the composition of students' classroom peer groups. Specifically, the course assignment increases the likelihood that students take courses with other students with low baseline quantitative skills, even by their senior year. One hypothesis, not directly testable in our data, is that the reduced-form findings of our paper mask two countervailing effects on student learning: a positive effect of receiving quantitative skills training, but a potentially negative peer effect.

There are two important caveats to these findings. First, the empirical strategy identifies the causal effect for students close to the passing cutoff. The average math SAT score among these students is 600–650, which is very high compared to scores of typical students in remedial, developmental or gatekeeper classes at many institutions.[3] We would also like to determine whether the course has a measurable impact on students with the lowest baseline skills, but this would require a different evaluation design. Second, a small college has, almost by definition, a small sample problem. We attempt to overcome this by pooling multiple years of data, facilitated by a common assessment and course assignment rule across 10 years. Even so, our statistical tests are not powerful enough to detect small effects on student outcomes (although we do identify effects of course-taking on subsequent peer-group composition).

These results generate additional research and policy questions. For example, if this course, and courses like it, do not have an impact on the academic performance of the students who are on the current threshold of being assigned to take the course, should the threshold be lowered, perhaps allowing more intensive targeting of resources toward students at the bottom of the skills distribution? If so, how can we credibly and feasibly evaluate whether such a program has a causal impact on outcomes? Finally, if the current method of delivering basic skills has an unintended consequence of shifting students' peer groups, can treatments be delivered in a way that avoids the potentially negative consequences of peer sorting?

The paper proceeds as follows. Section 2 discusses prior research on remedial, developmental and gatekeeper courses. Section 3 provides detailed background on the QR program. Section 4 describes the data, and Section 5 presents the details on the empirical strategy. Section 6 presents the results, and Section 7 concludes and discusses additional research and policy questions.

2. Prior research

This study has much in common with a small number of papers that use quasi-experimental techniques to evaluate the causal impact of remedial, developmental, and gatekeeper courses on student outcomes. Understanding the impact of these courses is very important: perhaps one-third of students entering post-secondary education are not ready for college-level work (Bettinger & Long, 2009). Tremendous resources are devoted to remedial education, but little is known about the causal impact of these programs on academic outcomes of at-risk students (Levin & Calcagno, 2008). Several recent papers apply a regression-discontinuity design in settings where placement in remedial programs is based on test score cutoffs. These papers compare the post-program differences in outcomes between students just below and just above the assignment threshold.

Calcagno and Long (2008) examine administrative data for 100,000 students at Florida's 2-year and 4-year colleges. All college students take a standardized test that measures basic skills, and are referred to remedial education courses based on the exam score. Martorell and McFarlin (2007) use administrative data on more than 450,000 2-year and 4-year college students in Texas where students are also referred to remedial education based on their score on an assessment exam relative to a strict cutoff. Lesik (2006, 2007) uses data from a single cohort entering one 4-year public university (approximately 1200 students), exploiting a similar discontinuity strategy.

The evidence from these papers is mixed. Lesik (2006, 2007) finds positive effects, with a remedial math course increasing the probability of passing a subsequent college-level math course and decreasing the probability of dropping out of college. Calcagno and Long (2008) find that while assignment to remediation appears to increase persistence to the second year and the total number of credits completed, it does not appear to increase the comple-

[3] In comparison, the average among all college-bound high school seniors in 2007 was 515 (College Board SAT, 2007). The average math SAT among a sample of Florida students in public 2-year and 4-year colleges was 490 (Calcagno and Long, 2008).

tion of college-level credits or eventual degree completion. Martorell and McFarlin (2007) find weak evidence that remediation improves the grades received in college-level math, but no evidence that remediation improves years of college completed, academic credits attempted, the receipt of a college degree, or labor market performance.

Bettinger and Long (2009) find different results using a different empirical strategy for administrative data from Ohio. They exploit the fact that different institutions employ different rules for referring students to remediation. Thus, students with a given test score might be referred to remediation at one institution, but not another. They use the distance between a student's home and state colleges to control for a student's selection of which institution to attend. They find that students who are induced to take remedial education classes by virtue of their institution's rules are more likely to persist in college and more likely to complete a degree than are students with similar test scores and background characteristics who were not induced to take remedial classes.

Finally, in addition to the quasi-experimental evidence cited above, there is some experimental evidence on programmatic interventions to improve academic outcomes of at-risk students. Angrist, Lang, and Oreopoulos (2009) report on the results of a randomized controlled trial at a large Canadian university concerned about drop out rates and poor achievement. Low-achieving students were randomly assigned to a control group and three treatment groups, including (1) a financial incentive (cash award of up to a full year's tuition); (2) academic support services including mentoring and supplemental instruction; and (3) a combination of both financial incentives and support services. Women's grades were significantly higher in the two groups that received financial incentives, and there is modest evidence that the combination of incentives and services was more effective than incentives alone.

Our paper differs from this research in several ways. First, the average "remedial" student in our sample is higher achieving than other samples, and exclusively female. Second, the content of the QR treatment is different: it consists of a single, intensively applied, and relatively homogeneous treatment (unlike the multiple, heterogeneous treatments applied across state university systems). Third, the student outcomes of interest are different. Prior research focuses on student retention, drop out rates, and course completion. Wellesley College has a very low drop out rate (McEwan and Soderberg, 2006), so we instead focus on the number of quantitative courses completed, and grades in these and other courses.

3. Background

3.1. The quantitative reasoning requirement

The Quantitative Reasoning (QR) requirement has two parts: (1) a "basic skills" component, to ensure that entering students are well prepared for quantitative coursework across the curriculum, and (2) an "overlay" component, to ensure that graduating students are proficient in mathematical, logical, and statistical tools (Taylor, 2006). A student satisfies the basic skills requirement by passing the QR assessment upon entering the College, or by passing a one-semester basic skills course ("Introduction to Quantitative Reasoning") in her first year at the college. Fulfilment of the QR basic skills requirement is a prerequisite for subsequent courses with quantitative content. To satisfy the second requirement, a student must pass a QR "overlay" course that emphasizes statistical analysis and interpretation of data. Some overlay courses are discipline-specific statistics courses; others are laboratory science courses in which students collect and analyze data and present their findings in labs.

3.2. Testing and course assignment

The QR assessment, administered during first-year orientation, consists of 18 questions, mainly open-response. Each response is graded as full-credit, half-credit, or no credit. A score of 9.5 (out of 18) or better is required to pass the assessment; a similar assessment and passing threshold have been employed since 1997. Students whose highest assessment score is between 9.5 and 12.0 are advised that they may opt to take the basic skills class to improve their QR skills. Students who score 9 or below are required to enroll in the basic skills course in their first year at the college.

Students can voluntarily take a re-test 1 day after taking the first test. We denote T_i^1 as student i's score on the first assessment and T_i^2 as her score on the optional re-test. In practice, few students opt to re-test if $T_i^1 > 9$, and few students voluntarily take the QR basic skills course if $\max(T_i^1, T_i^2) > 9$. If $\max(T_i^1, T_i^2) \leq 9$, then the course is mandatory.

3.3. Description of the treatment

Depending on the year, between 6 and 10% of the entering class is required to take the QR basic skills course. In the semester-long course, students review mathematical content from secondary school (including numeracy, algebra, linear and exponential modeling, graphing, geometry, basic probability and statistics, and formal logic). The students practice with these skills in a variety of authentic contexts, such as medical decision-making and personal finance. The basic skills course enhances students' capabilities in six areas: (1) reading and understanding quantitative information; (2) interpreting quantitative information and drawing appropriate references; (3) solving problems; (4) estimating answers and checking for reasonableness; (5) communicating quantitative information; and (6) recognizing the limits of mathematical or statistical models.[4]

This approach is designed to provide first-year students with the QR skills they will need in their subsequent quantitative coursework, including, but not limited to, the required QR overlay course. QR skills are also required in other courses with quantitative content (such as introduc-

[4] These are the six key QR areas identified by the Mathematics Association of America. They are consistent with many of the best-practice elements of courses identified by Koski and Levin (1998).

Table 1
Descriptive statistics.

	Full sample			Student did not take quantitative skills course			Student took quantitative skills course		
	Mean	S.D.	*N*	Mean	S.D.	*N*	Mean	S.D.	*N*
Took course	0.083		6311	0.000		5785	1.000		526
First assessment score (T^1)	12.616	2.985	6311	13.151	2.433	5785	6.729	2.004	526
$1\{T^1 \leq 9\}$	0.135		6311	0.060		5785	0.958		526
Student took re-test	0.090		6311	0.060		5785	0.416		526
Re-test score (T^2)	9.938	2.824	565	11.601	1.867	346	7.311	1.949	219
$1\{T^2 \leq 9\}$	0.375		565	0.017		346	0.941		219
Grade point average	0.000	1.000	6304	0.044	0.981	5778	−0.484	1.078	526
Proportion of low grades	0.047	0.100	6310	0.044	0.098	5784	0.074	0.121	526
"Overlay" course grade	0.000	1.000	4406	0.053	0.968	4126	−0.775	1.141	280
No. of courses taken with quant. skills prerequisite	3.057	1.705	6311	3.110	1.705	5785	2.473	1.604	526
Math SAT score	673.143	66.350	5887	680.853	60.862	5454	576.028	54.961	433
Verbal SAT score	684.776	69.804	5886	688.694	67.930	5453	635.427	74.202	433
ACT score	29.188	2.927	1582	29.587	2.625	1425	25.573	3.058	157
African-American	0.050		6311	0.033		5785	0.234		526
Latina	0.052		6311	0.043		5785	0.148		526
Asian or Asian-American	0.240		6311	0.258		5785	0.051		526
Other	0.212		6311	0.213		5785	0.203		526
White	0.446		6311	0.454		5785	0.363		526
Non-traditional student	0.026		6311	0.018		5785	0.103		526

Note: Standard deviations are not reported for binary variables. See text for details of sample and variables.

tory economics or physics courses) for which the basic skills class is a prerequisite.

The basic skills course is typically offered four times each year, with three sections in the fall and one in the spring. Class enrollments are usually capped at 15. The small class size allows for cooperative learning and individualized attention. Classes meet in "lecture" twice each week and in "lab" once each week. The "lecture" meetings are in a conference room so that students face each other at one large table and have ample opportunities to work with each other. The "lab" meetings are held in a computer laboratory and often involve the use of Excel to practice QR skills with real data. Anecdotally, students report that meeting three times each week in such a collaborative environment results in their getting to know each other better in this first-year class than in most other lower-level classes at the College. Thus, a potential concern is that weaker students who are "tracked" into the basic skills course will continue to attend courses and study with lower-achieving peers instead of students with relatively stronger QR skills.

4. Data

The treatment and assignment procedure have been consistently applied since 1998. Thus, we pool administrative data from 10 cohorts of degree-seeking undergraduates who took the QR assessment between Fall 1998—the second year of the QR program—and Fall 2007.[5] There are records for 6311 students, 8% of whom took the quantitative skills course in their first or second semester at the College. Table 1 reports descriptive statistics for the full sample, and subsamples of students who did or did not take the QR basic skills course. The mean score on the first intake test (T_i^1) is 12.6, about 1.2 standard deviations above the passing threshold of 9. Thirteen percent of students score at or below 9. Nine percent of students volunteered to re-test, of whom 97% did not pass the first assessment.

Table 1 reports the means of four measures of student outcomes. The first is a student's cumulative grade point average, calculated over all available course credits and reported as a *z*-score. One drawback of the GPA variable is that it reflects the roughly 90% of student grades with a recorded letter grade. The remaining 10% include instances in which students invoked a voluntary credit/no-credit option[6] (which must be chosen before an early semester deadline) and less common instances such as withdrawal after the course drop deadline, often interpreted as a failing grade.

Thus, the second outcome variable is the proportion of students' grades that are "low," defined as a C- or below. The variable allows us to impute low values for students that receive "no credit" (C- or below) or withdraw from a course, avoiding potential selection issues. The variable also focuses attention on the left tail of the grade distribution,

[5] This includes transfer students and older, non-traditional undergraduates who gain admission through a special program. It excludes a small number of other students (e.g. cross-registered or exchange students), even in the rare cases when they take the quantitative skills assessment.

[6] To get a "pass" with this option a student must get a C or higher in the course.

which includes disproportionate numbers of students with low quantitative skills. The third outcome variable is the letter grade in the mandatory QR "overlay" course, taken after fulfilling the basic skills requirement. It is also reported as a z-score.[7] The fourth variable is the number of (optional) quantitative courses taken by students over the course of their college careers. We define these as courses that list the QR basic skills components as a prerequisite, including introductory courses across the social and natural sciences.

Finally, we posit that students' exogenous assignment to the QR basic skills course also represents an exogenous assignment to a set of classroom peers. To assess the effects of the treatment on the composition of classroom peer groups over time, we calculate for each student—in their first, second, third, and fourth years—the proportion of their classmates in that year who ever took the QR basic skills course.

In Table 1, the mean differences in outcome variables favor students who did not take the QR basic skills course. This is unsurprising, given that course assignment is based on objective measures of pre-enrollment quantitative skills. However, it highlights the pitfalls in naïve attempts to estimate course treatment effects by simple mean comparisons, or even regression approaches that condition on a limited set of student background variables.

Other variables in Table 1 show that students are racially and ethnically diverse, and arrive at the College with high baseline achievement on standardized tests. However, students are clearly not assigned at random to the quantitative skills course. African-American and Latina students are over-represented, as are students with "low" SAT and ACT scores. It bears emphasis, however, that these scores are still high compared to national samples.

5. Empirical strategy

We wish to estimate the causal effect of the QR basic skills course on students' academic outcomes. A starting point is the following linear regression:

$$O_i = \beta_0 + \beta_i R_i + X_i\beta_2 + \varepsilon_i \quad (1)$$

where O_i is the outcome of student i, R_i is a dummy variable indicating whether a student has taken the course, X_i is a vector of student-specific control variables, and ε_i is an error term. One can interpret β_1 as the causal effect of taking a remedial course, but only if $\text{cov}(R_i,\varepsilon_i) = 0$. The assumption is implausible in settings where students choose, or are assigned to, remedial courses on the basis of imperfectly observed variables, such as ability or motivation, that also affect performance.

In contrast, course assignment at Wellesley College primarily occurs on the basis of observed intake assessments, in concert with a pre-determined passing threshold. To identify the course effect β_1, we rely upon sharp and exogenous variation in the probability of course-taking that occurs when students cross this threshold. Consider the following equations, estimated by two-stage least squares (TSLS) in the full sample of students (see Imbens & Lemieux, 2008; McEwan & Shapiro, 2008 for related discussion):

$$R_i = \alpha_0 + \alpha_1 \times 1\{T_i^1 \leq 9\} + f(T_i^1) + \upsilon_i \quad (2)$$

$$O_i = \beta_0 + \beta_1 R_i + f(T_i^1) + \omega_i \quad (3)$$

In Eq. (2), the first-stage, the course-taking dummy variable is regressed on a dummy variable indicating scores on the first assessment below the assignment threshold—the indicator function $1\{T_i^1 \leq 9\}$—which serves as the excluded instrument. Eq. (2) also controls for a smooth function of T_i^1, initially specified as a piecewise quadratic polynomial:

$$f(T_i^1) = \delta_1 T_i^1 + \delta_2 (T_i^1)^2 + \delta_3 \times 1\{T_i^1 \leq 9\} \times (T_i^1 - 9) + \delta_4 \times 1\{T_i^1 \leq 9\} \times (T_i^1 - 9)^2$$

We assess the fit of this functional form by visual inspection of the unsmoothed means of the dependent variable within discrete tests score values, and also by verifying that results are robust to alternate specifications of the piecewise spline of test scores. In all subsequent regressions, we adjust standard errors for clustering within the smallest discrete values of the assignment variables, to account for potential misspecification of functional form (Lee & Card, 2008).

We anticipate that a score below the passing cutoff will not perfectly predict course-taking ($0 < \widehat{\alpha}_1 < 1$), since a portion of students might choose to re-test and obtain a passing score. Further, a small number of students choose to take the course despite obtaining a barely passing score on the first assessment.

In the second-stage regression (Eq. (3)), β_1 identifies the causal effect of course-taking if $cov(1\{T_i^1 \leq 9\}, \omega_i) = 0$. Stated differently, students close to the passing threshold must be similar in observed and unobserved ways that affect outcomes, conditional on the smooth function of T_i^1. This seems likely, as long as students' precise scores on T_i^1—whether 9 or just above—have a random component. It would be less likely if students or administrators precisely manipulate T_i^1 in the vicinity of the threshold, perhaps inducing correlations between T_i^1 and student attributes (Lee, 2008; McCrary, 2008). This seems implausible in the present context since testing conditions are carefully monitored to avoid cheating, and the actual tests are graded without students' identifying data.

Still, we can explicitly test whether the observed characteristics X_i vary sharply around the threshold by estimating Eq. (2) with each X_i as a dependent variable. Further, estimates of course effects should be insensitive to controls for X_i. Finally, we can examine the histogram of V_i^1 to search for clustering of students on either side of the passing threshold, indicative of score manipulation.

Finally, it bears emphasis that the treatment effect may be heterogeneous across subpopulations of students. In this case, Eq. (3) yields estimates of local average treatment effect for students who: (1) obtain scores on T_i^1 that are close to 9, and (2) are induced to take the course by virtue of

[7] We estimated specifications using a binary variable indicating low-achievement in an overlay course, but these always yielded similar results, and are not reported in this paper.

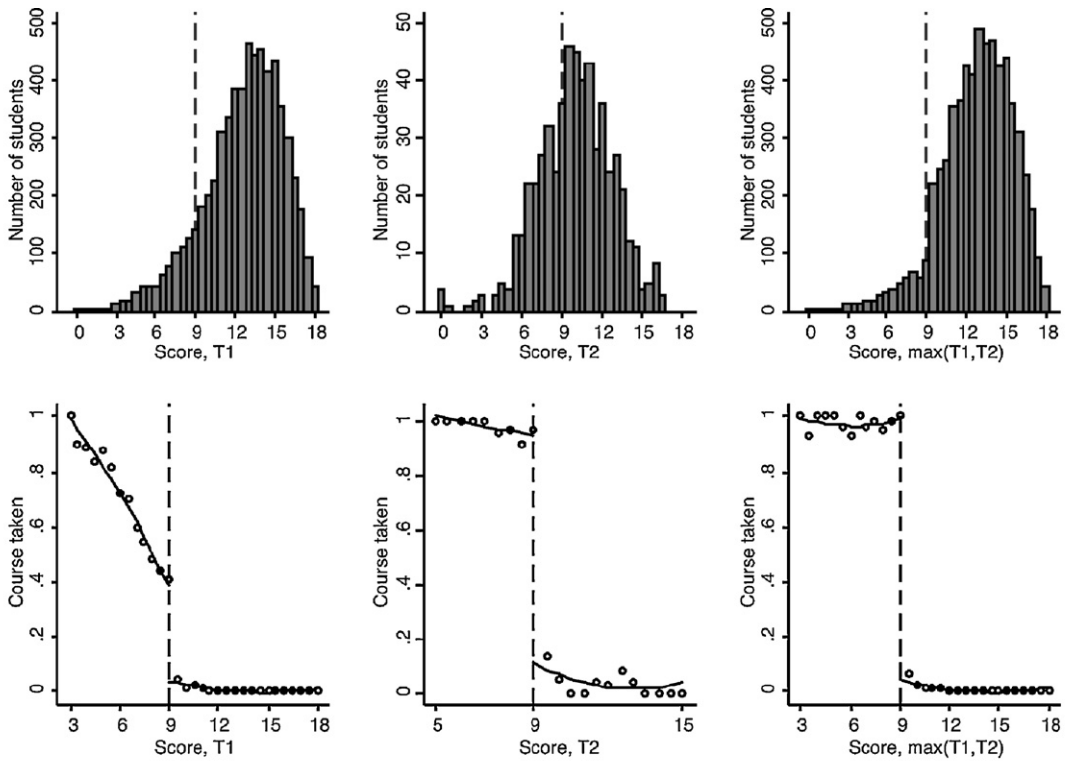


Fig. 1. Quantitative assessments and course-taking. Note: The top panels report histograms of three quantitative assessment variables. In the bottom panels, circles indicate the proportion of students taking the course within discrete values of each quantitative assessment variable. Lines indicate fitted values of quadratic polynomials estimated on either side of the cutoff.

scoring below the threshold (typically students who opted *not* to re-test, or did so unsuccessfully).

6. Results

6.1. Assignment to the quantitative skills course

We first confirm that course assignment followed the stated procedures. The top-left panel of Fig. 1 reports a histogram of the first assessment score (T_i^1). It is relatively smooth, and there is no evidence of bunching of observations to the right or left of the passing cutoff (indicated by a dotted line), which could indicate test score manipulation. In the bottom-left panel, circles illustrate mean values of the course-taking dummy variable, taken within bins of 0.5 points on the first assessment score (the smallest unit of T_i^1). The solid lines are fitted values of quadratic polynomials, estimated separately on either side of the passing threshold. The panel shows a sharp increase in the proportion of course-takers as scores decline from 9.5 to 9.[8] In column (1) of Table 2, we report estimates based on Eq. (2), the empirical counterpart to the panel. The estimated size of the break in course-taking is 0.35. It is highly significant ($T = 19.4$), and invariant to the inclusion of an additional set of student background controls (column 2). The upcoming results will rely on this source of variation in the course-taking probability to identify effects on student outcomes.

The middle and right-most panels of Fig. 1 suggest two other empirical approaches; we will argue that the first is less feasible and the second is inadvisable. In the upper-middle panel, we report the histogram of the re-test score (T_i^2), among the 9% of our sample that opted to take it. The much smaller sample size is evident from the *y*-axis scale and the more jagged distribution. The bottom-middle panel shows, as expected, that the re-test score results are binding, since the vast majority students with $T_i^2 \leq 9$ take the course. We could potentially leverage this source of variation to estimate course effects. Empirically, the analysis would be based on Eqs. (2) and (3), replacing T_i^1 with T_i^2 and limiting the sample to re-test students. While we report results from this specification as a robustness check, the much smaller sample size severely limits the precision of estimates.

The upper-right panel of Fig. 1 shows the histogram of $\max(T_i^1, T_i^2)$, the maximum assessment score obtained by a student on either the first test or the re-test. In some col-

[8] The panel also shows that students well below the threshold are more likely to take the course than students just below it. Students closer to the threshold are more likely to take the re-test, ensuring for some that $\max(T_i^1, T_i^2) > 9$ and that they can avoid taking the course.

Table 2
First-stage estimates and checks for smoothness of background variables around cutoff.

	Dependent variable:										
	Took quantitative skills course		Math SAT	Verbal SAT	ACT	African-Amer.	Latina	Asian/Asian-Amer.	Other	White	Older student
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
$1\{T^1 \leq 9\}$	0.349** (0.018)	0.349** (0.019)	−0.250 (3.110)	−2.442 (5.165)	0.571 (0.361)	−0.005 (0.021)	0.032 (0.023)	−0.005 (0.026)	−0.015 (0.038)	−0.006 (0.029)	0.017 (0.011)
Observations	6294	6294	5878	5877	1579	6294	6294	6294	6294	6294	6294
Adj. R^2	0.574	0.584	0.413	0.062	0.273	0.072	0.033	0.047	0.004	0.011	0.024
Controls?	No	Yes	No	No	No	No	No	No	No	No	No

* Statistical significance at 5%. Note: Robust standard errors are in parentheses, clustered on discrete values of T_i^1. All regressions control for a piecewise quadratic polynomial of T_i^1 (see text for details). Column (2) controls for math and verbal SAT scores, ACT scores, and dummy variables indicating missing values of these variables. Controls also include dummy variables indicating non-traditional students, groups of race and ethnicity, and the year in which the quantitative assessment was taken.

** Statistical significance at 1%.

lege settings, this is the only variable recorded in official databases, but its use in regression-discontinuity designs is potentially problematic.[9] Our histogram suggests why: there is a significant notch of the distribution, just below 9, that has been (non-randomly) removed. It is the result of voluntary re-test students increasing their maximum scores. The bottom-right panel confirms that almost all students with maximum scores below 9.5 take the course.

At first blush, the sharp variation in course-taking induced by the "maximum score" assignment variable seems to make it a good candidate for implementing a discontinuity design. However, the voluntary re-testing and re-sorting of students around the threshold could introduce correlations between the maximum test score and unexplained outcomes (e.g., less motivated students are less likely to voluntarily re-test, more likely to take the course, and less likely to obtain high grades). It is a special case of the common phenomenon of assignment variable manipulation in the regression-discontinuity design (McEwan & Shapiro, 2008; McCrary, 2008). In a robustness check, we will report estimates of Eqs. (2) and (3) that use $\max(T_i^1, T_i^2)$ as the assignment variable instead of T_i^1. The results suggest that course-taking could actually have an adverse effect on student outcomes, but these estimates are likely biased by endogenous re-testing.

6.2. Smoothness of student variables around cutoffs

Fig. 2 reports an important test of whether sharp variation in course-taking probabilities is exogenous. In each panel, the circles show mean values of student variables taken within discrete values of T_i^1, in addition to fitted values of piecewise quadratic polynomials. The panels reveal that students with lower assessment scores are more likely to have lower entrance exam scores, and more likely to be African-American. This is consistent with the descriptive statistics in Table 1, suggesting that course assignment rules disproportionately "tracks" relatively lower-ability and non-white students into separate classes.

However, there is no evidence of sharp breaks in these student characteristics close to the passing cutoff at 9. To the contrary, it appears that students just below and above the cutoffs are observably similar (and, by implication, similar in their unobserved attributes that affect student outcomes). Table 2 confirms this by reporting point estimates of the size of the difference at the cutoff; the estimates are based on Eq. (2) and successively use each student background variable as the dependent variable. The coefficients are generally small and statistically insignificant, lending credibility to our empirical strategy.

As the sample sizes in Tables 1 and 2 indicate, some students do not report SAT or ACT scores. This raises the possibility that sample selection in test-taking close to the threshold artificially creates the appearance of smoothness.

[9] As in this paper, Martorell and McFarlin (2007) focus on students' first test score as the assignment variable in a regression-discontinuity design. Calcagno and Long (2008) only have access to students' most recent placement score, but attempt to identify a subset of institutions where re-testing is not commonly employed.

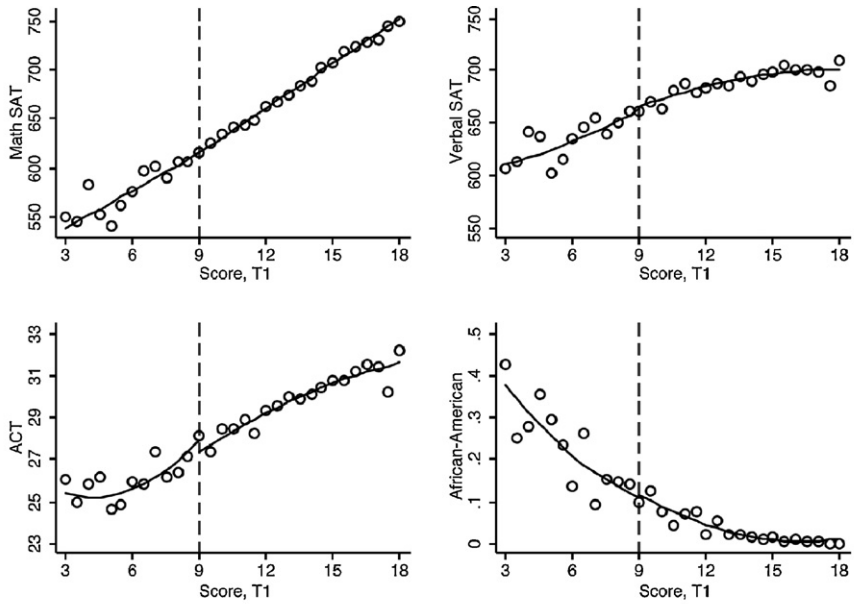


Fig. 2. Quantitative assessments and students' background variables. Note: Circles indicate the mean values of *y*-axis variables, taken within discrete values of the quantitative assessment variable. Lines indicate fitted values of quadratic polynomials estimated on either side of the cutoff.

In results not reported here, we compared the proportion of students reporting SAT and ACT scores near cutoffs, and found no evidence of a sharp break in sample selection near the threshold. In subsequent regression estimates that condition on student variables, we include dummy variables for missing values of these variables, to avoid eliminating any observations from regressions.

6.3. Effects on student outcomes

We use four measures of college performance, as described in Section 4. These include (1) each student's cumulative grade point average (GPA), (2) each student's proportion of courses with a low grade (C- or below), (3) each student's grade in the mandatory QR "overlay" course,

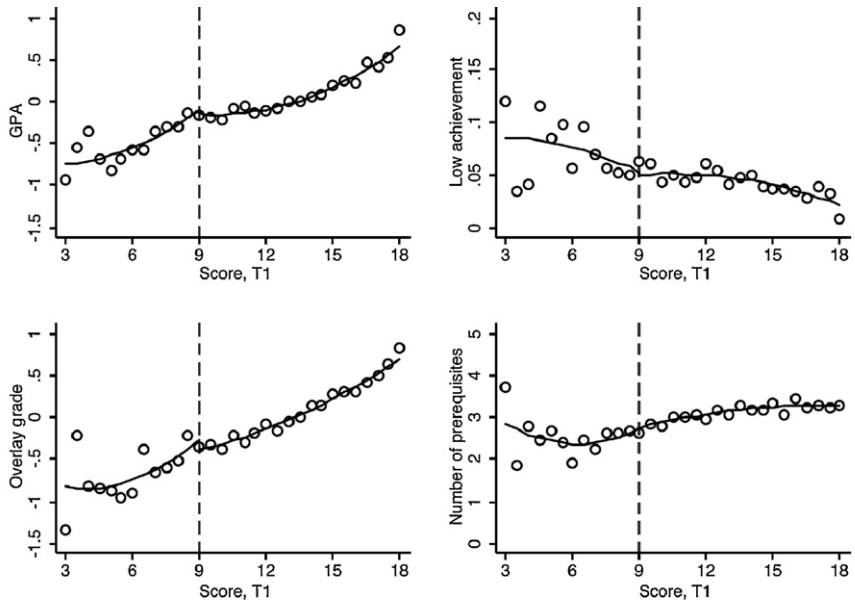


Fig. 3. Quantitative assessments and students' academic outcomes. Note: Circles indicate the mean values of *y*-axis variables, taken within discrete values of the quantitative assessment variable. Lines indicate fitted values of quadratic polynomials estimated on either side of the cutoff.

and (4) the number of subsequent courses taken with quantitative content, for which the basic skills component is a prerequisite.

Fig. 3 summarizes the main results, graphing each dependent variable against the first assessment score. Students with the highest assessment scores have GPAs that are about 1.5 standard deviations higher, on average, than students with the lowest scores. We are more interested, however, in whether students just below 9 (those more likely to take the course) have sharply higher GPAs than students just above. For GPA and other dependent variables, there is little visible evidence of breaks, suggesting no course effect on GPA for students close to the cutoff.

The empirical analogue to Fig. 3 is reported in the reduced-form estimates of Table 3, panel A (they are based on Eq. (2), using each student outcome as a dependent variable). The coefficients in odd columns (1)–(8) are the magnitudes of the breaks in Fig. 3; they are usually small and none are statistically significant. (The insignificant point estimate on overlay grade is a modest 12–13% of a standard deviation of the grade distribution.) The even columns add further controls for student variables, but the results are generally insensitive. These reduced-form estimates, reflecting the impact of scoring below 9, are akin to an intent-to-treat effect. Not all students necessarily comply with their initial course assignment, given the possibility of re-testing.

Thus, panel B reports two-stage least squares (TSLS) estimates, based on the joint estimation of Eqs. (2) and (3), that indicate the effect of actually taking the course on student outcomes, at least among the subgroup of students who comply with the initial assignment. It is simply a scaled-up version of the reduced-form coefficients. Not surprisingly, then, the TSLS estimates in columns (1)–(8), panel B are still mostly small, and all are statistically insignificant. Even so, the point estimate on overlay grade is now substantial (40% of a standard deviation).

To summarize, we find little evidence of a causal impact of the QR basic skills course on college outcomes measures. In general, however, the TSLS estimates are imprecisely estimated. For example, although we cannot say that the impact of exogenous assignment to the basic skills course on GPA for students around the threshold is statistically different from zero, we also cannot say that it is statistically different from an effect of 0.43 standard deviations.[10]

6.4. Evidence on classroom peer groups

Martorell and McFarlin (2007) raise the possibility that small or zero estimates of remedial class impacts could mask two countervailing effects: a positive effect of a course treatment (including the curriculum and instruction), and a negative effect of being assigned during one's first year to a class with lower-achieving peers.[11] For example, exogenous assignment to the course could encourage

Table 3
Reduced-form and two-stage least squares estimates.

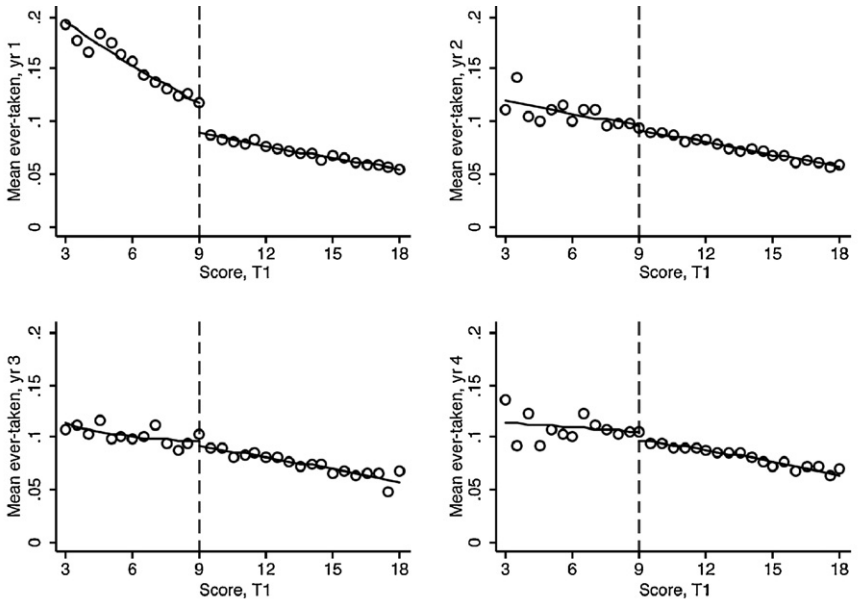
	Dependent variable:															
	Grade point average		Proportion low grades		Overlay grade		Number of quantitative courses taken		Prop. of "ever-takers" in first year		Prop. of "ever-takers" in second year		Prop. of "ever-takers" in third year		Prop. of "ever-takers" in fourth year	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Panel A: Reduced-form																
$1\{T^1 \leq 9\}$	0.027	0.003	0.003	0.003	0.123	0.128	−0.025	0.066	0.029**	0.030**	0.003	0.004	0.006	0.005	0.008**	0.008**
	(0.065)	0.035	(0.010)	(0.011)	(0.100)	(0.114)	(0.104)	(0.084)	(0.002)	(0.002)	(0.002)	(0.002)	(0.005)	(0.004)	(0.002)	(0.002)
N	6287	6287	6293	6293	4400	4400	6294	6294	6294	6294	5390	5390	3876	3876	3754	3754
Adj. R^2	0.057	0.167	0.010	0.031	0.095	0.135	0.019	0.199	0.344	0.418	0.125	0.213	0.071	0.127	0.074	0.123
Panel B: TSLS																
Took course	0.101	0.078	0.008	0.009	0.386	0.408	−0.073	0.190	0.082**	0.086**	0.010	0.012	0.017	0.013	0.024**	0.023**
	(0.188)	(0.182)	(0.029)	(0.031)	(0.319)	(0.373)	(0.297)	(0.248)	(0.007)	(0.005)	(0.006)	(0.007)	(0.012)	(0.012)	(0.007)	(0.007)
N	6287	6287	6293	6293	4400	4400	6294	6294	6294	6294	5390	5390	3876	3876	3754	3754
Controls?	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

* Statistical significance at 5%. Note: Robust standard errors are in parentheses, clustered on discrete values of T_i^1. In panel A, regressions control for a piecewise quadratic polynomial of T_i^1 (see text for details). Even columns control for math and verbal SAT scores, ACT scores, and dummy variables indicating missing values of these variables. Controls also include dummy variables indicating non-traditional students, groups of race and ethnicity, and the year in which the quantitative assessment was taken. In panel B, each cell is the coefficient from a separate TSLS regression, with the dummy variable $1\{T^1 \leq 9\}$ as the excluded instrument, following Eqs. (2) and (3). First- and second-stage regressions control for a piecewise quadratic polynomial of T_i^1.

** Statistical significance at 1%.

[10] See Table 3, panel B, column (2). The standard error is 0.182, multiplied by 1.96, and added to the coefficient estimate of 0.078.

[11] A growing literature uses experimental and quasi-experimental techniques to assess the impact of peers in post-secondary education (especially in quasi-random roommate pairings); this evidence is some-



Note: Circles indicate the mean values of y-axis variables, taken within discrete values of the quantitative assessment variable. Lines indicate fitted values of quadratic polynomials estimated on either side of the cutoff.

Fig. 4. Quantitative assessments and students' classroom peers during 4 years. Note: Circles indicate the mean values of *y*-axis variables, taken within discrete values of the quantitative assessment variable. Lines indicate fitted values of quadratic polynomials estimated on either side of the cutoff.

associations with classmates who, by definition, have below-average quantitative skills. This, in turn, could affect students' longer-term preferences for and performance in quantitatively oriented courses. These potential countervailing curriculum and peer effects may explain the zero impact we estimate for academic outcomes.

In the present setting, we cannot separately identify a true course effect from a peer effect (the prior estimates combine both). However, we can identify whether exogenous assignment to the basic skills course affected the subsequent composition of classroom peers. For each student, and in each of her college years (1–4), we calculate the proportion of her classmates during that year who ever took the basic skills course. The panels in Fig. 4 graph the average of these proportions against scores on the first quantitative assessment.

We first note that there is a clear negative correlation: in all years, students with higher assessment scores attend classes with lower proportions of QR course "ever-takers." Of course, this could reflect shared course preferences among students with similar assessment scores. We are more interested in whether exogenous assignment to the course exerts an additional influence on one's likelihood of attending class with "ever-takers" of the QR course. Among first-year students (the upper-left panel), students to the left of the passing cutoff attend class with a sharply higher proportion of "ever-takers." The large first-year effect is mechanical, since these peers include the basic skills course itself. But even in subsequent years, especially the fourth, a small gap seems to persist.

In panel B of Table 3 (columns (9)–(16)), the TSLS estimates suggest that taking the basic skills course increase the proportion of "ever-taker" classroom peers by about 8 percentage points in the first year. The effects are not statistically significant in years 2 and 3, but during students' senior years, the effect is just over 2 percentage points and statistically significant. To put this magnitude in context, the mean proportion of course "ever-takers" in fourth-year classes for the overall sample is 9%. Thus, early assignment to the skills course increases students' fourth-year "exposure" to classmates with lower quantitative skills by 22%. As with the results for academic performance, the discontinuity design only tells us what the effect is for students close to the threshold. The peer group shift might be larger or smaller for students at other points in the quantitative skills distribution.

Also bear in mind that we do not know whether this exogenous effect on classroom peer composition is larger or smaller than one would find with other first-year courses. Unfortunately, we have no credible source of exogenous variation with which to assess the impact of assignment to, for example, a particular introductory calculus class. What these results do suggest is that exogenous assignment to the QR basic skills course increases a student's classroom association with other students who have low quantitative skills, over and above levels produced merely by shared preferences or abilities. Of course, we also do not know whether this shifting of peer groups has a adverse impact on students' academic outcomes; we only know that the peer groups do shift due to assignment to the course—if the peer group shift lowers academic achievement, then

what mixed on the magnitude of peer effects and the appropriate functional form (McEwan and Soderberg, 2006).

Table 4
TSLS estimates with alternate specifications and samples.

	Dependent variable:							
	Grade point average	Prop. low grades	Overlay grade	No. of quant. courses taken	Prop. of "ever-takers" in first yr.	Prop. of "ever-takers" in second yr.	Prop. of "ever-takers" in third yr.	Prop. of "ever-takers" in fourth yr.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Baseline								
Took course	0.078 (0.182)	0.009 (0.031)	0.408 (0.373)	0.190 (0.248)	0.086** (0.005)	0.012 (0.007)	0.013 (0.012)	0.023** (0.007)
Observations	6287	6293	4400	6294	6294	5390	3876	3754
Panel B: Piecewise linear spline								
Took course	0.576** (0.161)	−0.018 (0.022)	0.670* (0.277)	−0.585* (0.207)*	0.074** (0.005)	0.005 (0.006)	0.005 (0.010)	0.012 (0.006)
Observations	6287	6293	4400	6294	6294	5390	3876	3754
Panel C: Piecewise cubic spline								
Took course	0.096 (0.170)	0.035 (0.031)	0.378 (0.340)	−0.029 (0.217)	0.092** (0.006)	0.009 (0.007)	0.020 (0.011)	0.020 (0.011)
Observations	6287	6293	4400	6294	6294	5390	3876	3754
Panel D: Restricted sample (± 3 points around threshold)								
Took course	0.213** (0.061)	−0.019 (0.012)	0.064 (0.383)	0.186 (0.218)	0.089** (0.006)	0.008 (0.013)	0.019* (0.008)	0.025** (0.006)
Observations	2268	2271	1456	2272	2272	1947	1404	1383
Panel E: Restricted sample (first quartile of math SAT)								
Took course	0.135 (0.171)	−0.006 (0.024)	0.552 (0.368)	0.206 (0.350)	0.086** (0.006)	0.017 (0.009)	0.018 (0.011)	0.021 (0.011)
Observations	1551	1553	1010	1554	1554	1359	989	1038
Panel F: Restricted sample (second quartile of math SAT)								
Took course	−0.061 (0.281)	0.010 (0.038)	0.105 (0.841)	−0.082 (0.409)	0.070** (0.009)	0.011 (0.014)	−0.015 (0.040)	0.034* (0.015)
Observations	1686	1686	1195	1686	1686	1457	1020	1033
Panel G: Re-test sample using T^2 as assignment variable								
Took course	−0.046 (0.118)	0.006 (0.017)	−0.032 (0.245)	−0.232 (0.319)	0.074** (0.004)	−0.001 (0.007)	−0.006 (0.007)	0.020 (0.011)
Observations	529	530	307	530	530	450	328	322
Panel H: Original sample, using max(T^1,T^2) as assignment variable								
Took course	−0.092 (0.047)	0.001 (0.009)	−0.157* (0.063)	0.104 (0.110)	0.081** (0.002)	0.006 (0.004)	0.004 (0.009)	0.014* (0.006)
Observations	6289	6295	4400	6296	6296	5391	3877	3755

Robust standard errors are in parentheses, clustered on discrete values of the assignment variable. Each cell is the coefficient from a separate TSLS regression, with the dummy variable $1\{T^1 \leq 9\}$ as the excluded instrument. Except for the noted modification in panels B and C, regressions control for a piecewise quadratic polynomial of T_i^1 (see text for details). All regressions control for math SAT scores, verbal SAT scores, ACT scores, and dummy variables indicating missing values of these variables. Controls also include dummy variables indicating non-traditional students, groups of race and ethnicity, and the year in which the quantitative assessment was taken.

* Statistical significance at 5%.

** Statistical significance at 1%.

that would at least partially offset positive effects of the course and help explain why we are estimating no impact of the QR course on academic outcomes.

6.5. Robustness

Table 4 presents additional estimates with varying specifications and in several subsamples. Panel A repeats the baseline TSLS results that include controls for student characteristics (taken from even columns of Table 3, panel B). Panels B and C show the results of fitting a piecewise linear spline and piecewise cubic spline, respectively, to the assessment score on either side of the threshold. Panel D shows the results of restricting the sample to those scoring within three points of the passing threshold. For student outcome variables in columns (1)–(4), there are some statistically significant coefficients, but little evidence of a robust pattern that would overturn our previous conclusions. For the peer variables in columns (5)–(8), in contrast, the results are substantively similar.

Panels E and F examine whether there are heterogeneous effects among different groups of students, focusing on the two lowest College quartiles of math SAT scores (since these students are the vast majority of students near the passing cutoff). Panel E shows the baseline specification for students with SAT scores in the lowest quartile, and Panel F shows the baseline results for students with SAT scores in the second quartile. There are no significant coefficients on the college performance variables, but with even stronger caveats about sample size and power. Again, the results on peer variables are robust.

Panel G reports an alternate specification based exclusively on the smaller re-test sample, as described earlier and illustrated in the middle panels of Fig. 1. Not surprisingly, there are no statistically significant coefficients on the performance variables. However, the peer coefficients are similar, and the fourth-year coefficient is at the margin of statistical significance. Finally, Panel H reports TSLS estimates within the original sample, but using the maximum assessment score (including the first and the optional retest). We argued above that this analysis could introduce bias from endogenous re-test sorting around the cutoff. While still largely inconclusive, the results reverse the signs of prior estimates (on GPA and the "overlay" grade).

7. Conclusions

We identify the causal effect of a course in quantitative skills on academic performance and classroom peer groups, using a regression-discontinuity design. Wellesley College requires that all first-year students take a quantitative skills assessment exam during orientation period. If students score below a passing cutoff, they must take a QR basic skills course during their first year. We compare outcomes for students just below and just above this cutoff. While students close to the cutoff are observably similar, they have sharply different probabilities of taking the course.

We find no impact of taking the QR basic skills course on outcomes that include overall GPA, the overall proportion of courses with low grades, the grade in the subsequent QR "overlay" course, and the number of subsequent courses taken with quantitative content. These conclusions only apply to the subgroup of students with assessment scores close to the passing cutoff. The conclusions are further tempered by the imprecision of many of the coefficients, and we cannot discard the possibility that we are simply unable to distinguish small effects in our dataset.

However, we do find robust effects of the course on classroom peer-group composition. Even in their senior years, students who are exogenously induced to take the QR basic skills course take courses with a higher fraction of other students who also took the QR course. This is not simply an artifact of students with similar academic abilities or preferences choosing courses that appeal to them.

Our results leave unanswered many questions for research and policy. The previous findings can be generalized to the subgroup of students who are close to the cutoff, and who comply with the initial course assignment. At Wellesley College, these students have fairly high math SAT scores. We cannot say whether the same results would be obtained among students with lower baseline quantitative skills (the same problem is faced by the larger-scale discontinuity studies discussed in Section 2). To do so would require a different research design, such as a randomized experiment in the spirit of Angrist et al. (2009), which was able to identify average effects among *all* students offered the program.

Absent such evidence, the results might at least seem to call for a reduction in the passing threshold, since resource savings could then be used to apply a more intensive version of the quantitative skills training (e.g., smaller class sizes or more instructional hours). However, we are very cautious about making such a recommendation. First, such a decision could unwittingly sacrifice benefits for excluded students, since our research design is not able to distinguish small or even modestly sized effects for students at the current threshold.

Second, such a decision would, at the margin, intensify "tracking" of first-year students by their baseline quantitative skills (and correlated variables, such as race). This could potentially alter the observed patterns of peer-group sorting in the longer-run, although it is not clear how this would affect student outcomes. On the one hand, it may further limit classroom mixing of students with heterogeneous skill levels and backgrounds (which is an implied goal of a diverse liberal arts college), and prevent positive classroom spillovers from higher-ability to lower-ability students. On the other hand, friendships formed in an even smaller and more intense basic skills course could build social capital and improve students' experiences at the College.

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), 136–163.

Bettinger, E. P., & Long, B. T. (2009). Addressing the needs of underprepared students in higher education: Does college remediation work? *Journal of Human Resources*, *44*(3), 736–771.

Calcagno, J. C., & Long, B. T. (2008). The impact of postsecondary remediation using a regression discontinuity design: Addressing endogenous sorting and noncompliance. Working Paper 14194, National Bureau of Economic Research.

College Board SAT. (2007). 2007 College-Bound Seniors: Total Group Profile Report. College Board. Downloaded February 22, 2009 from: http://www.collegeboard.com/prod_downloads/about/news_info/cbsenior/yr2007/national-report.pdf.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615–635.

Koski, W. S., & Levin, H. M. (1998). *Replacing remediation with acceleration in higher education: Preliminary report on literature review and initial interviews*. Stanford, CA: National Center for Postsecondary Improvement, Stanford University.

Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. house elections. *Journal of Econometrics*, *142*(2), 675–697.

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

Lesik, S. A. (2006). Applying the regression-discontinuity design to infer causality with non-random assignment. *The Review of Higher Education*, *30*(1), 1–19.

Lesik, S. A. (2007). Do developmental mathematics programs have a causal impact on student retention? An application of discrete-time survival and regression-discontinuity analysis. *Research in Higher Education*, *48*(5), 583–608.

Levin, H. M., & Calcagno, J. C. (2008). Remediation in the community college: An evaluator's perspective. *Community College Review*, *35*(3), 181–207.

Martorell, P. & McFarlin, I. (2007). Help or hindrance? The effects of college remediation on academic and labor market outcomes. Unpublished manuscript, RAND and University of Texas at Dallas.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*(2), 698–714.

McEwan, P. J., & Shapiro, J. S. (2008). The benefits of delayed primary school enrollment: Discontinuity evidence using exact birth dates. *Journal of Human Resources*, *43*(1), 1–29.

McEwan, P. J., & Soderberg, K. A. (2006). Roommate effects on grades: Evidence from first-year housing assignments. *Research in Higher Education*, *47*(3), 347–370.

Taylor, C. (2006). Quantitative reasoning at Wellesley College. In R. Gillman (Ed.), *Current Practices in Quantitative Literacy* (pp. 141–146). Mathematics Association of America.