Title

Do learning communities increase first year college retention? Evidence from a randomized control trial

Permalink

https://escholarship.org/uc/item/3t43v85b

Authors

Azzam, Tarek
Bates, Michael D
Fairris, David

Publication Date

2022-08-01

DOI

10.1016/j.econedurev.2022.102279

Copyright Information

Do Learning Communities Increase First Year College Retention?
Evidence from a Randomized Control Trial

Tarek Azzam, Michael D. Bates, and David Fairris

May, 2022

Abstract:

In this paper, we estimate the impact of a learning community on first-year college retention at a four-year public research university using a randomized control trial (RCT) for those students who opt into the experiment. Intent-to-treat and local-average-treatment-effect estimates reveal no discernable programmatic effects. We also generate estimates of program impact using observational techniques and find estimated impacts that are positive, large and statistically significant. We explore a variety of selection processes to better understand the differences between the RCT and observational estimates and to reflect on the generalizability of the RCT results for various other populations of interest. Non-random selection into the experimental sample accounts for the major difference in the two estimates and also cautions against generalizing the RCT result for populations outside the experiment. Keywords: higher education, experimental design, generalizability, selection on unobserved variables.

Contact: Azzam: Gevirtz Graduate School of Education, University of California, Santa Barbara, Santa Barbara, CA, 93106 (email: tarekazzam@ucsb.edu); Bates: (corresponding author) Department of Economics, University of California, Riverside, Riverside, CA 92521 (email: mbates@ucr.edu); Fairris: Department of Economics, University of California, Riverside, Riverside, CA 92521 (email: dfairris@ucr.edu). Acknowledgements: We acknowledge the able research assistance of Melba Castro and Amber Qureshi. David Fairris acknowledges support from the "Fund for the Improvement of Post-Secondary Education" at the U.S. Department of Education, grant number P116B0808112. The contents of this paper do not necessarily represent the policy of the Department of Education. This experiment is registered at the Registry for Randomized Controlled Trials under the number AEARCTR-0003671. The authors are prepared to provide all data and code for purposes of replication. This work was conducted under exempted IRB approval through the University of California, Riverside.

Introduction

In the past few decades, colleges have responded to the challenge of improving first-year college retention by creating freshmen year learning communities (Pitkethly and Prosser, 2001; Kyndt et al., 2017; Xerri, Radford, and Shacklock, 2018).[1] Learning communities bring together small groups of students, typically into thematically-linked courses for at least one term during freshmen year, in the hopes that students will better engage with course material, support one another socially and academically, and thereby enhance academic success, first-year retention, and ultimately graduation. An independent study in 2010 by the John N. Gardner Institute for Excellence in Undergraduate Education found that 91% of reporting institutions claimed to possess a learning community of some form or another at their institution (Barefoot, Griffin, and Koch, 2012).

We utilize a randomized control trial (RCT) design, linked to student record data, to explore the extent to which a First-Year Learning Community (FYLC) program at a four-year research university increases first-year student retention. At the outset, and during the period for which our analysis takes place, students voluntarily enrolled in the program, and so randomization occurs within this self-selected student population. We offer “intent to treat” (ITT) estimates of the effect of being randomized into treatment from among students in this population. Though there is relatively high compliance with the randomization, some students who were randomly assigned to the program ended up not taking it, and some students who were not assigned to the program made their way into the program nonetheless. Due to this two-sided noncompliance with the randomization we also estimate the “local average treatment effect” (LATE) of the program’s impact among those compliers who are moved to treatment or non-treatment by the randomization. The ITT and LATE estimates of program impact reveal no

[1] First-year retention rates vary significantly across higher education institutions and institutional types. For full-time students, first-year retention rates are close to 80% at four-year public and private institutions, and close to 50% at two-year institutions (U.S. Department of Education, 2017). At elite four-year institutions, first-year retention can be as high as 99%, whereas at lesser-known regional institutions that award four-year degrees, first-year retention rates can be as low as 40% (U.S. News and World Report, 2018).

statistically significant effect on first-year retention. This is the first study of which we are aware to generate estimates from an RCT design of the impact of a learning community on first-year retention at a four-year university.

In the absence of randomization, many researchers in this literature employ observational analyses to study program impact, conditioning on variables available in student records and using methods such as OLS, nonlinear estimation, and propensity score matching on the full sample of freshmen to estimate the effect of the FYLCs on first-year retention. We mimic this exercise to explore whether these estimates closely match the results found in the RCT. We find stark differences in estimated impacts between the two approaches, with the observational methods revealing large and statistically significant benefits of the program and the RCT uncovering little discernable impact.

We explore a variety of selection processes to better understand the differences between the RCT and observational estimates and to reflect on the generalizability of the RCT results for various other populations of interest. As it happened, this last question is not a matter of idle curiosity; shortly after the completion of our analysis, and based largely on the positive results of an in-house observational analysis of program impact, the FYLC was expanded to populations beyond those who opt-in voluntarily.

We begin this exploration by comparing retention rates, based on both observable and unobservable attributes, for the untreated within the experiment and the larger group of students who did not select into the experiment at all. Next, because there was a group of students who made their way into treatment even though they did not participate in the RCT, we consider whether they have similar outcomes to those who participated in the randomization before receiving treatment. And, finally, we consider students who voluntarily enrolled in the experiment but did not comply with the randomization by either not showing up to their assigned treatment or by crossing over from assigned control to taking treatment.

Our results reveal that those students who express a desire to enroll in the RCT are, in many observed respects, selected from more vulnerable segments of the student population –

they tend, for example, to have lower high-school GPAs, lower SAT scores, and come from less-advantaged backgrounds. However, we also show that the experimental population possess unobserved characteristics – presumably, things like grit, determination, focus, and commitment – which make them even more likely to succeed in college than their peers who did not enroll in the study. This positive selection on unobserved variables holds for both those who do and do not receive treatment. Moreover, we find no evidence of nonrandom noncompliance within the experiment.

Nonrandom selection into the RCT accounts for the positive observational estimates of the program's impact on retention. This positive selection also clearly demonstrates that the experimental population is not representative of the remainder of the freshman class. Given this positive selection on unobserved variables persists among both treated and untreated, we believe it would be unreasonable to assume the LATE generalizes to the population of students outside the experiment. Indeed, earlier work establishes the absence of unobserved differences in the outcome as a criterion for the generalizability of results. However, the absence of selection into noncompliance within the experiment suggests the LATE is likely to generalize to the entire experimental population.

The paper is organized as follows: First, we describe in greater detail the learning community literature, the FLYC at this institution, the nature of the randomized control trial design, and the data to be used in the analysis. Second, we describe the empirical methodology, followed by the results. The final section offers a summary discussion and conclusion.

Background and Data

While there are many published evaluations of first-year learning experience programs broadly defined and still more conducted by in-house researchers, the set of studies focused on such first-year learning communities is restricted.[2] Of those focused on learning communities

[2] See Barefoot, et al. (1998) and Pascarella and Terenzini (2005) for early reviews and Angrist, Lang, and Oreopoulis, (2009), Bettinger and Baker, (2014), and Paloyo, Rogin, and Siminski, (2016) for more recent examples of evaluations of first-year experience programs more generally.

specifically, some observational studies utilize advanced techniques, such as propensity score matching (Clark and Cundiff, 2011; Provencher and Kassel, 2019), instrumental variables (Pike, Hansen, and Lin, 2011; Hansen and Schmidt, 2017), Heckman's two-step procedure (Hotchkiss, Moore, and Pitts, 2006), and longitudinal analysis (Millea et al. 2018). However, in each, the exogeneity assumptions necessary for causal interpretation of the results may be problematic.

There are some RCT studies that also estimate the impact of learning communities on various programmatic outcomes. Two of these studies estimate the impact of remedial learning communities on retention rates in two-year community college settings (Scrivener et al. (2008) and Visher et al. (2012). Both find small positive effects on performance in remedial courses, though no effects on first-year retention. Interestingly, Scrivener et al. (2008) find in a two-year follow-up study that program participants were 5 percentage points more likely still to be pursuing their degree than control group members. However, causal effects identified in the community college setting are likely to differ from those at four-year institutions. Community colleges typically draw differentially from the academic and soft-skills distributions. Four-year universities also tend to provide more opportunities for a community to develop naturally through on-campus housing and additional extra-curricular programs. Consequently, the effects of learning communities on retention at four-year institutions warrants further examination.

The third RCT evaluation of learning communities provides the closest study to the one at hand. Russell (2017) examines the effects of experimental study groups at the Massachusetts Institute of Technology. While the overall effects on program participants are of mixed sign, small in magnitude, and noisy, subgroups of participants do display large, positive, marginally statistically significant program effects on some outcomes such as GPA and majoring within a STEM field. First-year retention was not an outcome variable that was evaluated in this study and effects on male, racial majority, and high income students are not reported.

The First-Year Learning Community (FYLC) we study exists in the largest college of a major, four-year public research university with now over 20,000 students. The student population is very diverse, ethnically, racially, and socio-economically. Nearly 40 percent of entering freshmen are from underrepresented backgrounds (African American, Native American, and Chicano/Latino), over 50% would qualify as first-generation college graduates, and roughly one-third come from family incomes under $30,000. While in many respects the university is representative of four-year public, and even some private, institutions of higher education, in other respects, especially in regards to the diversity of the student population, it is arguably less representative.

The FYLC program began on a small scale, with roughly 200 students, at a time when the campus was growing rapidly and there was a sense on campus that students – and freshmen in particular – were facing larger and more impersonal classes. The program is modeled after learning community programs in which two or more courses are linked around a specific theme (Laufgraben, Shapiro and Associates, 2004; Kuh, Kinzie, Schuh, Whitt and Associates, 2005; Zhoa and Kuh, 2004). The general format may vary across institutions but the basic idea is similar and the intention is the same: that students will better engage with course material, support one another socially and academically, and thereby enhance academic success, first-year retention, and ultimately graduation. Increasing first-year retention rates was the primary institutional measure of the FYLC's success according to the founding Associate Dean and Director of the program under study.

The FYLC program at this institution possesses a variety of curricular and extra-curricular activities for enrolled freshmen. The primary curricular piece is a set of year-long thematic courses – each course based, for example, on a theme such as "Human Rights" or "Justice" – and participating students enroll in one thematic course for their entire freshman year. This course constitutes one of four or perhaps five courses taken by a student each term. A FYLC course is smaller than an average lecture course at the University – 75 students rather than the typical 100-500 students – and is taught by three different ladder-ranked faculty members, typically from three different disciplinary fields, for each of the three freshman year terms. As with most large lecture courses, students in FYLC courses break up into "discussion sections" once a week with a graduate student teaching assistant. However, every effort is made to have both the group of students and the T.A. remain together throughout all three terms of the course. In addition to these curricular aspects, the program offers extra-curricular events – including trips to museums and college basketball or soccer games – throughout the year. Students in the FYLC program are obligated to engage in behavioral activities that are meant to enhance academic success – such as meeting with professors and teaching assistants, separately, at least once during each term – and to attend both academic and cultural events on campus, such as a faculty lecture and a concert.

With the help of a Fund for the Improvement of Post-Secondary Education (FIPSE) grant from the Department of Education, student capacity in the FYLC was increased over a two-year period in order to provide sufficient sample size for analysis. The random assignment feature was institutionalized in the following way: Program staff solicited intent to participate commitments from incoming freshmen following communications about the program to both parents and students prior to freshman orientation. Every entering freshman received the same information about the program and was encouraged to enroll in the lottery in order to participate. The goal was to receive expressions of interest by 1,000 incoming freshmen each year, 450 of whom would then be randomly assigned to the available program seats and the others would be assigned to the control condition. This would allow us to detect an effect of about 0.05 change in first-year college retention at a power of 0.9, similar to that detected in Scrivener et al. (2008).

The new random assignment regime roughly approximates the old program implementation procedure, but with several differences that conceivably could have affected program participation and program outcomes. Under the former regime, program participants were essentially drawn from among the self-selected student population (i.e., those who would have expressed an intent to enroll had they been asked) on a "first-come, first-served basis" during consecutive course enrollment sessions at freshmen orientation. Under the new regime, treatment is randomly assigned from the self-selected population and given permission to enroll in the program in advance of orientation. Under the old regime, non-participants were either unaware of the program or found that the FYLC classes were filled if they tried to enroll in them during orientation. Under the new regime, the control population knew of the program and elected to be enrolled in it, but was never notified that they had “lost” the lottery, only possibly finding out that this was the case if they attempted and were unable to enroll in the program during course enrollment at orientation. Finally, under random assignment students and parents were given greater opportunity to discuss the program before being given an opportunity to enroll.

A fundamental assumption of RCTs is that the behavior of the control population is not affected by the knowledge that they are part of an experiment in which they will be compared to the treated population and so act differently than they normally would to overcome a perceived disadvantage. As is the case in many experimental studies, we cannot guarantee that this assumption holds in the present case. However, several factors lead us to believe that it might. First, study participants were unaware of the RCT analysis itself – that there would be treated and control samples and an evaluation of certain outcomes on each. Instead, they were informed that random assignment would be used to allocate scarce space to students interested in the program. Second, there are various opportunities on campus for students to enhance their academic performance and foster greater connections with fellow students and the institution more generally, however, none of these programs is in great supply and none resembles the full activities and support available through the FYLC program.

Among the widely available opportunities for enhancing academic performance on campus, there is tutoring in select courses and note-taking and test-taking skills tutorials conducted at the Learning Center. Freshmen who are struggling academically at the end of their first term are offered a unit-bearing course that imparts some of these same skills and requires that students engage more academically by, for example, visiting their professors during office hours. There are also plenty of opportunities to become more socially connected to fellow students, such as fraternities and sororities, and to the institution more generally, such as concerts and art openings. However, apart from the Honors Program, which accepts a very small number of high-performing freshmen to engage with ladder faculty in small-group seminar settings, there is nothing resembling the FYLC program for freshmen at this institution. The Honors Program solicits applications and admits students well before FYLC applicants are informed of their allocation status.

Data for this analysis come from student records on the two freshman cohorts during the years for which the program capacity was increased by virtue of the federal grant. A unique feature of our analysis is that in addition to retention and demographic information for the self-selected population who applied to be part of the program, we also gather information on the remainder of the freshman class who at the outset expressed no interest in program participation. Having information on the non-experimental population is unfortunately rare in RCT designs. We use this additional information to shed light on the nature of various selection issues which are impossible to explore without it.

Table 1: Student Background Characteristics

	Assigned Control	Assigned Treatment	Difference	Lottery Sample	Non-lottery Sample	Difference
High-school GPA	3.46	3.46	0.01 (0.02)	3.46	3.53	-0.07*** (0.01)
SAT math	494.25	498.65	4.40 (6.15)	496.57	544.40	-47.83*** (3.63)
SAT writing	491.42	496.40	4.98 (5.77)	494.04	508.33	-14.29*** (3.29)
SAT verbal	488.00	491.14	3.14 (5.88)	489.65	502.39	-12.73*** (3.31)
Female	0.68	0.69	0.01 (0.02)	0.69	0.50	0.19*** (0.01)
1st generation	0.63	0.62	-0.01 (0.02)	0.62	0.56	0.07*** (0.01)
Low income	0.60	0.62	0.01 (0.02)	0.61	0.56	0.05*** (0.01)
Lives on Campus	0.74	0.75	0.01 (0.02)	0.75	0.71	0.04*** (0.01)
N	741	824	1,565	1,565	6,566	8,131

Low income is defined as family income below $30,000. Robust standard errors are in parentheses.

We begin by aggregating the two cohorts into a single sample for the purpose of analysis.

This yielded a sample of 8,131 students, 1,565 of whom applied to be part of the FYLC, and 824 of whom were chosen through the lottery system to be part of the program. In addition to first-year retention (where, 1=returned for a second year at this institution, and 0=did not return), we have a host of student background characteristics from student records that are used as control variables in the analyses to follow. Table 1 lists these characteristics variables and shows their means for three primary populations of interest.

None of the background variables is meaningfully or statistically significantly different across those assigned to the treatment or control. However, this is decidedly not the case when we compare students who self-selected into the lottery with those who chose not to enter the lottery. The Table 1 results reveal that these two groups are statistically different with regard to every observed background characteristic. Moreover, with the exception of being proportionately substantially more female and slightly more likely to live on campus, the ways in which the lottery students differ would suggest they possess greater vulnerability to attrition between the first and second year of college. They possess lower SAT scores (nearly 10 percent below average for math), slightly lower high-school GPAs, and they are substantially more likely to be a first-generation college student and from a low-income family.[3]

Table 2: Populations within the setting

Name	*Condition*	*Population Share*
Compliant Treated	*R=1, Z=1, D=1*	*8.0*
Compliant Control	*R=1, Z=0, D=0*	*7.8*
No-shows	*R=1, Z=1, D=0*	*2.1*
Crossovers	*R=1, Z=0, D=1*	*1.4*
Never-ever takers	*R=0, D=0*	*79.3*
Nonrandomized takers	*R=0, D=1*	*1.3*

Notes: R indicates participation in the RCT, Z indicates assignment to treatment, and D indicates receipt of treatment (enrollment in the first-year-learning community).

As mentioned above, there are three important instances of migration between assigned groups in the data. Of the 824 students initially assigned to the treatment group, 170 did not attend any of the program courses or services, and are thus referred to as “no shows.” Of the 741

[3] We discuss this matter further below and show that each of these traits is correlated with lower retention in Table A1 in Appendix A.

students assigned to the control group, 108 students migrated to treatment (as "crossovers") and enrolled in FYLC courses (with permission of the program director, and presumably as a partial replacement for those no-shows from the assigned treated group). Finally, 117 of the 6,566 students who did not initially express an interest in the program migrated to treatment (also with permission of the program director), and are referred to as "nonrandomized takers." Table 2 describes the various populations of freshmen within the setting of our analysis.

None of these groups is a random draw from the entering freshman class. As shown in Table A2 in Appendix A, those who ultimately receive treatment are in some instances statistically significantly different from those in their original assignment category on almost every observable dimension. While none of these violations of initial assignment bias the "intent to treat" estimates of program impact, they do present complications in estimating the effects of treatment itself. However, their presence also provides opportunities to explore the extent to which our estimated LATE can be generalized to the entire experimental sample and to the remaining freshman class.

The possibility of nonrandom sample selection into RCTs has received earlier attention.[4] Our analysis contributes to a small but growing literature that combines RCTs with outcome data from observational settings to examine selection into experiments on the basis of unobserved variables.[5] The closest methodological work to the method we introduce is Hartman et al. (2015). Hartman et al. (2015) propose a test comparing the outcomes (adjusted for observed covariates) of those who receive treatment within the randomized sample to those who receive

[4] For an early example see Hausman and Wise (1979) and for recent examples see Cole and Stuart (2010), Andrews and Oster (2018), Ghanem, Hirshleifer, Ortiz-Becerra (2018), and Bo and Galiani (2021).

[5] See Tian and Pearl (2000), Bartlett et al. (2005), Prentice et. al. (2005), Karlan and Zinman (2009), Alcott (2015), Altidag et al. (2015), Gechter (2015), Hartman et al. (2015), Lise et al. (2015), Sianesi (2017), Galiani, et al. (2017), and Walters (2018) for examples.

treatment otherwise. However, their "strong ignorability of treatment assignment" may fail from selection on unobserved variables either into the experiment or into treatment in the observational setting.[6] Nonrandom selection into treatment in observational settings is common, and by itself has no bearing on the validity of RCT results. In contrast, nonrandom selection on unobserved variables, which may include differences in responsiveness to treatment, directly calls into question the representativeness of the RCT (see Janzing, Peters, and Schölkopf, 2017; and Bo and Galiani, 2021). Our test for nonrandom selection into RCTs is designed to help researchers identify nonrandom selection on unobserved variables specifically into experiments so that we can better discern whether an RCT has a representative sample, and whether its results are likely to generalize.

Empirical Methodology

We divide the empirical analysis into three sections.

Analysis 1: Estimating treatment effects using the RCT

In the first analysis, we utilize the RCT design to identify the intent to treat effect of the FYLC on first-year retention, as well as the average treatment effect on those who complied with the randomization. We maintain four assumptions which are commonly invoked when using randomized control trials. First, we assume proper randomization within the experiment, such that unobserved variables including responsiveness to treatment do not differ in expectation between those assigned to treatment and those assigned to the control group. Second, we assume no independent effect of assignment within the experiment, ruling out disappointment or compensatory effects. Third, we assume monotonicity such that assignment to treatment does not lower the probability that any participant enters the treatment. Finally, we maintain the stable

[6] The strong ignorability of treatment assignment assumption from Hartman et al. (2015) holds that the expectation of the potential outcome given treatment status and participation are equal between experimental and non-experimental participants. Violation from selection into treatment in the observational setting or selection into the experiment itself would prohibit their methods from recovering the population parameter, which is their primary objective. However, this test is not informative regarding the generalizability of RCT.

unit treatment value assumption, which holds that individuals' responsiveness to treatment is unaffected by the number of others who also receive treatment (Rubin, 1980). Given these assumptions, we are able to estimate the causal "intent to treat" effects of the program using standard approaches.

Due to the ease of interpretation, we begin by estimating a linear probability model using ordinary least squares among the population who selected into the lottery according to the following specification:

$$Y_i = \alpha + Z_i\beta_{ITT} + \boldsymbol{X_i}\boldsymbol{\gamma} + \epsilon_i, \tag{1}$$

where Y_i indicates whether student i remained in school the following year, Z_i indicates whether individual i entered and won the lottery, and $\boldsymbol{X_i}$ is a rich vector of student background characteristics discussed in the "Data" section above. As causal identification does not hinge on the covariates we conduct the analysis both with and without conditioning on $\boldsymbol{X_i}$. We prefer to include these controls because doing so generally provides more efficient estimates that remain consistent.[7] We repeat the exercise using logit to respect the binary nature of the dependent variable under a quasi-maximum likelihood estimation (QMLE) framework to obtain heteroscedasticity robust standard errors (Gourieroux, Monfort, and Trognon, 1984).

Were compliance with the lottery perfect, the average intent to treat estimate would also provide an estimate of the average effect of treatment for the experimental sample. However, due to the two-way non-compliance, estimates of the intent to treat may be misleading regarding the efficacy of treatment, because they ignore contamination of the treatment and control groups. We attempt to uncover the average effect of the treatment on the compliers using 2SLS with the lottery as an instrumental variable for enrollment in the FYLC. In the non-linear specification, we use a control function approach in which we treat the endogeneity in *FYLC* by adding the first-stage residuals in the logit estimation of equation (1) following Vytlacil (2002) and

[7] However, the inclusion of covariates may introduce finite sample bias, which may give reason to prefer the nonparametric approach. For references see Yang and Tsiatis (2001), Tsiatis et al. (2008), Schochet (2010), and Lin (2013).

Wooldridge (2014).[8] While this procedure provides us with internally-valid, causal estimates of the effect of treatment, without further assumptions these estimates hold only for the whose treatment status is determined by the randomization.

Analysis 2: Estimating treatment effects using an observational analysis

In the second analysis, we conduct a conventional observational analysis of program impact on the entire treated population, including the compliant treated, the crossovers and the nonrandomized takers. The control population includes the compliant control, the no-shows and the never-ever takers. The observational designs we consider are still commonly used by in-house institutional researchers and appear in much of the earlier-published program evaluation studies of first-year learning communities, in the context of both voluntary and mandated enrollment.

We first estimate the effect of enrollment in the FYLC on first-year retention using unconditional OLS regressions, covariate adjusted OLS regressions, and logit QMLE analysis. This analysis is similar to the that used to identify our average intent to treat estimates except that here we use the full sample of freshman entrants and treatment is measured by an indicator for enrollment in the FYLC (D_i) instead of by an indicator for winning the lottery (Z_i). We supplement this analysis by adding propensity score matching techniques, which are used by Clark and Cundiff (2011), for example, to evaluate the efficacy of a FYLC without random assignment. We estimate the average treatment effect on the treated by averaging over the difference between the retention of each treated student and the retention of the student in the remaining population who is most similar to the treated student, but did not receive treatment. We adopt the standard practice of using logit to estimate the propensity scores. We perform all analyses both on the full sample as well as the sample in which estimated propensity scores are between 0.1 and 0.9, out of respect for the overlap assumption and in accordance with the rule of thumb provided by Crump, et al. (2009).[9] We bootstrap the standard errors to account for

[8] Since the included residuals are estimated, the standard errors we use for inference must account for possible estimation error. Consequently, we bootstrap both stages to estimate the standard errors.

[9] We illustrate the overlap in these populations in Figure A1 in Appendix A.

estimation error. In all analyses, we use the same vector of observed covariates that is used in *Analysis 1* above.

Analysis 3: Decomposition, selection and generalizability

In the third analysis, we explore a variety of selection processes within the full population setting. We use a decomposition framework to better understand the differences in the RCT and observational estimates and to reflect on the external validity of the RCT results to populations beyond the compliers.

We offer a decomposition analysis of different forms of selection that can account for the difference between the RCT and observational results. Selection on unobserved variables may originate from selection into the experiment, selection into treatment of the nonrandomized takers outside the experiment, or through noncompliance within the RCT itself.[10] We begin by temporarily ignoring the imperfect compliance within the RCT, estimating the following difference-in-differences decomposition of the OLS estimate in which the intercept and treatment coefficient from the OLS estimation are allowed to differ according to whether students participated in the RCT:

$$Y_i = \boldsymbol{\alpha} + \boldsymbol{X_i\gamma} + R_i\pi_1 + D_i\pi_2 + D_i \times R_i\pi_3 + \varepsilon_i. \tag{2}$$

The reference group is composed of those who did not sign up for the RCT and never received treatment – that is, the never-ever takers. The coefficient π_1 on the RCT indicator (R_i) picks up the difference in retention between the never-ever takers and the compliant controls plus no-shows.[11] D_i is an indicator for the treated population, and the coefficient on the stand alone D_i indicator (π_2) picks up the difference in outcomes between the treated (nonrandomized takers) and untreated (never-ever takers) from outside the experiment. The estimated coefficient on the interaction of D_i with R_i (π_3) picks up the additional difference in outcomes for those treated within the experiment – the compliant treated and crossovers versus the compliant controls and

[10] See Table 2 for greater clarity on the populations we consider.

[11] An assumption inherent in the causal identification strategy of RCTs is that the behavior of the randomized control group that relates to the various outcome measures is not altered in reaction to its control condition and eventual comparison with a treated population.

no-shows – beyond the difference in outcomes between the treated and untreated outside of the RCT. Thus, π_3 is the difference-in-differences estimate, $\pi_2 + \pi_3$ provides the estimated impact of treatment within the RCT, and $\pi_1 + \pi_3$ is the difference in outcomes between the treated groups, one inside and the other outside the RCT.

This decomposition can illuminate whether the difference in results of the RCT and observational analyses are driven by different estimated effects within the nonrandomized and randomized populations or by differences in the populations themselves. For instance, imagine a setting in which the observational estimate of programmatic impact on retention are larger than those from the RCT. In that setting, observing $\widehat{\pi_2} > 0$ and $\widehat{\pi_3} < 0$ may indicate that the difference in estimates is driven by larger estimated treatment effects in the nonrandomized population than in the randomized population. A non-zero $\widehat{\pi_3}$ may indicate heterogeneous treatment effects within and outside of the RCT. However, it may just as easily also result from selection into treatment among the nonrandomized, or selection into the experiment on other unobserved variables. In contrast, $\widehat{\pi_1}$ shows the difference in expected outcomes between two groups of untreated students, and $\widehat{\pi_1} + \widehat{\pi_3}$ is the difference in expected outcomes between those who receive treatment within and outside the experiment, respectively. Statistically significant coefficients here would seem to indicate the difference in observational and RCT results may be driven by selection into the RCT itself.

We also use the results from estimating equation (2) to reflect on matters of external validity. If $\widehat{\pi_1}$ and $\widehat{\pi_1} + \widehat{\pi_3}$ are both small and statistically insignificant, this offers some assurance that there is no sign of selection into the experiment, as any differences in unobserved characteristics would load on these parameters. With both the treated and untreated within the RCT closely representing the treated and untreated individuals in the broader population, we may be more comfortable generalizing the RCT results to the larger population outside the experiment. If $\widehat{\pi_1}$ and $\widehat{\pi_1} + \widehat{\pi_3}$ are of opposite signs, this may indicate selection into treatment outside of the experiment, which prevents us from learning about selection into the experiment itself. Thus, a statistically significant $\widehat{\pi_1}$ *or* a statistically significant $\widehat{\pi_1} + \widehat{\pi_3}$ alone cannot indicate nonrandom selection into the RCT. However, finding that $\widehat{\pi_1}$ and $\widehat{\pi_1} + \widehat{\pi_3}$ are both statistically significant and possess the same signs suggests selection into the experiment on unobserved variables, which may include heterogeneous treatment effects.[12] In all of these cases, the experimental sample would not be deemed representative. Given that $\widehat{\pi_1} + \widehat{\pi_3}$ incorporates possible differences in treatment effects by RCT participation, we believe it would be unreasonable to assume homogeneous treatment effects under these circumstances. Indeed, under the definition of external validity found in both Janzing, Peters, and Schölkopf (2017) and Bo and Galiani (2021), selection on unobserved variables into a particular experiment is sufficient to indicate external invalidity.

In the above analysis, we assume that compliance with assigned randomization within the experiment is perfect. To address the various selection processes associated with this noncompliance, we further decompose the OLS effects using indicators to separate out the six populations outlined in Table 2 as shown in equation (3) below:[13]

$$Y_i = \boldsymbol{\alpha} + \boldsymbol{X_i\gamma} + R_i\delta_1 + D_i\delta_2 + D_i \times R_i\delta_3 + Z_i \times R_i\delta_4 + Z_i \times D_i \times R_i\delta_5 + \varepsilon_i. \qquad (3)$$

Here, since Z_i indicates assignment to treatment, $Z_i \times R_i$ allows us to separate out the no-shows from the compliant controls and $Z_i \times D_i \times R_i$ allows for separation of the compliant treated from the crossovers. Accordingly, $\widehat{\delta_1}$ compares the conditional outcomes of the compliant controls and never-ever takers, $\widehat{\delta_2}$ compares the nonrandomized takers to the never-ever takers, $\widehat{\delta_3}$ gives the additional estimated impacts of treatment for the crossovers relative to the estimated impacts for the nonrandomized takers, $\widehat{\delta_4}$ compares conditional outcomes for the no-shows and compliant controls, and $\widehat{\delta_5}$ indicates whether estimated non-compliance is significantly different for those who were treated versus those who were not treated.

From these coefficients we can compare any two populations. For instance $\widehat{\delta_1} + \widehat{\delta_4}$ and $\widehat{\delta_1} + \widehat{\delta_3}$ may be relevant to assessing selection into the RCT population. $\widehat{\delta_1} + \widehat{\delta_4}$ provides the

[12] See Appendix B for proof of this statement. As this is a joint hypothesis, we follow Brinch et al. (2017) in using the larger of the two p-values for statistical inference. We provide the corresponding rationale in Appendix C.

[13] This exercise is an extension of Huber's (2013) test for ignorability of non-compliance. Similar examinations of selection into compliance can be found in Angrist (2004), Black et al. (2017), Brinch et al. (2017), Kowalski (2016, 2018), and Bertanha and Imbens (2019). We add to these the sample participation margin.

comparison of no-shows to never-ever takers, both of whom avoid treatment though the no-shows at least expressed an interest by enrolling in the RCT. And $\widehat{\delta_1} + \widehat{\delta_3}$ compares crossovers to nonrandomized takers. Here again we need both sums to have the same sign to indicate nonrandom selection on unobserved variables into the RCT.

Applying Huber (2013) to this context, testing for the significance of $\widehat{\delta_4}$ and $\widehat{\delta_4} + \widehat{\delta_5}$ indicates whether noncompliance within the RCT is ignorable. Due to the exogeneity of assignment, having either be significant is sufficient to reject ignorable noncompliance. However, were both to be small and statistically insignificant, we may be comfortable generalizing the LATE to the remainder of the RCT population, even if we cannot generalize beyond it.

Empirical Results

Analysis 1: The ITT and the LATE estimates of program effect from the RCT design appear in Panels A and B, respectively, of Table 3. The ITT estimates are not altered in any meaningful way by the introduction of controls, and are the same whether estimated by OLS or logit QML. The quantitative magnitude of the ITT – a roughly two percentage point increase in the retention probability – is not insubstantial, but the estimates have large standard errors and none are close to being statistically different from zero at any conventional threshold. Some may worry about the lack of power due to a binary outcome in this case. As a result, we also perform a similar analysis with GPA as the outcome variable, which possesses increased power. We find no statistically significant effect of the FYLC on grades as well.[14]

Panel B gives the LATE estimates, while Panel C provides the first-stage estimates, which reveal that the randomization provides a strong instrumental variable in explaining

[14] With regard to the analysis of 1st year GPA, at a power of 0.9 our desired sample size would allow us to detect an effect of 0.07 grade points. Our data contain 2nd year cumulative GPA for just the first cohort. For analysis on 2nd year GPA at a power of 0.9 our desired sample size would allow us to detect an effect of 0.10 grade points. We include the RCT results of the FYLC program on GPA in Table A2 in the Appendix A.

variation in FYLC participation. The estimated local average treatment effect of the program is roughly one percentage point larger than the intent to treat estimates, but once again these estimates are imprecisely estimated and thus statistically insignificantly different from zero.

The control function residuals in column 3 of Panel B preview results from a portion of the analysis presented below in Analysis 3. The coefficient estimate is small and far from statistically significant. Thus, we fail to reject the null hypothesis of ignorable noncompliance. This provides the first piece of reassurance that the compliers do not appear to be systematically selected among those who opt into the experiment.

Table 3: RCT estimates

	(1)	(2)	(3)
	Retention	Retention	Retention
Panel A: ITT effects of winning lottery on first-year retention (reduced form estimates)			
Assigned FYLC	0.019	0.018	0.018
	(0.015)	(0.015)	(0.014)
Panel B: Estimated LATEs of FYLC on 1st year retention (2nd Stage estimates)			
FYLC	0.029	0.027	0.027
	(0.022)	(0.022)	(0.022)
Residuals			-0.004
			(0.029)
Panel C: Effect of lottery assignment of treatment status (1st stage estimates)			
Assigned FYLC	0.648^{***}	0.648^{***}	0.648^{***}
	(0.019)	(0.019)	(0.019)
Observations	1565	1565	1565
Retention Mean	0.910	0.910	0.910
Controls	No	Yes	Yes
Model	LPM	LPM	QML

The first two column report results from linear models whereas column (3) reports estimates from nonlinear estimation. Logit was used in QML estimation. The control function residuals used with QML in panel B were estimated using OLS. Column (1) is an unconditional estimate whereas columns (2) and (3) include baseline covariates. Robust standard errors in parentheses. Bootstrap standard errors with 500 replications were used for inference in QML control function estimation. *** p<0.01, **p<0.05, * p<0.1.

Analysis 2: If the evaluation of program impact had not relied on random assignment, but rather had utilized an observational research design, how would the estimated program impact have differed? We present the results from observational approaches to estimate the program impact where the treated, including crossovers, nonrandomized takers in addition to the compliant treated, are compared to non-participants that include the compliant control, no-shows, and never-ever takers in Table 4.

Contrary to the findings from the RCT design, the Table 4 results reveal an estimated coefficient on the treatment variable in the observational analysis that is positive and statistically significant regardless of specification or procedure invoked. Moreover, the estimated quantitative impact is large – ranging from a 2.7 to 5.2 percentage point gain in retention probability by virtue of participation in the FYLC. Furthermore, in Panel B we restrict attention to the observations in which the overlap is thick (with propensity scores ranging from 0.1 to 0.9). Here the estimated effects are even larger with coefficient estimates all over 5 percentage points with similar p-values ranging from 0.03 to less than 0.001.

Table 4: Observational analysis estimates of program effects.

	(1)	(2)	(3)	(4)	(5)
Panel A: Full sample					
FYLC	0.038***	0.049***	0.052***	0.044**	0.027*
	(0.010)	(0.011)	(0.013)	(0.020)	(0.016)
Observations	8131	8131	8131	8131	8131
Mean	0.91	0.91	0.91	0.91	0.91
Controls	No	Yes	Yes	Yes	Yes
Estimation	OLS	OLS	Logit	PSM ATT	PSM ATE
Panel B: Sample restricted on propensity score					
FYLC	0.050***	0.052***	0.058***	0 .050***	0.054***
	(0.013)	(0.013)	(0.017)	(0.023)	(0.015)
Observations	3816	3816	3816	3816	3816
Mean	0.88	0.88	0.88	0.88	0.88
Controls	No	Yes	Yes	Yes	Yes
Estimation	OLS	OLS	Logit	PSM ATT	PSM ATE

Robust standard errors in parentheses. Bootstrap standard errors with 500 replications were used for inference on propensity score matched estimates of the treatment on the treated. The restricted sample uses only observation for which there is overlap with propensity scores greater than 0.1 and less than 0.9. For PSM we present the estimated average treatment on the treated as well as estimates of the ATE. *** p<0.01, ** p<0.05, * p<0.1.

Based on the observed differences among the treated and control populations and the way in which retention probabilities are negatively correlated with those differences, analysts employing such observational analyses might be tempted to hypothesize that the observational results are *underestimates* of true program impact. Indeed, such reasoning underlies the bounds of ATEs proposed in Andrews and Oster (2018). However, as we restrict the sample to that for which there is more overlap on observed covariates, the observational estimates universally grow away from the estimated LATE. We explain the basis for this finding in an exploration of the differences between the observational and RCT results below.

Analysis 3: What accounts for the differences in the observational and RCT results? And to what extent are the RCT findings generalizable to other populations of interest? Table 5 presents the results from estimating equation (2) – in columns (1) and (2) – and equation (3) – in columns (3) and (4). Below these results in the table, we give the relevant population comparisons of interest to our analysis and discussed in the methodology section above. Looking at the population comparisons for the column (2) results, we see that among those who receive no treatment, those who participate in the experiment are 3.8 percentage points more likely to persist in college (p-value of 0.001). Here, neither comparison group was treated; the difference is in selection. Next, we examine the coefficients on RCT and RCTxFYLC to compare retention between the nonrandomized takers and the treated participants in the experiment. We find that those who entered the RCT are 6.4 percentage points more likely to persist (p-value of 0.054). This difference is despite the fact that both received treatment.

Turning to the results of the fully interactive model presented in columns (3) and (4), we see, in rows three through six of the population comparisons, that accounting for noncompliance within the experiment does little to change our results above; we continue to observe substantial positive selection into the experimental sample among both the treated and untreated. Because those who enter the RCT exhibit higher persistence than those who do not, among both treated and untreated individuals, the evidence indicates positive selection on unobserved characteristics into the RCT, which drives the differences between the observational and RCT results.[15]

Table 5: Full-sample decomposition testing selection into and within the RCT

	(1)	(2)	(3)	(4)
	Retention	Retention	Retention	Retention
RCT (R)	0.028**	0.038***	0.026**	0.038***
	(0.011)	(0.011)	(0.013)	(0.013)
FYLC (D)	-0.016	-0.002	-0.016	-0.002
	(0.033)	(0.032)	(0.033)	(0.032)
RCT x FYLC	0.038	0.026	0.036	0.026
	(0.036)	(0.035)	(0.044)	(0.044)
RCT x Assigned FYLC (Z)			0.009	0.002
			(0.025)	(0.025)
RCT x Assigned FYLC x FYLC			-0.003	-0.001
			(0.038)	(0.038)
		Population Comparisons		
$\widehat{\pi_1}$ comparing NET vs (CC+NS)	0.028**	0.038***		
	[0.014]	[0.001]		
$\widehat{\pi_1} + \widehat{\pi_3}$ comparing NRT vs (CT+CO)	.066**	.064*		
	[0.050]	[0.054]		
$\widehat{\delta_1}$ comparing NET vs CC			.026**	.038***
			[0.039]	[0.003]
$\widehat{\delta_1} + \widehat{\delta_4}$ comparing NET vs NS			0.035	0.039*
			[0.128]	[0.079]
$\widehat{\delta_1} + \widehat{\delta_3}$ comparing NRT vs CO			0.062	0.064
			[.141]	[0.128]
$\widehat{\delta_1} + \widehat{\delta_3} + \widehat{\delta_4} + \widehat{\delta_5}$			0.067**	0.065*
comparing NRT vs CT			[0.049]	[0.055]
$\widehat{\delta_4} + \widehat{\delta_5}$ comparing CO vs CT			0.005	0.001
			[0.852]	[0.969]
N	8131	8131	8131	8131
Controls	No	Yes	No	Yes

All results are from OLS regressions and present the decomposition of the results from Table 5. Robust standard errors in parentheses. P-values of the linear combination above are in brackets. NET=never-ever takers; CC=compliant controls; NS=no-shows; NRT=nonrandomized takers; CT=compliant treated; and CO=crossovers. Columns (2) and (4) add controls. Columns (3) and (4) add indicators assignment to the FYLC and the same indicator interacted with participating in it. The omitted category for these regressions is composed of those never-ever takers who do not enter the RCT and do not enter the FYLC.

[15] Following the conservative inference described in Appendix C, we reject the joint hypothesis of no selection or opposed signs with p-values of 0.050 or 0.054 depending on whether we adjust for covariates. Confidence intervals are even tighter using randomization inference as shown in Appendix D.

*** $p<0.01$, **. $p<0.05$, * $p<0.1$.

Combined with the negative selection into the experiment based on observed characteristics, reported in the descriptive statistics section above and in appendix Table A1, this analysis leads to a rather striking conclusion: While students who select into the experiment possess observed correlates – such as lower SAT scores, lower high-school GPAs, and be substantially more likely to be a first-generation college student and from a low-income family -- which render them vulnerable with regard to first-year retention, their observed vulnerabilities appear to be combined with unobserved characteristics – such as grit, determination, and focus – that more than make up for these observed weaknesses.

Finally, regarding the issue of external validity of the RCT, the large positive selection on unobserved variables into the experiment suggests that the RCT results lack external validity for the non-experimental population. However, we find little difference between those who comply with their assignments and those who do not.[16] The coefficient on FYLC shows the difference between the no-shows and the assignment-compliant untreated is 0.002 (with a p-value of 0.959). Among the treated the last row of Table 5 shows that the difference between the crossovers and compliant treated participants is 0.001 (with a p-value of 0.969). These results reveal no discernable selection within the experiment itself, suggesting that the ATE for the experimental sample is unlikely to differ from the LATE and that the RCT findings are likely to extend to the noncompliant crossovers and no-shows.

Conclusions

We utilize an RCT design to estimate the impact of a learning community on first-year college retention for those students who self-select into the program. The results are the first of their kind to employ an RCT to address this question at a four-year college or university. We find that both the “intent to treat” and the “local average treatment effect” estimates of program

[16] We further directly test for ignorable noncompliance described in Huber (2013) in Table A4 in the Appendix A. We find little difference between those who comply with their assignments and those who do not both among the treated and untreated, suggesting such noncompliance is ignorable.

impact are small and statistically insignificantly different from zero. The first-year learning community program at this institution had no measureable causal effect on student retention into the second year of college for the treated population.

Next, we turn to an analysis of program impact using observational approaches, which are still employed widely by in-house institutional researchers. The results reveal an estimated impact on retention that is positive, large and statistically significant regardless of specification or procedure (OLS or propensity-score matching). Further analysis reveals that the driving force behind the difference in the observational and experimental estimates of program impact is the statistically meaningful difference in retention prospects between those who do and do not enter the lottery. Interestingly, we find that, comparatively, lottery participants (whether or not they receive treatment) possess observed characteristics – including lower high-school GPAs and lower SAT scores – that render them less likely to return for sophomore year, but unobserved characteristics – presumably, things like grit, determination and focus – that statistically and quantitatively more than make up for these observed vulnerabilities in terms of retention prospects.

Finally, we explore issues related to external validity of the RCT results, focusing on different possible populations of interest, and using statistically significant differences in unobserved propensities to persist as our criterion for external invalidity. While the RCT findings may serve as a causal and unbiased estimate of program impact for the students who self-selected into the study, nonrandom selection into the program cautions against generalizing these results to the population who did not enroll in the experiment. We also test for generalizability of the RCT findings to other populations within the experiment who failed to comply with the random assignment – i.e., the group that migrated from treatment to control and the group that did the reverse. We find little difference in unobserved propensities to persist for these two groups, and so it seems reasonable to generalize the "local average treatment effect" estimate of program impact to the remaining experimental population.

References

Allcott, Hunt. "Site selection bias in program evaluation." The Quarterly Journal of Economics 130, no. 3 (2015): 1117-1165.

Altindag, Onur, Theodore J. Joyce, and Julie A. Reeder, 2015. *Effects of Peer Counseling to Support Breastfeeding: Assessing the External Validity of a Randomized Field Experiment*. No. w21013. National Bureau of Economic Research.

Andrews, Isaiah, and Emily Oster. 2017. *Weighting for External Validity*. No. w23826. National Bureau of Economic Research.

Angrist, Joshua D. "Treatment effect heterogeneity in theory and practice." The economic journal 114, no. 494 (2004): C52-C83.

Angrist, Joshua, Daniel Lang, and Philip Oreopoulos. 2009. "Incentives and services for college achievement: Evidence from a randomized trial." *American Economic Journal: Applied Economics* 1, no. 1: 136-63.

Barefoot, Betsy O., Betsy Q. Griffin, and Andrew K. Koch. 2012. "Enhancing student success and retention throughout undergraduate education: A national survey." *Gardner Institute for Excellence in Undergraduate Education*.

Barefoot, Betsy O., Carrie L. Warnock, Michael P. Dickinson, Sharon E. Richardson, and Melissa R. Roberts. 1998. *Exploring the Evidence: Reporting Outcomes of First-Year Seminars. The First-Year Experience. Volume II. Monograph Series, Number 25*. National Resource Center for the First-Year Experience and Students in Transition, 1629 Pendleton St., Columbia, SC 29208.

Bartlett, C., L. Doyal, S. Ebrahim, P. Davey, M. Bachmann, M. Egger, and P. Dieppe. 2005. "The causes and effects of socio-demographic exclusions from clinical trials." *Health Technology Assessment (Winchester, England)* 9, no. 38: iii-iv.

Berger, Roger L. "Multiparameter hypothesis testing and acceptance sampling." Technometrics 24, no. 4 (1982): 295-300.

Bertanha, Marinho, and Guido W. Imbens. 2019. "External validity in fuzzy regression discontinuity designs." *Journal of Business & Economic Statistics*: 1-39.

Bettinger, Eric P., and Rachel B. Baker. 2014. "The effects of student coaching: An evaluation of a randomized experiment in student advising." *Educational Evaluation and Policy Analysis* 36, no. 1: 3-19.

Black, Dan, Joonhwi Joo, Robert LaLonde, Jeffrey A. Smith, and Evan Taylor. 2017. "Simple tests for selection: Learning more from instrumental variables." *CES IFO, Working paper No 6392.*

Bo, Hao, and Sebastian Galiani. "Assessing external validity." *Research in Economics* (2021).

Brinch, Christian N., Magne Mogstad, and Matthew Wiswall. 2017. "Beyond LATE with a discrete instrument." *Journal of Political Economy* 125, no. 4: 985-1039.

Clark, M. H., and Cundiff, Nicole L. 2011. "Assessing the effectiveness of a college freshman seminar using propensity score adjustments." *Research in Higher Education* 52, no. 6 (2011): 616-639.

Cole, Stephen R., and Elizabeth A. Stuart. 2010. "Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial." *American Journal of Epidemiology* 172, no. 1: 107-115.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2009. "Dealing with limited overlap in estimation of average treatment effects." *Biometrika* 96, no. 1: 187-199.

Galiani, Sebastian, Patrick J. McEwan, and Brian Quistorff. "External and internal validity of a geographic quasi-experiment embedded in a cluster-randomized experiment." In *Regression discontinuity designs: Theory and applications*, pp. 195-236. Emerald Publishing Limited, 2017.

Gechter, Michael. "Generalizing the Results from Social Experiments: Theory and Evidence." *Working Paper* (2016).

Ghanem, Dalia, Hirshleifer, Sarojini, and Ortiz-Becerra, Karen. 2018. Testing Attrition Bias in Field Experiments. Unpublished manuscript, University of California, Riverside.

Hansen, M.J. and Schmidt, L., 2017. The synergy of and readiness for high-impact practices during the first year of college. Journal of the First-Year Experience & Students in Transition, 29(1), pp.57-82.

Hartman, Erin, Richard Grieve, Roland Ramsahai, and Jasjeet S. Sekhon. "From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178, no. 3 (2015): 757-778.

Hausman, Jerry A., and David A. Wise. 1979. "Attrition bias in experimental and panel data: the Gary income maintenance experiment." *Econometrica*: 455-473.

Hotchkiss, Julie L., Robert E. Moore, and M. Melinda Pitts. 2006. "Freshman learning communities, college performance, and retention." *Education Economics* 14, no. 2: 197-210.

Huber, Martin. 2013. "A simple test for the ignorability of non-compliance in experiments." *Economics Letters* 120, no. 3: 389-391.

Janzing, Dominik, Peters, Jonas, and Schölkopf, Bernhard, 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT press.

Karlan, Dean, and Jonathan Zinman. 2009. "Observing unobservables: Identifying information asymmetries with a consumer credit field experiment." *Econometrica* 77, no. 6: 1993-2008.

Kowalski, Amanda. 2016. *Doing more when you're running late: Applying marginal treatment effect methods to examine treatment effect heterogeneity in experiments.* Working Paper 22362, National Bureau of Economic Research, URL http://www.nber.org/papers/w22362.

Kowalski, Amanda E. 2018. *How to Examine External Validity Within an Experiment*. No. w24834. National Bureau of Economic Research.

Kuh, George D., Jillian Kinzie, John H. Schuh, and Elizabeth J. Whitt. 2011. *Student success in college: Creating conditions that matter*. John Wiley & Sons.

Kyndt, E., Donche, V., Trigwell, K. and Lindblom-Ylänne, S., 2017. Why theory, research and practice matter. Higher Education Transitions: Theory and Research, p.306.

Laufgraben, J. L., Shapiro, N. S., & Associates. 2004. "The what and why of learning communities." In J. L. Laufgraben, N. S. Shapiro, & A. (Eds.), *Sustaining and Improving Learning Communities*:1-13. San Francisco: Jossey-Bass.

Lin, Winston. 2013. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique." *The Annals of Applied Statistics* 7, no. 1: 295-318.

Lise, Jeremy, Shannon Seitz, and Jeffrey Smith. 2015. "Evaluating search and matching models using experimental data." *IZA Journal of Labor Economics* 4, no. 1: 16.

Millea, M., Wills, R., Elder, A. and Molina, D., 2018. What matters in college student success? Determinants of college retention and graduation rates. Education, 138(4), pp.309-322.

Paloyo, Alfredo R., Sally Rogan, and Peter Siminski. 2016. "The effect of supplemental instruction on academic performance: An encouragement design experiment." *Economics of Education Review* 55: 57-69.

Pascarella, Ernest T., and Patrick T. Terenzini. 2005. "How college affects students: A third decade of research." 571-626.

Pike, Gary R., Michele J. Hansen, and Ching-Hui Lin. 2011. "Using instrumental variables to account for selection effects in research on first-year programs." *Research in Higher Education* 52, no. 2: 194-214.

Pitkethly, Anne, and Michael Prosser. 2001. "The first year experience project: A model for university-wide change." *Higher Education Research & Development* 20, no. 2: 185-198.

Prentice, Ross L. , Robert Langer, Marcia L. Stefanick, Barbara V. Howard, Mary Pettinger, Garnet Anderson, David Barad, J. David Curb, Jane Kotchen, Lewis Kuller, Marian Limacher, Jean Wactawski-Wende, (2005). "Women's Health Initiative Investigators, Combined Postmenopausal Hormone Therapy and Cardiovascular Disease: Toward Resolving the Discrepancy between Observational Studies and the Women's Health Initiative Clinical

Trial," *American Journal of Epidemiology*, Volume 162, Issue 5, 1 September, Pages 404–414, https://doi.org/10.1093/aje/kwi223

Provencher, A. and Kassel, R., 2019. High-impact practices and sophomore retention: Examining the effects of selection bias. Journal of College Student Retention: Research, Theory & Practice, 21(2), pp.221-241.

Romano, Joseph P., and Michael Wolf. "Stepwise multiple testing as formalized data snooping." Econometrica 73, no. 4 (2005): 1237-1282.

Rubin, Donald B. 1980. "Comment." *Journal of the American Statistical Association* 75, no. 371: 591-593.

Russell, Lauren. 2017. "Can learning communities boost success of women and minorities in STEM? Evidence from the Massachusetts Institute of Technology." *Economics of Education Review* 61: 98-111.

Scrivener, S., D. Bloom, A. LeBlanc, C. Paxson, and C. Sommo. 2008. "A good start: Two-year effects of a freshmen learning community program at Kingsborough Community College. New York, NY: MDRC."

Schochet, Peter Z. 2010. "Is regression adjustment supported by the Neyman model for causal inference?" *Journal of Statistical Planning and Inference* 140, no. 1: 246-259.

Sianesi, Barbara. 2017. "Evidence of randomisation bias in a large-scale social experiment: The case of ERA." *Journal of Econometrics* 198, no. 1: 41-64.

Tian, Jin, and Judea Pearl. 2000. "Probabilities of causation: Bounds and identification." *Annals of Mathematics and Artificial Intelligence* 28, no. 1-4: 287-313.

Tsiatis, Anastasios A., Marie Davidian, Min Zhang, and Xiaomin Lu. 2008. "Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach." *Statistics in medicine* 27, no. 23: 4658-4677.

U.S. Department of Education. 2017. "*Digest of Education Statistics 2017.*" National Center for Education Statistics. Washington, D.C.

U.S. News and World Report. 2018. https://www.usnews.com/best-colleges/rankings/national-universities/freshmen-least-most-likely-return.

Visher, Mary G., Michael J. Weiss, Evan Weissman, Timothy Rudd, and Heather D. 2012. Wathington. "The Effects of Learning Communities for Students in Developmental Education: A Synthesis of Findings from Six Community Colleges." National Center for Postsecondary Research.

Vytlacil, Edward J., 2002. "Independence, Monotonicity, and Latent Index Models: An Equivalence Result." *Econometrica*, 70(1) 331–41.

Walters, Christopher R. 2018. "The demand for effective charter schools." *Journal of Political*

Economy 126(6) 2179-2223.

Wooldridge, Jeffrey M. 2014. "Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables." Journal of Econometrics 182, no. 1: 226-234.

Xerri, M.J., Radford, K. and Shacklock, K., 2018. Student engagement in academic activities: A social support perspective. Higher education, 75(4), pp.589-605.

Yang, Li, and Anastasios A. Tsiatis. 2001. "Efficiency study of estimators for a treatment effect in a pretest–posttest trial." The American Statistician 55, no. 4: 314-321.

Young, Alwyn. "Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results." The Quarterly Journal of Economics 134, no. 2 (2019): 557-598.

Zhao, Chun-Mei, and George D. Kuh. 2004. "Adding value: Learning communities and student engagement." Research in higher education 45, no. 2: 115-138.

Appendix for online publication:

Appendix A: Additional figures and tables

Figure A1: Overlap in the propensity scores by treatment status

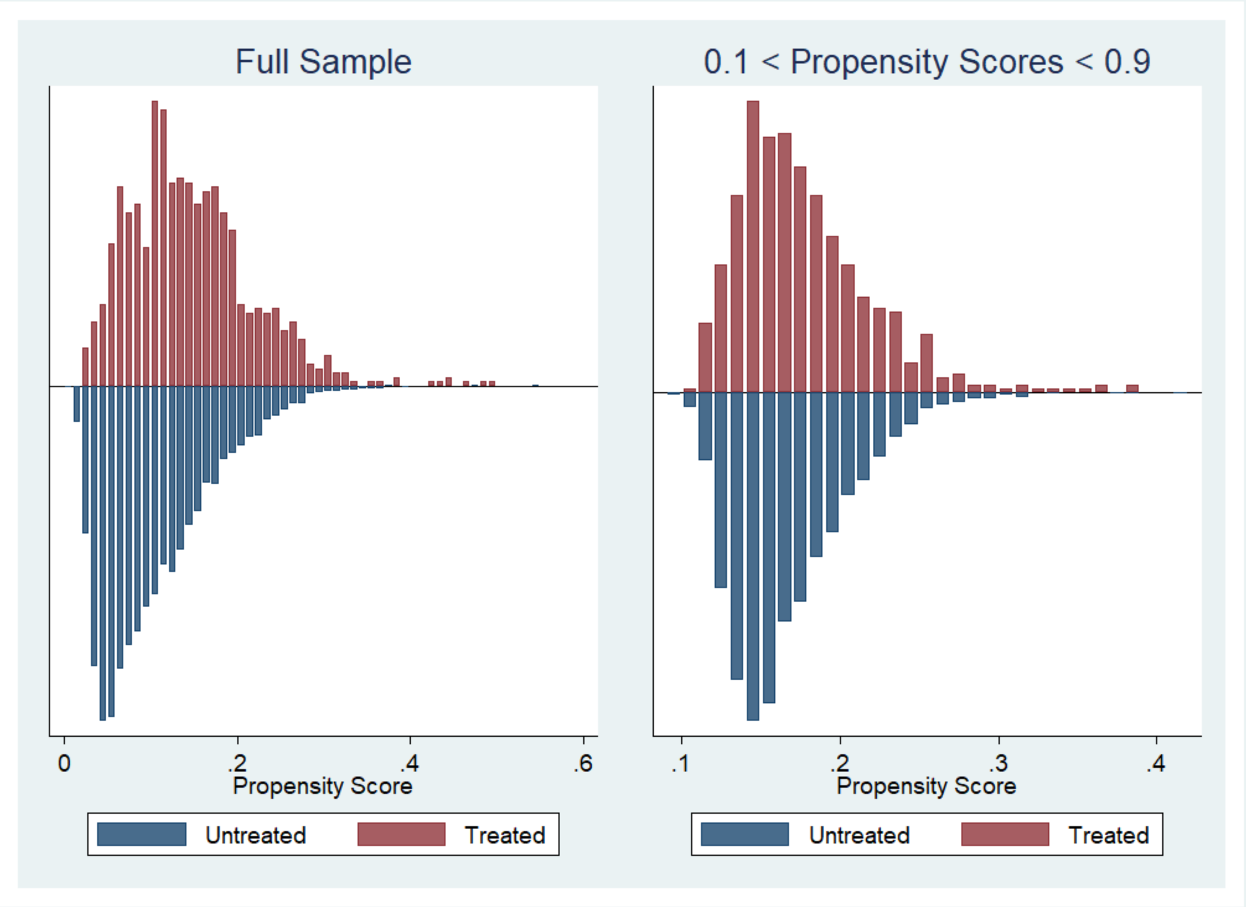


Figure notes: Propensity scores estimated using logit.

Table A1: Observed predictors of 1st year college retention

	(1) Retention	(2) Retention	(3) Retention
High-school GPA	0.06***	0.07***	0.07***
	(0.010)	(0.011)	(0.012)
SAT Math	0.01*	0.01**	0.01**
	(0.005)	(0.005)	(0.006)
SAT Writing	0.00	0.00	0.00
	(0.007)	(0.007)	(0.007)
SAT Verbal	0.01	0.01	0.01
	(0.006)	(0.007)	(0.007)
On Campus	0.04***	0.04***	0.04***
	(0.009)	(0.009)	(0.010)
Female	0.01	0.01	0.01
	(0.008)	(0.008)	(0.009)
First Generation	-0.02**	-0.02**	-0.02**
	(0.008)	(0.009)	(0.009)
Low Income	-0.00	-0.01	-0.00
	(0.008)	(0.009)	(0.009)
Cohort	0.00	0.00	0.00
	(0.007)	(0.008)	(0.008)
N	8131	7252	6449

SAT scores are divided by 100 for presentation. Robust standard errors are in parentheses. All regressions use OLS and also include cohort indicators and indicators for missing covariates. Column (1) is estimated using the whole sample. Column (2) is estimated with only the untreated. Column (3) is estimated using those who do not enter the RCT and do not receive treatment. *** p<0.01, ** p<0.05, * p<0.1.

Table A2: Observable differences between treated and untreated in different populations

	(1)	(2)	(3)
	FYLC	FYLC	FYLC
High-school	0.010	-0.007	-0.007**
GPA	(0.041)	(0.036)	(0.003)
SAT math	-0.059***	-0.053***	-0.010***
	(0.021)	(0.017)	(0.002)
SAT writing	-0.023	-0.022	0.001
	(0.028)	(0.025)	(0.003)
SAT verbal	0.066**	0.050**	0.006**
	(0.027)	(0.023)	(0.003)
Female	0.053	0.051*	0.013***
	(0.033)	(0.027)	(0.003)
1st generation	0.013	-0.001	0.009**
	(0.035)	(0.033)	(0.004)
Low income	0.072**	0.014	-0.006*
	(0.035)	(0.033)	(0.004)
Lives on	0.017	0.059**	0.003
campus	(0.034)	(0.028)	(0.004)
N	824	741	6572

SAT scores are divided by 100 for presentation. Robust standard errors are in parentheses. All regressions use OLS and also include cohort indicators and indicators for missing covariates. Column (1) is estimated with those assigned to treatment. Column (2) is estimated with those assigned to the control group. Column (3) is estimated using those who do not enter the RCT. *** p<0.01, ** p<0.05, * p<0.1.

Table A3: RCT estimates of the effects on GPA

Panel A: Intent to treat effects of winning lottery on first and second year GPA (reduced form estimates)

	(1) 1st Year GPA	(2) 1st Year GPA	(3) 2nd Year GPA	(4) 2nd Year GPA
Won lottery	0.016	0.018	0.015	-0.016
	(0.030)	(0.027)	(0.038)	(0.036)

Panel B: Estimated LATEs of FYLC on 1st and 2nd year GPA (2nd Stage estimates)

FYLC	0.024	0.027	0.022	-0.018
	(0.045)	(0.042)	(0.053)	(0.053)

Panel C: OLS 1st stage estimates of the effect of winning the lottery on FYLC participation

Won lottery	0.649***	0.649***	0.706***	0.709***
	(0.020)	(0.019)	(0.028)	(0.027)
Observations	1489	1489	662	662
GPA Mean	2.812	2.812	2.901	2.901
Controls	No	Yes	No	Yes

All estimates are from linear regressions. Columns (1) and (3) are unconditional estimates whereas columns (2) and (4) include baseline covariates. 1st GPA includes FYLC course grade. 2nd year GPA only exists in our data for the earlier cohort. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A4: Testing for selection of compliers within the RCT

	(1)	(2)
Won lottery	0.009	0.003
	(0.025)	(0.025)
FYLC	0.019	0.025
	(0.029)	(0.030)
Won lottery x FYLC	-0.003	-0.002
	(0.038)	(0.038)
Observations	1565	1565
Controls	No	Yes
Sample	Lottery	Lottery

All results are from OLS regressions. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Column (1) presents the results from a simple regression of retention on indicators for winning the lottery, entering the FYLC after entering the lottery, and winning the lottery and entering the FLYC. In column (2), we add controls.

Appendix B: Selection into the RCT proposition and proof

Let Y_{0i} and Y_{1i} be outcomes in states of the world where individual i does not and does receive treatment. Consider the following unconditional regression in under perfect compliance with the experiment:

$$Y_i = \alpha + R_i\pi_1 + D_i\pi_2 + D_i \times R_i\pi_3 + \varepsilon_i, \tag{4}$$

such that $\pi_1 = E(Y_0|R = 1, D = 0) - E(Y_0|R = 0, D = 0)$, $\pi_2 = E(Y_1|R = 0, D = 1) - E(Y_0|R =, D = 0)$, $\pi_3 = E(Y_1|R = 1, D = 1) - E(Y_0|R = 0, D = 0) - [E(Y_0|R =, D = 0) - E(Y_0|R = 0, D = 0)] - E(Y_1|R = 0, D = 1) - E(Y_0|R = 0, D = 0)$, and $\pi_3 + \pi_1$= $E(Y_1|R = 1, D = 1) - E(Y_1|R = 0, D = 1)$.

Proposition: $\pi_1, \pi_3 + \pi_1 > 0$ or $\pi_1, \pi_3 + \pi_1 < 0$ can only be justified through heterogeneous effects, selection on unobserved variables into the experiment or both under (A1) proper randomization and (A2) no independent effect of assignment within the experiment, and (A3) the stable unit treatment value assumption.

Proof by contradiction: Suppose also (A4) no selection on unobserved variables into the experiment, $[E(Y_0|R = 1) = E(Y_0|R = 0)$ and $E(Y_1|R = 1) = E(Y_1|R = 0)]$, and (A5) homogenous effects of treatment, $[E(Y_1 - Y_0) = E(Y_1 - Y_0|R = j, D = k)$ for $j = 0, 1$ and $k = 0, 1]$. Without loss of generality consider the case where $\pi_1, \pi_3 + \pi_1 > 0$. Under assumptions (A1) – (A4), we rewrite the inequalities regarding π_1 and $\pi_3 + \pi_1$ as the following:

$$\pi_1 = (1 - P)^{-1}\{P[E(Y_0|R = 0, D = 1) - E(Y_0|R = 1)]\} > 0, \tag{B1}$$

$$\pi_1 + \pi_3 = P^{-1}\{(1 - P)[E(Y_1|L = 0, D = 0) - E(Y_1|R = 1)]\}>0. \tag{B2}$$

Further under no selection on unobserved variables into the experiment (A4), $E(Y_0|R = 0, D = 1) > E(Y_0|L = 1) > E(Y_0|R = 0, D = 0)$. Under homogenous treatment effects (A5), $E(Y_1|R = 0, D = 1) > E(Y_1|R = 1) > E(Y_1|R = 0, D = 0)$, which contradicts equation (B2).

Appendix C: Joint hypothesis inference:

We wish to perform inference over the hypothesis that $\pi_1, \pi_3 + \pi_1 > 0$ or $\pi_1, \pi_3 + \pi_1 > 0$. We follow Brinch et al. (2017) in evaluating the hypothesis that both coefficients are the same sign with three steps. First, we test whether both treated and untreated lottery participants have higher outcomes than the treated and untreated non-experimental populations respectively. We may conduct this intersection test by performing two separate one-sided t-tests in which we control for the familywise error rate associated with having multiple hypotheses. We again follow Brinch et al. (2017) in using the conservative Bonferroni correction in which the corrected p-value on the hypothesis that both inequalities hold versus the alternative that at least one inequality does not hold is two times the largest of the p-values from the two one-sided t-tests. Bonferroni tests are known to be under-powered and researchers may gain power by instead using a step-down procedure as in Romano and Wolf (2005). Secondly, we repeat the procedure for the test of the joint hypothesis that both treated and untreated experimental participants experience lower outcomes than their respective non-experimental comparisons. Finally, we test whether ($\pi_1, \pi_3 + \pi_1 > 0$) is included in the union of two subsets. As each of the two subsets is formed by the intersection of two one-sided tests, we may apply the result from Berger (1982), which shows that the p-value for the test that selection into the experiment among the treated and untreated is either both positive or both negative is the lower of the two corrected p-values from the first two steps. Because p-values for one-sided tests are half the p-values for the two-sided tests, in this application, the Bonferroni correction and application of Berger (1982) leaves us with the largest p-value of the original two-sided t-test in cases where both coefficients are the same sign.

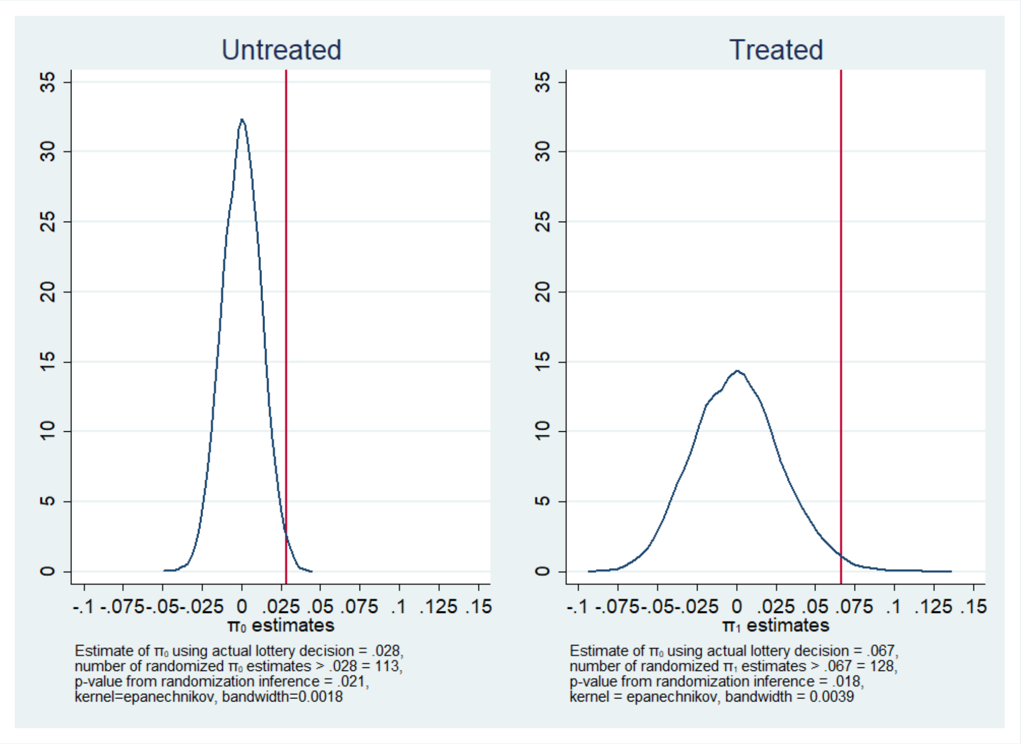
Appendix D: Randomization Inference

As a robustness exercise, we couple the decomposition presented in the main text with an analogous nonparametric randomization inference exercise. Here, we test the sharp hypothesis that there are no differences in outcomes between those who enter and those who do not enter the experiment within each treatment status. We do this using two different approaches following Young (2019). In each of 10,000 repetitions, we randomly assign each individual within the treated and non-treated populations to the "lottery" according to the binomial distribution, keeping the shares of the treated and untreated populations who enter the lottery constant at 87 percent and 11 percent respectively. In the first approach, we find the average differences ($\widehat{\pi_{0p}}$ being the average difference in retention by lottery participation for those who do not receive treatment and $\widehat{\pi_{1p}}$ serving as the same for the treated) between the placebo lottery assigned groups. We then compare the differences in retention observed under the actual lottery participation decisions to the distribution of placebo differences we observe under random assignment of "lottery participation." The share of placebo differences whose magnitudes are more extreme than the magnitude of the difference using actual lottery assignment may sensibly be interpreted as the p-values of the differences using actual lottery participation. Secondly, we do the same using the t-statistics on the difference rather than the difference itself. These approaches avoid possible finite sample bias and apply minimal assumptions or structure to the data, while providing valid and transparent inference.

Figure D1 presents the distribution of estimated differences in retention among the untreated (on the left) and among the treated (on the right) when "lottery participation" is randomly assigned in each of 10,000 repetitions. We show the estimated difference in retention using the actual lottery participation using a red vertical line. Following Young (2018), we repeat the exercise using the t-statistics presented in Figure D2.

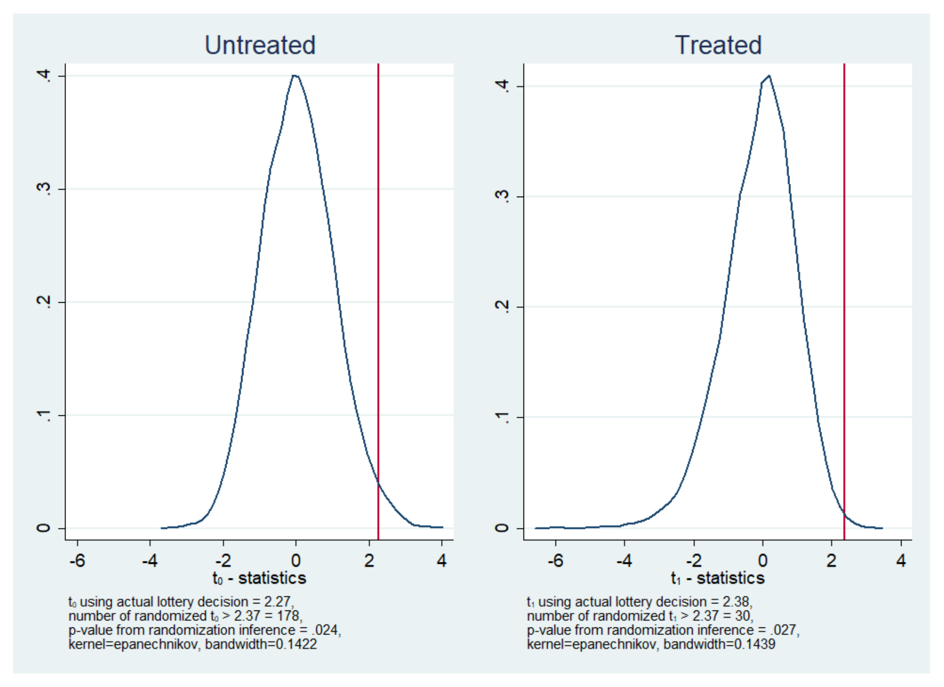
In each case the red line lies on the far-right side, indicating that the realized differences in retention between those who actually do and do not participate in the experiment are unlikely to result from pure chance. As described above, the p-values from our randomization tests correspond closely to the share of squared differences (or F-statistics) from the placebo assignment that are larger than the squared differences (or F-statistics) that arrived at using the actual realization of lottery assignment. Among the untreated, the corresponding p-values using squared raw differences and F-statistics are 0.021 and 0.024. Among the treated, the corresponding p-values are 0.018 and 0.027. In both cases we find strong positive selection into the experimental sample and reject the null of no selection. The distribution of squared differences and F-statistics are shown in Figures D3 and D4 and Table D1 provides the nonparametric unconditional differences in retention rates between the experimental and non-experimental populations stratified by treatment status with the accompanying p-values as well as the mean and the first, fifth, tenth , fiftieth, nintieth, ninty-fifth, and ninty-ninth percentile of the placebo differences when lottery participation is randomly assigned. In each case, randomization inference provides similar or smaller p-values than those constructed from the Huber-White robust standard errors.

Figure D1: Distribution of placebo $\hat{\pi}$ where "lottery participation" is randomly assigned



Notes: Binomial random assignment to lottery participation with probabilities of inclusion in the lottery by treatment status set at 0.11 for the untreated and 0.87 for the treated reflecting the shares observed in the data. Distributions constructed from 10,000 repetitions. The red vertical lines denote the differences in the mean retention between experimental and non-experimental populations within treatment status.

Figure D2: Distribution of placebo t-statistics where “lottery participation” is randomly assigned

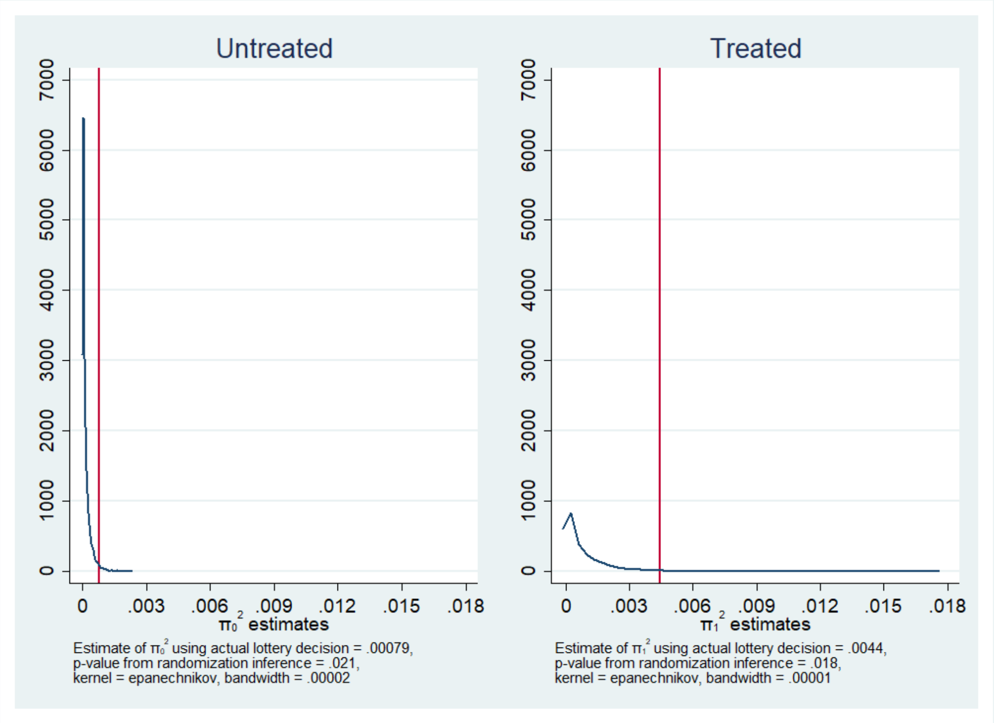


t_0 using actual lottery decision = 2.27,
number of randomized t_0 > 2.37 = 178,
p-value from randomization inference = .024,
kernel=epanechnikov, bandwidth=0.1422

t_1 using actual lottery decision = 2.38,
number of randomized t_1 > 2.37 = 30,
p-value from randomization inference = .027,
kernel=epanechnikov, bandwidth=0.1439

Notes: Binomial random assignment to lottery participation with probabilities of inclusion in the lottery by treatment status set at 0.11 for the untreated and 0.87 for the treated reflecting the shares observed in the data. Distributions constructed from 10,000 repetitions. The red verticle lines denote the differences in the mean retention between experimental and non-experimental populations within treatment status.

Figure D3: Distribution of squared placebo coefficients

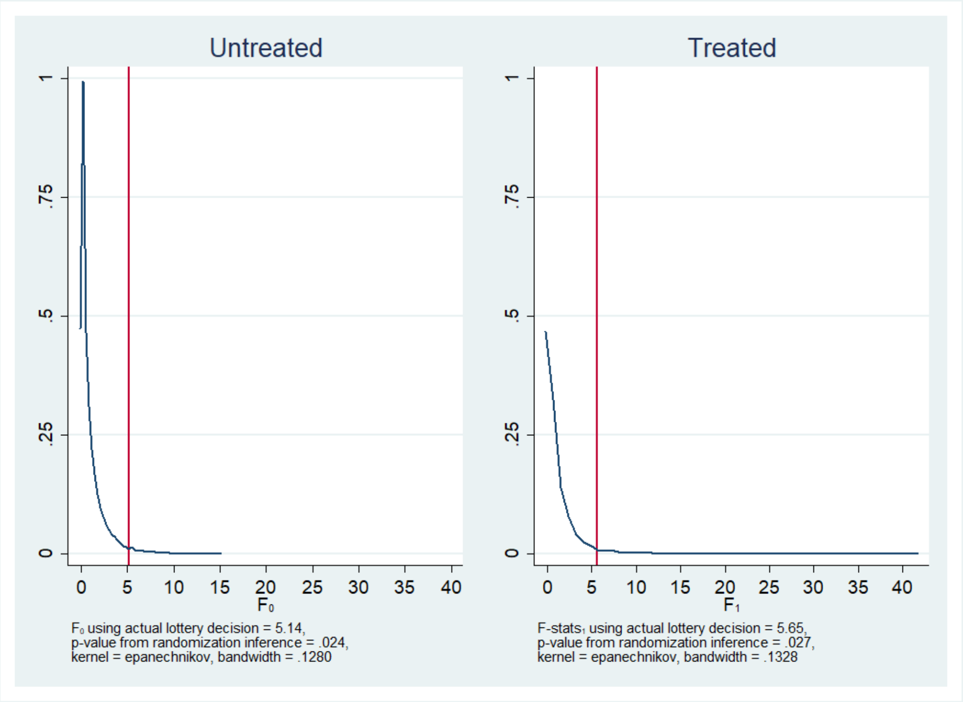


Estimate of π_0^2 using actual lottery decision = .00079,
p-value from randomization inference = .021,
kernel = epanechnikov, bandwidth = .00002

Estimate of π_1^2 using actual lottery decision = .0044,
p-value from randomization inference = .018,
kernel = epanechnikov, bandwidth = .00001

Notes: Binomial random assignment to lottery participation with probabilities of inclusion in the lottery by treatment status set at 0.11 for the untreated and 0.87 for the treated reflecting the shares observed in the data. Distributions constructed from 10,000 repetitions. The red verticle lines denote the squared differences in the mean retention between experimental and non-experimental populations within treatment status.

Figure D4: Distribution of F-statistics



Notes: Binomial random assignment to lottery participation with probabilities of inclusion in the lottery by treatment status set at 0.11 for the untreated and 0.87 for the treated reflecting the shares observed in the data. Distributions constructed from 10,000 repetitions. The red verticle lines denote the differences in the mean retention between experimental and non-experimental populations within treatment status.

Table D1: Nonparametric randomization testing results

Statistics	(1) estimate	(2) p-value	(3) mean	(4) p1	(5) p5	(6) p10	(7) p50	(8) p90	(9) p95	(10) p99
Coefficients										
Untreated										
Actual $\widehat{\pi_0}$	0.028	0.021								
Placebo $\widehat{\pi_0}$			0.000	-0.028	-0.020	-0.016	0.000	0.016	0.021	0.029
Treated										
Actual $\widehat{\pi_1}$	0.067	0.018								
Placebo $\widehat{\pi_1}$			0.000	-0.061	-0.044	-0.035	-0.000	0.037	0.048	0.069
t-statistics										
Untreated										
Actual t_0	2.27	0.024								
Placebo t_0			0.047	-2.120	-1.536	-1.212	0.025	1.333	1.745	2.516
Treated										
Actual t_1	2.38	0.027								
Placebo t_1			-0.100	-2.984	-1.950	-1.489	-0.008	1.164	1.472	2.032

Notes: Binomial random assignment to lottery participation with probabilities of inclusion in the lottery by treatment status set at 0.11 for the untreated and 0.87 for the treated reflecting the shares observed in the data. Distributions constructed from 10,000 repetitions. P-values constructed from the share of squared placebo estimated coefficients (t-statistics) greater than the squared actual estimated coefficients (t-statistics). The distribution of these squared statistics are shown in figures B3 and B4.