Do Learning Communities Increase First Year College Retention? Testing Sample Selection and External Validity of Randomized Control Trials

Tarek Azzam, Michael D. Bates, and David Fairris
September, 2019

Abstract:

Voluntary selection into experimental samples is ubiquitous and leads researchers to question the external validity of experimental findings. We introduce tests for sample selection on unobserved variables to discern the generalizability of randomized control trials. We estimate the impact of a learning community on first-year college retention using an RCT, and employ our tests in this setting. We compare observational and experimental estimates, considering the internal and external validity of both approaches. Intent-to-treat and local-average-treatment-effect estimates reveal no discernable programmatic effects, whereas observational estimates are significantly positive. The experimental sample is positively selected on unobserved characteristics suggesting limited external validity.

Contact: Azzam: Gevirtz Graduate School of Education, University of California, Santa Barbara, Santa Barbara, CA, 93106 (email: tarekazzam@ucsb.edu); Bates: 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

The goal of much empirical work in economics is to estimate a credibly causal (i.e., internally valid) effect, which also applies to the entire population of interest (i.e., is externally valid). In an effort to obtain credibly causal estimates, researchers typically exploit naturally occurring exogenous variation or employ randomized control trials (RCTs). However, even when the internal validity of estimates is credible, the same estimates often hold only for a localized subpopulation and may lack external validity (Imbens and Angrist, 1994). Whether the purpose of the research is to test hypotheses derived from theory, explain empirical regularities, or inform policy making, it is often desirable to extend the causal estimates beyond the population for which the estimates directly apply.[1] In RCTs, the experimental sample sometimes does, but often does not, directly correspond to the population of interest. As a result, the generalizability of the RCT results to the remaining population is pertinent to whether a given theory generally holds, the degree to which one mechanism explains a broader phenomenon, or whether a particular policy should be expanded or scaled back.

In this paper, we offer direct, widely applicable, and straightforward tests for selection into the experiment on the basis of unobserved characteristics and heterogeneous responsiveness to treatment.[2] We do so by comparing outcomes for those who do not receive treatment, by whether they participate in the experiment. We do the same among the treated populations comparing those who receive treatment by randomized assignment to those who received treatment without participating in the randomization. While each test may be suggestive by itself, we subsequently provide the conditions in which the coefficient magnitude and the rejection of the null hypotheses of homogeneous average outcomes within treatment status directly contradict claims of external validity. The benefit to researchers is that we illustrate a way to provide

[1] While the categorization of purposes of experiments is provided by Roth (1986), this broader point has been written about eloquently in Angrist, Imbens, and Rubin (1996), Heckman and Vytlacil (2005), Deaton (2009), Heckman and Urzua (2010), Imbens (2010), and Deaton and Cartwright (2017) as well as elsewhere.

[2] Our tests combine those from Hartman et al. (2015) and Sianese (2017) and relate to those from Huber (2013), Brinch et al. (2017), Black et al. (2017), Kowalski (2018), and Bertanha and Imbens (2019), many of which come out of the marginal treatment effects literature born out of Heckman and Vytlacil (2005).

concrete evidence of the external validity of their experimental results.

We explore these matters in the context of an analysis of the efficacy of a freshmen year learning community in increasing first year college retention at a large, four-year public research university. 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).

In the past decade or so, colleges have responded to the challenge of improving first year college retention by creating first year learning communities, which are viewed by higher education institutions and researchers as central to enhanced first-year retention and thus to graduation (Pitkethly and Prosser, 2001). 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 an RCT design to explore the extent to which the 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, freshmen students voluntarily enrolled in the program. We offer “intent to treat” (ITT) estimates of the effect of being randomized into treatment from among the self-selected 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 who comply with randomization. The ITT and LATE estimates of program impact reveal no 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 higher education institution.

Next, we consider the generalizability of these results. As a pretest, we follow Black et al. (2017) to test for selection on unobserved variables into compliance (or noncompliance) with the randomization.[3] Finding little evidence of nonrandom noncompliance, we then turn to implementing our novel tests addressing the question of generalizability of the RCT results to a possible broader population of interest – beyond those who self-select into the experiment. Non-representative experimental samples may originate through selection processes that are either researcher generated (as discussed in Allcott, 2015) or participant generated (as in the case at hand). In order to address the external validity of their work, many who design and implement RCTs attempt to randomly sample from the population or show the degree to which their experimental sample is representative of the broader population of interest. Despite researchers' best efforts to achieve a representative experimental sample, participants are often self-selected, even if only in granting consent. Many of the most influential experiments within economics rest on voluntary selection into the study. For example, in conducting the Perry Preschool experiment, researchers recruited and enlisted children from surrounding neighborhoods in Ypsilanti, Michigan who had IQ scores which ranged from 75 to 85 (Weikart et al. 1978). Individuals randomized in the National Supported Work Demonstration used in LaLonde's (1986) evaluation of observational methods, the Moving to Opportunity housing voucher experiment (Goering et al., 1999), and the Oregon health insurance experiment (Finklestein et al., 2012) are all self-selected into the experimental sample to some degree.[4]

[3] These tests are closely related to those in Huber (2013), Brinch et al. (2017), and Bertanha and Imbens (2019).
[4] In the National Supported Work Demonstration the majority of participants were drawn from those who had signed up, but were not enrolled in other government programs, but remaining slots were filled by "walk-in" enrollees (MDRC, 1980).

This self-selection can render RCT estimates externally invalid for a broader population. Selection may follow the Roy model, where those who benefit most from treatment select into the RCT, or a "reverse-Roy" selection process, where those who select into the RCT would do well even in the absence of treatment. Regardless of the process, in the presence of heterogeneous effects, nonrandom sample selection into an experiment may inhibit the generalizability of experimental results to a broader population. Moreover, similarity between the experimental sample and the remaining population on observed characteristics does not guarantee that their responsiveness to the intervention will be the same. Differences in unobserved characteristics may exist and present more persistent problems for estimation.

In order to consider these differences, we must first define the population of interest. The composition of the population of interest depends on the question and audience. Researchers often utilize RCTs to answer general or theoretical questions, which typically pertain to a broader population than the RCT sample. For instance, does neighbor quality, affect residents' educational, health, or employment outcomes (as in Ludwig et al., 2008)? Does early education lead to future educational and economic success (as in Weikart, 1998)? In general, for a pure policy evaluation in which policy makers are interested only in the effect of the policy on those who currently select into it, those who select into the study constitute the population of interest. However, for some policy evaluations the population of interest is much broader. Were the program expanded or the selection criterion changed, the program outcomes for these new entrants are also of interest.[5]

This is true of the FYLC program in several respects. First, a small share of the students who received treatment were late arrivals, and so never participated in the randomization process. Naturally, we should be interested in the effects of the program on this group of students who stand apart from the experimental population. Second, in later years, following the period of our analysis, the institution extended the program to nearly 90% of freshmen in the college in which the program originated. As a result, the population of interest for policy evolved, making

[5] This could arguably pertain to LaLonde (1986) and Finklestein et al., (2012) among others.

the external validity of our RCT results of some importance.

We implement our tests for sample selection on unobserved variables and responsiveness to treatment to examine the generalizability of our results to this larger student population by testing for selection (on unobserved characteristics) into the experiment. As with many experiments with human subjects, enrollment in the RCT was voluntary, and so there are questions of nonrandom selection on unobserved variables into the self-selected population. However, unlike many RCT-designed studies, we possess information on the non-experimental population as well as those who selected into the experiment. This enables us to explore the extent of otherwise unobserved differences between the experimental sample and the broader population of interest.

Our results reveal that those students who express a desire to enroll in the program are, in many observed respects, 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, further analysis reveals that the experimental population also 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. The magnitude of the selection into the RCT clearly raises concerns regarding the generalizability of the RCT results to these larger populations of interest.

A final contribution of the paper addresses LaLonde's (1986) seminal work on "within-study design." We begin by using the data to perform a comparison of the experimental results with those from standard observational approaches, which have been used both by institutional researchers and some academics to estimate the effects of first year learning communities. We also consider the internal and external validity of both experimental and standard observational approaches in order to reflect on what we learn from such within-study designs in general.

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.[6] Of those focused on learning communities specifically, some observational studies utilize advanced techniques, such as propensity score matching (Clark and Cundiff, 2011), instrumental variables (Pike, Hansen, and Lin, 2011), and Heckman's two-step procedure (Hotchkiss, Moore, and Pitts, 2006). However, in each, the exogeneity assumptions necessary for causal interpretation of the results are problematic.

There are three 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,

[6] See Barefoot, et al. (1998) and Pascarella and Terenzini (2005) for early reviews and Angrist, Lang, and Oreopoulis, (2009), Bettinger and Baker, (2014), Paloyo, Rogin, and Siminski, (2016) for more recent examples.

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 began on a small scale and included approximately 200 students from a population of roughly 4,000 incoming freshmen. During its founding there was a growing sense on campus that students – and freshmen in particular – were facing larger and more impersonal classes as enrollments had increased substantially during the preceding decade. The proposed first year learning community had several goals, but one of the most important was to increase first to second year retention rates of freshman students by offering them a small learning community experience in what was rapidly becoming a large research university setting.

The basic structure of the program is a year-long, theme-driven sequence of courses, structured study sessions, peer mentoring, and extra-curricular activities designed to foster academic achievement and socialization, and thereby to increase retention rates for freshmen participants. The FYLC is modeled after coordinated studies 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 – for example, the courses may all take place in the first term of freshman year as opposed to being spread out over the entire year, as is the case with the FYLC – 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.

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 doubled over two years. 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 student received the same information about the program and was encouraged to enroll in the lottery to be in the program. The goal was to receive expressions of interest by 1000 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).[7]

The new random assignment regime roughly approximates the old program implementation procedure, but with several differences that could have conceivably affected program participation and program outcomes pre- and post-random assignment. 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 summer enrollment sessions. Under the new regime, participants are randomly assigned from the self-selected population. Non-participants among the self-selected population under the old regime were simply unaware of the program or found that the FYLC classes were filled if they tried to enroll. Under the new regime, the control group was notified that they had not been chosen to participate in the program, perhaps giving them further encouragement to seek out alternative first-year experiences or disappointing them and thereby leading to behaviors that would not have occurred under the previous regime. Additionally, students and parents were given greater opportunity to discuss the program before expressing an interest in the program under random assignment.

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

[7] Some may worry about the lack of power due to a binary outcome. As a result, we also perform similar analysis with GPA as the outcome variable. 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 contains 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. The results for GPA are similar to those for first year retention. We find no statistically significant effects of the FYLC despite the increased power.

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	1565	1565	6566	8131

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 8131 students, 1565 of whom applied to be part of the FYLC, and 824 of which 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 self-selected out of 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.[8]

As mentioned above, there are three important instances of migration between assigned groups in the data.[9] Of the 824 students initially assigned to the treatment group, 170 (or 21%) did not attend any of the program courses or services. There is also contamination in the control sample in this randomized control trial, as 108 students (15%) assigned to the control group enrolled in FYLC courses (presumably as a partial replacement for those no-shows from the assigned treated group). Finally, 117 of 6,566 students (2%) who did not initially express interest in enlisting in the program and did not enter into the lottery eventually entered the program.

None of these groups is a random draw from the assigned treatment group. As shown in Table A1 in the Appendix, those who ultimately receive treatment are in some instances statistically significantly different from those in their original assignment category on almost every observable dimension. However, none of these violations of initial assignment bias the "intent to treat" estimates of program impact, though they do present complications in estimating the effects of treatment itself. However, their presence also provides opportunities for exploring the extent to which our estimated LATE can be generalized to the estimation sample or the entire population of interest.

[8] We discuss this matter further below and show that each of these traits is correlated with lower retention in Table A3 in the appendix.

[9] We present a figure depicting these various subpopulations in Figure A1 of the appendix.

Empirical Methodology

We divide our empirical analysis into three sections. First, we utilize the RCT design to identify the intent to treat effect of the FYLC on 1st year retention, as well as the average treatment effect on the treated. Second, we test for selection on unobserved characteristics between compliers and always-takers, between compliers and never-takers (non-random attrition), and (most importantly) for non-random selection into the experiment. Third, we compare the results from our RCT to estimates that would be obtained using standard observational methods, with a focus on the external and internal validity of each. More detail about each set of analyses is given below.

Analysis 1: Estimating treatment effects using the RCT

Randomization among the experimental group provides two groups of similar size; those assigned to treatment and those assigned to the control group. These two groups should be in expectation identical with respect to both observed and unobserved pre-determined characteristics. Accordingly, we may 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 OLS among the population who selected into the lottery according to the following specification:

$$Retention_i = \alpha + won_i\beta_{ITT} + \boldsymbol{X_i\gamma} + \epsilon_i, \quad (1)$$

where $Retention_i$ indicates whether student i remained in school the following year, won_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 $\boldsymbol{X}$.[10] 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).

[10] The inclusion of covariates may provide efficiency, but introduce finite sample bias. For summary of discussion see Lin (2013).

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 Wooldridge, 2014.[11] While this procedure provides us with internally-valid, causal estimates of the effect of treatment, without further assumptions these estimates hold only for the compliers who received treatment because they won the lottery. We may wonder whether there is nonrandom selection into these compliers and whether the estimated LATE generalizes to the average treatment effect among the whole experimental sample and, perhaps even more importantly, among the larger population of interest. We take up these issues in Analysis 2.

Analysis 2: Testing for selection on unobserved characteristics and external validity

In this set of analyses, we examine the extent to which our RCT results may generalize. In so doing, we first apply the tests as described in Black et al. (2017) to judge whether we can detect nonrandom selected noncompliance with the randomization within the experimental sample. We then introduce tests for whether there is nonrandom selection into the experimental sample on unobserved characteristics, and provide minimal conditions under which we may interpret the selection as directly contradicting the external validity of the RCT.

Existing tests for selection on unobserved variables and external validity within sample

To formalize these tests, let D indicate treatment (FYLC participation), Y be the outcome (first-year retention), and Z denote the binary randomized assignment (whether or not the lottery assigned an individual to participate in the FYLC). We add to this familiar framework, L, as an indicator for participation in the experiment. Much of the earlier treatment effects literature as well as Huber (2013), Brinch et al. (2017), Black et al. (2017), Kowalski (2018), and Bertanha and

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

Imbens (2019) considers only the population for which L=1. However, we are often interested in generalizability to a population broader than the sample, particularly in experimental settings. As a result, we must add two additional groups to the typical division of the sample among compliers, always-takers, and never-takers. Namely, we add the "late-takers" who take-up the treatment despite not entering the lottery, and the "never-ever-takers" who do not enter the lottery and do not take the treatment. Again, we maintain the monotonicity assumption that there are no defiers.[12]

Following Black et al. (2017), we model the conditional expectation of Y as a function of the treatment conditional on $\boldsymbol{X}=\boldsymbol{x}$. We will later use the unconditional outcome when addressing the external validity interpretation of the results. We write the model in familiar linear form with unobserved heterogeneous intercepts as well as heterogeneous effects of treatment:

$$Y_i = \boldsymbol{X_i}\boldsymbol{\gamma} + D_i b_i + \varepsilon_i, \tag{2}$$

where $\boldsymbol{X_i}$ denotes a vector of observed characteristics. Here, ε_i represents the unobserved heterogeneous intercept, while $b_i = \beta + e_i$ represents the heterogeneous responsiveness to treatment, which are centered on the ATE, β. Note that the model implicitly assumes that neither *Z* nor *L* directly affects the outcome variable, though selection into treatment may depend on both. Accordingly, we summarize the groups that comprise our sample, and write the expected outcome for each subsample conditional on $\boldsymbol{X}=\boldsymbol{x}$ in Table 2.

In order to test for nonrandom selection into compliance on the basis of unobserved heterogeneity we test whether the mean heterogeneous fixed errors and heterogeneous effects differ across populations using side-by-side comparisons. For instance, the difference between complacent controls and no-shows may be expressed as $E(Y_i|\boldsymbol{x}, D=0, L=1, Z=0) - E(Y_i|\boldsymbol{x}, D=0, L=1, Z=1) = \bar{\varepsilon}_c - \bar{\varepsilon}_{ns}$. Accordingly, we test whether this difference is zero in the following conditional mean function for the sample that enters the lottery but does not take up treatment:

$$E(Y_i|D_i=0, L_i=1, Z, \boldsymbol{X}) = Z_i\pi_{01} + \boldsymbol{X_i}\boldsymbol{\gamma_{01}}. \tag{3}$$

[12] We must also assume the "stable unit treatment value assumption (SUTVA)" from Rubin (1980) which holds that individuals' responsiveness to treatment is unaffected by the number of others who also receive treatment. This assumption may be restrictive as increases in scale may affect the quality of instructors and mentors providing services. However, we consider these concerns secondary to the selection effects present in our context.

Because the no-shows are composed only of never-takers and the control group of never-takers and compliers, this test ultimately assesses whether the compliers differ systematically on the basis of unobserved characteristics from never-takers. Thus, if we reject the null hypothesis that $\pi_{01} = 0$, then selection on unobserved characteristics may be problematic. We repeat the exercise among those in the experimental sample who receive treatment. Performing a standard t-test on the coefficient on the instrument tests whether $\bar{e}_t + \bar{\varepsilon}_t - (\bar{e}_{co} - \bar{\varepsilon}_{co})$ is nonzero. Whereas the former test examines whether there is selection into attrition, the latter tests for selection into the crossovers and also factors in possible heterogeneous treatment effects.

Table 2 Sample composition

Name	*Conditional outcomes*	*Type composition*
Complacent Treatment	$E(Y_i\|\boldsymbol{x}, D = 1, L = 1, Z = 1) = \boldsymbol{x\gamma} + \beta + \bar{e}_t + \bar{\varepsilon}_t$	*Compliers and always-takers*
Complacent Control	$E(Y_i\|\boldsymbol{x}, D = 0, L = 1, Z = 0) = \boldsymbol{x\gamma} + \bar{\varepsilon}_c$	*Compliers and never-takers*
No-shows	$E(Y_i\|\boldsymbol{x}, D = 0, L = 1, Z = 1) = \boldsymbol{x\gamma} + \bar{\varepsilon}_{ns}$	*Never-takers*
Crossovers	$E(Y_i\|\boldsymbol{x}, D = 1, L = 1, Z = 0) = \boldsymbol{x\gamma} + \beta + \bar{e}_{co} + \bar{\varepsilon}_{co}$	*Always-takers*
Never-ever-takers	$E(Y_i\|\boldsymbol{x}, D = 0, L = 0, Z = 0) = \boldsymbol{x\gamma} + \bar{\varepsilon}_n$	*Never-ever-takers*
Late-takers	$E(Y_i\|\boldsymbol{x}, D = 1, L = 0, Z = 0) = \boldsymbol{x\gamma} + \beta + \bar{e}_l + \bar{\varepsilon}_l$	*Late-takers*

Kowalski (2016), summarizes the test proposed by Brinch et al. (2017) remarking that these tests reject the null of external validity if the sign of the selection into compliance for the untreated is opposite the sign of selection into compliance among the treated. Kowlaski (2016) demonstrates that either ancillary assumption from Brinch et al. (2017) "weak monotonicity of the untreated outcomes in the fraction treated, [or] weak monotonicity of the treated outcomes in the fraction treated" is sufficient for directly testing the external validity of compliers for whom the estimates hold to the remainder of the experimental sample. Under this assumption, if the compliers are significantly and monotonically selected from both the never-takers and always-takers, she rejects the external validity of the LATE.

Proposed tests for selection on unobserved variables into experimental samples

Here, we build off of the previously mentioned literature focused on selection into the local population *within* an estimation sample, by providing to our knowledge the first tests for external

validity of experimental results to a broader population from which the experimental sample originates. The possibility of non-random sample selection of RCTs has received earlier attention.[13] Andrews and Oster (2018) model the decision to participate in a study based only on observed characteristics and approximate weights based on these approximations to provide bounds on the population average treatment effect. However, we worry that observed data will not fully capture the non-random selection. We test for selection into the estimation sample on the basis of unobserved variables by comparing outcomes by whether they participate in the experiment conditional on treatment status.

Comparison of outcomes between the experimental control and the untreated population has been conducted in Hartman et al. (2015), Lise, Seitz, and Smith (2015), Sianesi (2017), Galliani, McEwan, and Quistorff (2017), and Walters (2018). Of these Hartman et al. (2015) and Sianesi (2017) are closest to the work at hand. Sianesi (2017) develops nonparametric tests for randomization bias that compare the outcomes of the control group to those who opt out or were directed away from the study, similar to some of the tests we propose. Sianesi finds substantial selection on unobserved variables into the Employment Retention and Advancement experiment in the United Kingdom. Under the assumption of homogeneous average responsiveness to treatment across those who select into or out of the study (CIA-β), Sianesi attributes any differences in unobserved characteristics as being the result of the randomization itself. It seems difficult to maintain that responsiveness to treatment would be the same (CIA-β) across populations that differ significantly on unobserved characteristics. Indeed Galliani, McEwan, and Quistorff (2017) use a similar test as a placebo test for external validity.

Hartman et al. (2015) proposes a test comparing the outcomes of those who receive treatment within the randomized sample to those who receive treatment otherwise, thus incorporating heterogeneous responsiveness to treatment into the tests. When such "essential heterogeneity" (Heckman et al., 2006) is integrated into the tests, maintaining homogenous slopes

[13] For an early example see Hausman and Wise (1979) and for a recent example see Ghanem, Hirshleifer, and Ortiz-Becerra (2018).

despite significant differences on total unobserved heterogeneity seems unreasonable.

While each test may be suggestive on its own, we gain more information over existing tests by combining tests among both the treated and untreated. It is only through combining tests on the treated that we can directly assess external validity. We subsequently provide the conditions in which the coefficient magnitudes and the rejection of the null hypotheses directly contradict claims of external validity.

To introduce our tests for selection into the experimental sample on the basis of unobserved heterogeneity, we first provide a parametric framing to provide the intuition behind our straightforward tests. We then adopt a potential outcomes framework over simple and more complicated settings to demonstrate what our tests for selection reveal concerning the external validity of the estimates, and what assumptions would justify such an interpretation.

Such an examination is demanding on the data. In our application we benefit from access to the contemporaneous universe from the institution. However, only a representative sample of outcome data is necessary. While such data is large absent from existing work, we highlight the use for such data and remark that analysis samples are endogenous to experimental research designs. We begin by splitting the population into just four groups: those who entered the lottery and received treatment; those who entered the lottery and did not receive treatment; those who did not enter the lottery and did not receive treatment; and those who did not enter the lottery but did receive treatment. This last group – the late-takers – may not be present in all settings, but they are certainly not unique to our experiment.[14] They may be present in any randomized evaluation of an existing program, and may be absent from many experimental design because the usefulness to experimental design of a representative sample of treated individuals is not generally understood. The late-takers are not necessary in order to test for selection on unobserved variables, but are necessary for directly testing external validity.

[14] The compliers who were moved by housing demolitions in Jacobs (2004) and Chyn (2018) are essentially late-takers to the Moving-to-Opportunity compliers from Goering et al. (1999) and Chetty et al. (2016). Late-takers are also present in the data underlying the evaluation of the efficacy of Teach for America in Glazerman, Mayer, and Decker (2006) and in the large-scale class-size experiment of Tennessee STAR analyzed in Folger and Breda (1989), Krueger and Whitmore (2001), and Chetty et al. (2013) among many others.

Let $E(\varepsilon_i|D=0,L=0)=\bar{\varepsilon}_{00}$, $E(\varepsilon_i|D=0,L=1)=\bar{\varepsilon}_{01}$, $E(\varepsilon_i|D=1,L=0)=\bar{\varepsilon}_{10}$, and $E(\varepsilon_i|D=1,L=1)=\bar{\varepsilon}_{11}$. Likewise, let $E(e_i|D=1,L=0)=\bar{e}_{10}$ and $E(e_i|D=1,L=1)=\bar{e}_{11}$. Accordingly, the difference in outcomes conditional on $\boldsymbol{X}=\boldsymbol{x}$ within treatment status is given by the following:

$$E(Y_i|\boldsymbol{x},D=1,L=1)-E(Y_i|\boldsymbol{x},D=1,L=0)=\bar{e}_{11}+\bar{\varepsilon}_{11}-\bar{e}_{10}-\bar{\varepsilon}_{10} \quad (4)$$

$$E(Y_i|\boldsymbol{x},D=0,L=1)-E(Y_i|\boldsymbol{x},D=0,L=0)=\bar{\varepsilon}_{01}-\bar{\varepsilon}_{00} \quad (5)$$

If we restrict the sample to those who do not receive treatment, an indicator for participation in the experiment would absorb any mean differences in unobserved characteristics between those who do and do not participate in the experiment. Thus, we can test for selection into the experiment on the basis of such unobserved characteristics by conducting a simple t-test on the estimated coefficient on *L* in the regression of *Y* on $\boldsymbol{X}$ and *L* with this restricted sample:

$$E(Y_i|D_i=0,L,\boldsymbol{X})=L_i\pi_0+\boldsymbol{X_i\gamma_0}. \quad (6)$$

So long as the treatment and non-treatment do not differ by participation in the lottery and for any setting of covariates there is a chance to see each state of treatment, a substantially or significantly non-zero $\widehat{\pi_0}$ provides evidence of selection on unobserved characteristics into the experiment, making the claim of external validity of the experimental results to the non-experimental population difficult to accept. The intuition is simple, since neither received treatment, any differences in outcomes must be due to differences in selection. Further the sign and magnitude of $\widehat{\pi_0}$ demonstrates the extent and direction of the selection bias. Granted that some who did not enter the lottery made their way into treatment, we may conduct an additional test on the remaining sample, restricted to those who do receive treatment. A simple t-test on the coefficient of lottery participation among the treated provides a summative test of whether $\bar{e}_{11}+\bar{\varepsilon}_{11}-\bar{e}_{10}-\bar{\varepsilon}_{10}$ equals zero. Thus, we test whether those who do not enter the experiment differ from those who do on the basis of unobserved characteristics and heterogeneous effects.

Another way to approach the issue of selection on unobserved variables into the experiment is to compare the populations who select into treatment after enrolling in the experiment against those who select into the treatment without enrolling in the experiment, as well as doing the same

for those who choose not to take treatment at all. The idea here is that if in the natural world participation in treatment is voluntary, and the selection processes into (or out of) treatment are similar within and outside of the experimental setting, we can reveal whether participation in the experiment alters the findings. These comparisons will lack the power of the earlier tests, but with sufficient sample size may allow us more insight into the comparability of each population.

Nonparametric testing for selection and external validity beyond the experimental sample

In order to show what these tests reveal and the assumptions upon which our interpretation of the results rely, we revisit the potential outcomes framework where Y is the observed outcome, Y_1 is the outcome that would be manifested under treatment, and Y_0 is the outcome that would be manifested without treatment. As before, L=1 denotes participation in the lottery, Z=1 denotes being selected for treatment by the lottery, and D=1 indicates receipt of treatment. Let P (P= E(D|L=0)) stand for the share of those who do not participate in the lottery, but do receive treatment. Here we relax the assumption that **X** linearly enters the model and work with nonparametric unconditional means, to provide additional robustness and because we are focusing on responsiveness to treatment rather than selection on unobserved covariates.

We maintain throughout, that the randomization was carried out properly. That is $E(Y_1|L=1,Z=1) = E(Y_1|L=1,Z=0)$ and $E(Y_0|L=1,Z=1) = E(Y_0|L=1,Z=0)$. Second, we maintain that being selected for the control (or treatment) has no effect on the outcome independent of treatment status. Both of these assumptions are standard to interpreting experimental results.

We would like to test whether $E(Y_1|L=1) = E(Y_1|L=0)$ and $E(Y_0|L=1) = E(Y_0|L=0)$, but our data only contains realizations of the outcome (Y) in conjunction with realizations of lottery participation (L), treatment assignment (Z), and treatment status (D).

We begin with the simple case in which compliance with the randomization is perfect. It is clear that if the treatment status is homogeneous among the non-experimental population, then in performing a standard t-test comparing E(Y|L=1,D=0) to E(Y|L=0,D=0), we directly test whether there is selection into the experiment on the potential level of the outcome under no treatment with no further assumptions. Likewise, when the entire non-experimental population

receives treatment, comparing E(Y|L=1,D=1) to E(Y|L=0,D=1) provides a direct test of selection into the experiment on the potential level of outcome and responsiveness to treatment. However, rejecting the null hypothesis of no selection into the randomized sample on either potential outcome does not directly show a lack of external validity.

Testing of the external validity of RCT results requires data from the non-randomized population to contain both treated and untreated observations. The self-selection into or out of treatment among those who do not participate in the experiment requires us to make an additional assumption in order to interpret whether selection into the experiment is problematic for its generalizability. One reasonable candidate may be that of weakly monotonic selection by potential outcome:

Additional Assumption 1: If $E(Y_0|L=0,D=1) >> E(Y_0|L=0,D=0)$, then $E(Y_1|L=0,D=1) \geq E(Y_1|L=0,D=0)$ and if $E(Y_0|L=0,D=1) << E(Y_0|L=0,D=0)$, then $E(Y_1|L=0,D=1) \leq E(Y_1|L=0,D=0)$.[15]

If it were not for differences in potential outcomes between the 'if' and 'then' clauses, it would necessarily be true, as there cannot simultaneously be positive and negative selection into treatment among the non-experimental populations. However, whereas Y_0 only refers to selection on the unobserved outcome at baseline, Y_1 builds in both selection on the unobserved outcome at baseline and responsiveness to treatment. We do not believe this assumption is very restrictive, as it permits all cases where selection into treatment among the non-experimental population on unobserved variables is positively correlated responsiveness to treatment, and even permits cases where the two selection processes are opposed, so long as in those cases the selection on responsiveness to treatment is not so large as to reverse the overall direction of the nonrandom selection.[16] It does rule out instances where among the non-experimental population, differences in responsiveness to treatment among the treated and untreated are larger in

[15] We believe that the magnitude of the difference matters in this case as well as the precision with which it is estimated.

[16] Though applied to a different margin (participation in an experiment rather than assignment to treatment), it is similar to the least restrictive assumption applied in Kowalski (2016) to test the external validity of LATE to the remaining population with the estimation sample.

magnitude and opposite signed as the differences in unobserved levels of the outcome.

Assuming weakly monotonic selection into treatment in the nonexperimental population allows us to focus on differential treatment effects. We stratify by realized treatment status and write the expected differences in realized outcomes across the experimental and non-experimental populations as the following:

$$E(Y|L=1,D=0)-E(Y|L=0,D=0)= \quad (7)$$
$$E(Y_0|L=1)-E(Y_0|L=0)+P[E(Y_0|L=0,D=1)-E(Y_0|L=1)],$$
$$E(Y|L=1,D=1)-E(Y|L=0,D=1)= \quad (8)$$
$$E(Y_1|L=1)-E(Y_1|L=0)+(1-P)[E(Y_1|L=0,D=0)-E(Y_1|L=1)].$$

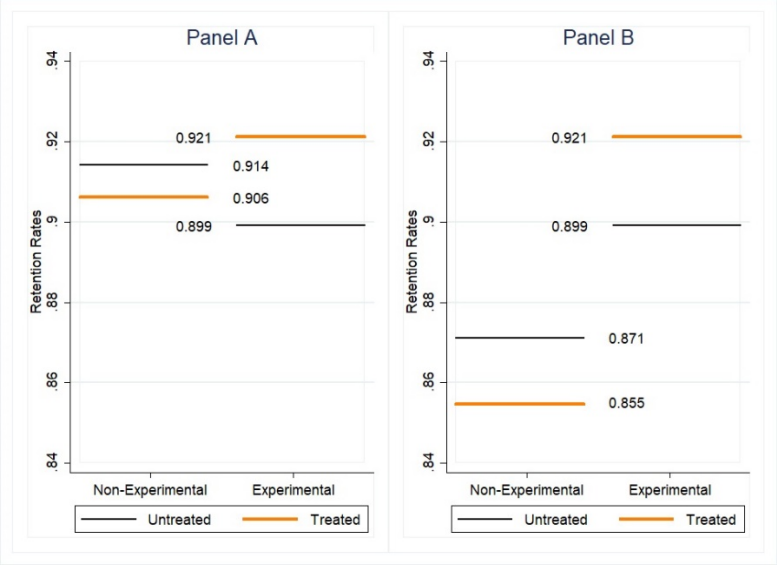
Equation (7) compares the expected outcomes among those who did not receive treatment by whether they participated in the lottery. The first difference on the right-hand side of equation (7) directly examines whether there is selection into the lottery on the basis of potential outcome in the absence of treatment. The latter difference could be nonzero either from selection into the lottery or selection into treatment in the non-randomized population.

Equation (8) compares the expected outcomes among those who did receive treatment by whether they participated in the lottery. Here, the first difference directly examines whether there is selection into the lottery on the basis of potential outcome in the event that both populations were to receive treatment. Again, the latter difference could be nonzero either from selection into the lottery or selection into treatment in the non-randomized population. Taken together, the two tests may demonstrate how problematic selection into the experiment is.

Figure 1 illustrates two possible scenarios of selection into the experimental sample on outcomes in order to describe what each might mean for the external validity of these hypothetical results. Selection into the experimental sample could be in the same direction among both the treated and untreated populations (as in Panel B) or in opposite directions (as in Panel A). In both panels it appears that treatment effects are of differing magnitudes and sign

among experimental nonparticipants than among experimental participants. However, these differences in the gaps between the treated and untreated could be driven by selection into treatment among the experimental nonparticipants. Consequently, while intuitively troubling, we cannot conclude from these results that the RCT findings are externally invalid.

Figure 1: Selection into experimental sample and external validity



In contrast, Panel B depicts a case where selection into the experiment is of common sign among the treated and untreated. Referring back to equation (7), the inequality among the untreated could be due to those who select into the lottery having higher potential outcomes on average than those who do not participate, or to those who select into treatment, but not the lottery, having higher than average potential outcomes than the remaining non-lottery population (thus pushing down the average among the untreated non-participants). Similarly, referring back to equation (8) the inequality of outcomes among the treated could be due to those who select into the lottery having on average higher potential outcomes than those who do not, or due to those who do not select into treatment nor the lottery having on average higher potential outcomes under treatment than those who do select into treatment but did not enter the lottery. However, under weakly monotonic selection on potential outcomes among those who do not enter the lottery, we cannot simultaneously maintain that those who chose to receive treatment are both positively and negatively selected on their propensity to persist in college. Thus, in order to reject random selection on unobserved variables into the experiment, our tests require:

1) significant differences between the experimental and nonexperimental populations in the outcome, and 2) differences that are consistent in sign between the treated and untreated groups.

Turning to external validity, leaning on Kowalski's (2016, 2018) examination of external validity *within* experimental samples, under the assumption that either potential outcome is monotonically related to probability of treatment, we can interpret these same results as indicative of external invalidity. However, this assumption is strong and leads us to reject external validity even when the differences in treated and untreated average outcomes is identical between experimental participants and nonparticipants. We gain further evidence of external invalidity if selection into the experiment is the same sign among both the treated and untreated, and the difference between equations (8) and (7) is large in magnitude, applying no additional assumptions. The most reassuring case for generalizing experimental results occurs with very small differences in outcomes of opposed sign between experimental participants within each treatment status, particularly if they are precisely estimated. If the sign of selection into randomization differs by treatment status, such differences may well be caused by selection into treatment among nonparticipants of the experiment.

In the case at hand, the presence of both no-shows and crossovers indicates that compliance with the randomization is not perfect. We accordingly maintain the standard monotonicity assumption from Imbens and Angrist (1994). In order to examine external validity in the presence of such noncompliance we adopt a second additional assumption, namely ignorability of noncompliance:

Additional Assumption 2: $E(Y|L=1,D=0) = E(Y_0|L=1,Z=1,D) = E(Y_0|L=1,Z=0,D)$ and $E(Y|L=1,D=1) = E(Y_1|L=1,Z=1,D) = E(Y_1|L=1,Z=0,D)$.

On its face, this assumption is nontrivial and likely does not hold in many instances. However, though we cannot exactly observe whether it holds, we may gain insight into its plausibility by conducting the tests previously described at the beginning of this section. It would be unreasonable to expect generalizability out of experimental sample in the presence of significant selection into compliance within the experimental sample. Accordingly, we only apply the formal tests for external validity if the differences in outcomes between the complacent controls and no-shows and between the complacently treated and crossovers near zero.

Analysis 3: Within study design

In this section, we conduct a conventional observational analysis of program impact on the treated population. We conduct this analysis with two purposes in mind. First, we compare the observational results with the experimental estimates to explore issues of bias in conventional observational designs where the population of interest may be only those students who voluntarily enroll in the experiment. 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 enrolment.

Secondly, we conduct this within-study design to reflect on within-study designs themselves in the context of our tests for external validity, where the population of interest extends beyond those who self-selected into the experiment. Differences in results between the two approaches may originate from a lack of internal validity or a lack of external validity of either observational or RCT approaches. We perform a decomposition of the observational results to provide evidence for the cause of any divergence in results from these two approaches.

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 analysis used to identify our average intent to treat estimates except that these analyses use the full sample of freshman entrants and treatment is measured by an indicator for enrollment in the FYLC instead of by an indicator for winning the lottery.

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 also report estimates of the average treatment effect for comparability.

We use the same vector of observed covariates as we use as controls as first described in

Analysis 1. Appendix Table A3 shows that in particular high-school GPA, math SAT score, living on campus, and (negatively) being a first generation college student each predict persisting in college, while Appendix Table A1 shows that students are inversely selected for retention into the treatment on these same dimensions. We adopt the standard practice of using logit to estimate the propensity scores. We bootstrap the standard errors to account for estimation error.

Validity of this and similar techniques require two assumptions; overlap and ignorability. The overlap assumption requires that, for any setting of observed characteristics, there is a chance the individual could be in either the treatment or control group. We can examine the overlap assumption through the estimated propensity scores. Figure A2 presents histograms of these estimated propensity scores split by treatment status. Crump, et al. (2009) provide a rule of thumb that observations with propensity scores above 0.9 and below 0.1 should be discarded. We accordingly perform all analyses both on the full sample as well as this trimmed subsample. The ignorability assumption requires sufficient information in the control variables such that there would be no expected difference in retention between those who receive treatment and those who do not in the absence of treatment ($E(Y_i|\boldsymbol{x}, D = 1) - E(Y_i|\boldsymbol{x}, D = 0) = 0$).

We consider what we learn from comparing the two approaches in light of the tests for selection and external validity. For ease of explanation let us compare the estimates from OLS to those from the RCT. Note that $plim\widehat{\beta_{OLS}} = ATE + Bias_{OLS}$ and $plim\widehat{\beta_{RCT}} = LATE$, assuming proper randomization. That is if we assume overlap (or only consider the population on which the overlap is thick) and ignorability the two estimators would converge to different parameters. We can relate the two parameter according to the following:

$$ATE = LATE \times P(compliers) + E(e_i|compliers = 0) \times \big(1 - P(compliers)\big). \quad (9)$$

We typically cannot observe whether the ignorability assumption holds, and suspect the OLS estimate may be biased. Thus, in comparing the two estimates, we observe the following:

$$plim\widehat{\beta_{OLS}} - plim\widehat{\beta_{RCT}} = E(e_i|compliers = 0) - LATE + \frac{Bias_{OLS}}{1-P(compliers)} \quad (10)$$

Equation (10) nicely demonstrates the comparison in results provides a mixture of the possible external invalidity of the RCT ($E(e_i|compliers = 0) - LATE$) and a scaled measure of the

possible bias in the OLS estimate. Thus, differences in estimates alone do not imply that the RCT estimate is to be preferred.

Finally, we conduct a decomposition of the full-sample OLS estimate by regressing first-year retention on participation in the experiment, assignment to treatment, treatment, the interaction of participation and treatment assignment, and the triple interaction between participation assignment to treatment, and receipt of treatment. This exercise illustrates which selection processes drive the results and from where the deviations from the ATE originate.

Empirical Results

Analysis 1

Table 3: RCT estimates

	(1) Retention	(2) Retention	(3) Retention
Panel A: ITT effects of winning lottery on first year retention (reduced form estimates)			
Won lottery	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)			
Won lottery	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

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.

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 variation in FYLC participation. The estimated impacts of the program in the second-stage regression analysis increase in quantitative magnitude – by roughly one percentage point – compared to 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 some of the analysis presented in Analysis 2 below. 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.

Analysis 2: Table 4 presents the results of an analysis examining selection into both the experimental sample and the complier subset within that sample. We will use these results to examine the external validity of the RCT LATE. Panel A applies tests from the existing literature to explore whether the compliers systematically differ from the always-takers and never-takers. Panels B and C apply our proposed tests for selection into the experimental sample among the untreated and treated populations respectively.

Columns (1) and (2) of Panel A compare the retention probabilities of no-shows and the untreated population. Comparing the estimated coefficient on being randomly selected for participation in the FYLC program (i.e., having “won” the lottery) across the two columns, there is no statistically significant change in the magnitude of the estimate and thus no detectable substantive difference in the impact of controlling for observed characteristics across the two populations as regards their retention prospects. Moreover, the estimated coefficient on “won” in the column (2) results with controls is statistically insignificantly different from zero, implying

no detectable substantive difference across the two populations regarding the impact of unobserved characteristics on retention.

Table 4: Testing for selection into and within the lottery

	(1)	(2)	(3)	(4)
Panel A: Test for nonrandom attrition and noncompliance within the lottery				
Won	0.009	0.000	0.005	0.005
	(0.025)	(0.026)	(0.029)	(0.029)
Observations	803	803	762	762
Controls	No	Yes	No	Yes
Sample	Control + No-shows	Control + No-shows	Treated + Crossovers	Treated + Crossovers
Treatment status	Untreated	Untreated	Treated	Treated
Panel B: Test for selection into the experiment among the untreated				
Lottery	0.028**	0.039***	0.035	0.040*
	(0.011)	(0.011)	(0.023)	(0.023)
Observations	7252	7252	6619	6619
Controls	No	Yes	No	Yes
Sample	Control + No-shows + Never-ever-takers	Control + No-shows + Never-ever-takers	No-shows + Never-ever-takers	No-shows + Never-ever-takers
Treatment status	Untreated	Untreated	Untreated	Untreated
Panel C: Test for selection into the experiment among the treated				
Lottery	0.067*	0.063*	0.062	0.062
	(0.034)	(0.033)	(0.042)	(0.045)
Observations	879	879	225	225
Controls	No	Yes	No	Yes
Sample	Treated + Crossovers + Late-takers	Treated + Crossovers + Late-takers	Crossovers + Late-takers	Crossovers + Late-takers
Treatment status	Treated	Treated	Treated	Treated

All results are from OLS regressions. Robust Huber-White standard errors in parentheses. *** p<0.01, **p<0.05, * p<0.1.

Columns (3) and (4) do the same, but exploring selection issues regarding the retention probabilities of the treated and crossovers populations – crossovers, being those who migrated from the control population to become treated despite losing the lottery. The results are similar; we see little difference in retention propensities across the crossovers and treated populations based on differences in either observed or unobserved background characteristics. As the

coefficient estimates are in the same direction, and more importantly qualitatively small, following Kowalski (2016), Brinch et al. (2017), and Kowalski (2018), we fail to reject the hypothesis of homogeneous treatment effects within the experimental sample. Thus, the Panel A results find little reason to worry about the two migrations within the experiment (despite the differences on observed variable among migrants), and provide some reassurance that the RCT LATE may generalize to the entire sample who selected into the experiment.

Columns (1) and (2) of Panel B explore the extent to which those who selected into the lottery, but were untreated, differ regarding the probability of retention from the "never-ever takers" (i.e., those who did not select into the lottery and did not later become treated as late-takers). Column (1) provides the unconditional estimates, such that the reported coefficient provides the nonparametric difference in the mean outcomes of the untreated by whether or not they participated in the lottery. Column (2) conditions on predetermined observed student characteristics. While this approach may introduce finite sample bias, it is also generally more efficient (though not noticeably in this case) and focuses attention on the differences on unobserved variables. The fact that the coefficient on Lottery is statistically and economically significantly positive in both specifications indicates that those who enter the lottery are more likely to persist in college regardless of the program, as neither population in these regressions took part in the FYLC. The fact that the magnitude of the coefficient grows from 0.028 (p-value = 0.013) to 0.039 (p-value = 0.001) with the addition of covariates indicates that lottery participants are negatively selected on observed characteristics – something we indicated in the comparison of background characteristics across these two populations in the "Background and Data" section above. However, the positive selection into the lottery based on unobserved characteristics is more pronounced than the negative selection on observed variables.

Columns (3) and (4) of Panel B test for differences across the never-takers and the never-ever-takers in retention probabilities. The former expressed an interest in the lottery but, having won, decided not to participate in the program, whereas the latter also did not participate in the program but never expressed a desire to do so. Neither group was treated; the difference is in selection into the lottery. Once again, we find evidence of positive selection on unobserved characteristics among those who entered the lottery. With the smaller sample size, these estimates are less precise, but the magnitudes are roughly comparable to those of columns (1) and (2). From column (4), we estimate that the never-takers are 4 percentage points more likely to persist beyond the first year (p-value = 0.077) than are the never-ever-takers. The Panel B results indicate that there is positive selection into the lottery based on unobserved characteristics for the untreated population, and thus that the RCT findings of program impact cannot be generalized to the students who elect not to participate in the lottery and who maintain that commitment.

In Panel C we turn to selection into the lottery among the treated populations. Columns (1) and (2) compare retention probabilities for the treated population that selected into the lottery and those late-takers who expressed no interest in the program initially, but later changed their minds and were admitted into the FYLC. We find that, among the treated, those who entered the lottery are roughly 6 percentage points (p-value = 0.062) more likely to persist than those who came into the program as late-takers.[17]

In columns (3) and (4), we compare two final treated groups – the crossovers and late-takers – both of whom were treated and migrated from initially assigned or chosen positions in order to receive treatment. Once again, the central distinguishing feature of these two groups is

[17] Confidence intervals are even tighter using randomization inference as shown in Appendix B.

the initial decision to participate in the lottery. While the results reveal no statistically significant difference in retention probabilities across these two groups at conventional thresholds, the magnitude of the difference owing to unobserved characteristics is very large (equivalent to the estimate in the first two columns). This is likely due to the much-reduced sample size.

What do these results mean for the external validity of the RCT? To summarize, the results from Panel A reveal no discernable selection into the compliers. The estimates are very small and far from statistically significant. This analysis accordingly provides reassurance that the additional assumption of ignorable noncompliance may hold. Secondly, the analysis from Panels B and C reveal substantial positive selection into the experimental sample among both the treated and untreated. Under the assumption that potential outcomes are monotonic with respect to the probability of treatment, as in Kowalski (2016, 2018), we would reject the external validity of the RCT estimates, as the assumption implies that $E[Y_0|D=1,L=0] \geq E[Y_0|\ L=1]$ or $E[Y_1|D=0,L=0] \geq E[Y_1|\ L=1]$.

Imposing only our earlier assumptions (most critically monotonicity of selection on potential outcomes), we attempt to reconcile the difference in outcomes between treated and non-treated students within and outside of the experiment. We begin by taking the difference between equation 7 and 8 from section 3, providing the following:

$$E(Y|L=1,D=1) - E(Y|L=0,D=1) - [E(Y|L=1,D=0) - E(Y|L=0,D=0)] =$$
$$E(Y_1|L=1) - E(Y_0|L=1) - [E(Y_1|L=1) - E(Y_0|L=0)] \quad (11)$$
$$+(1-P)[E(Y_1|L=0,D=0) - E(Y_1|L=1)] - P[E(Y_0|L=0,D=1) - E(Y_0|L=1)].$$

If the RCT is externally valid, then $E(Y_1|L=1) - E(Y_0|L=1) - [E(Y_1|L=1) - E(Y_0|L=0)] = 0$. Yet, from table 4 we estimate that $E(Y|L=1,D=1) - E(Y|L=0,D=1) = 0.067$, and that $E(Y|L=1,D=0) - E(Y|L=0,D=0) = 0.028$, making the estimated

differences in differences depicted in equation 11 and shown directly in column 3 of Table A4 equal to 0.038. Further, in our data the share of the nonexperimental population that receives treatment (P) is approximately 0.018. Under the standard experimental assumptions of proper randomization and no independent effects of assignment the data reveals estimates of $E(Y_1|L = 1) = 0.921$ and $E(Y_0|L = 1) = 0.899$.

We achieve the maximum value of equation 11 under external validity by using the maximum possible value of $E(Y_1|L = 0, D = 0)$ and minimum possible value of $E(Y_0|L = 0, D = 1)$. Under monotonic selection on potential outcome, $E(Y_1|L = 0, D = 0)$ cannot exceed $E(Y_1|L = 1)$ or 0.921, as we observe positive selection into the experiment among both the treated and untreated individuals.[18] The minimum possible retention rate in the absence of treatment in the nonexperimental sample may be as extreme as zero following Lee (2009). Substituting these values into equation 11 provides the following:

$$0+0.982(0.921-0.921)-0.018(0-0.899)=0.016. \quad (12)$$

Note that the maximum difference we can generate in the absence of differential treatment effects between the experimental and nonexperimental populations under external validity is less than half the difference we actually observe in the data.[19] Though this difference is not statistically distinguishable from zero, with selection into treatment explaining at most so little of the difference in outcomes, it seems extremely unlikely that we would realize as positive treatment effects were we to substitute a random sample from the nonexperimental population into the experiment in place of those who volunteered to participate.

[18] A violation of this assumption seems unlikely and would run counter to economic intuition, as a violation would imply the ATE for the never-ever-takers is so much larger than the ATE among the compliers, that it reverses the positive selection into the experiment on outcome levels.

[19] Without the monotonicity in potential outcomes assumption, in order to explain the total difference, we would need to assume extreme adverse selection into treatment among the nonexperimental sample. The never-ever-takers have over 270 percent higher treatment effects than the compliers within the RCT.

Analysis 3: 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? Further, what can we learn about competing research designs by such a comparison, when the parameter of interest pertains to a larger population than that on which we have strong exogenous variation? We present the results from observational approaches to estimate the program impact where the treated, including crossovers and late-takers, are compared to non-participants that include both the control, no-shows, and never-ever-takers in Table 5.

Table 5: 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.

Contrary to the findings from the RCT design, the Table 5 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. However, because selection into treatment largely transpires through selection into the RCT, we know from the results in Panel C of Table 4 that this estimate is biased due to self-selection on unobserved characteristics.

Curiously, 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. However, as we restrict the sample to that for which there is more overlap on observed covariates, the observational estimates universally grow. While students who select into the experiment may be vulnerable with regard to observed correlates regarding first-year retention, this vulnerability is combined with unobserved characteristics that more than make up for their observational vulnerabilities.

We emphasize what is driving the differences between observational and RCT estimates by decomposing the findings from Table 5 into selection into each of the six populations. As the exercise is an extension of the tests for nonrandom noncompliance outlined in Huber (2013) and is similar to Analysis 2 we report the estimates in Table A5 in the Appendix. The analysis shows that the only statistically meaningful difference in retention prospects is between those who do and do not enter the lottery (p-values of 0.038 and 0.003 respectively).

How do we accordingly assess the experimental and observational approaches? Here, the RCT seems to provide a credible estimate of the average causal effect of the first-year learning community on those who choose to participate in the experiment, while the observational approaches provide only inconsistent results. However, in this context the observational estimates fail to provide a plausibly causal estimate due to selection into the experiment itself. The majority of our treated population received treatment by selecting into the study. As a result, the nonrandom selection that may compromise the external validity of the experimental results directly undermines the internal validity of the observational estimates. Considering this strong

selection into the experiment and consequently into treatment, it is difficult to claim that either estimate – the experimental or the observational – captures the population average effect of the FYLC on first-year college retention. Thus, the take-away from the difference in observational and experimental results is not the universal superiority of experimental approaches, but rather that differences in results from different approaches are symptomatic of persistent problems in uncovering the population parameter. By applying tests for external validity, we can determine whether the RCT delivers a valid estimate of this elusive parameter in each specific context.

Conclusions

We began our analysis with an RCT design to estimate the impact of a learning community on first-year college retention for those who select into the study. The results are the first of their kind to employ an RCT to address this question at a large, four-year research university. We find that both the "intent to treat" and the "local average treatment effect" estimates of program 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 issues related to external validity of the RCT results. There were significant migrations from the assigned populations in the experimental sample. In conducting existing tests for whether the assignment of compliers differs substantially from the never-takers or always-takers on the basis of unobserved propensities to persist, we find little difference. As a result, it seems reasonable to generalize the "local average treatment effect" estimate of program impact to the remaining experimental population.

However, when we add tests for whether the experimental sample is representative of a broader population of interest – for example, the entire freshman class – we find that those who enter the lottery, and thereby express initial interest in the first-year learning community program, are quite different from those who elect not to enter the lottery. In particular, we find that lottery participants (whether or not they receive treatment) possess unobserved characteristics that lead them to be far more likely, statistically and quantitatively, to return for a

second year of college compared to those who decline to participate in the lottery. Thus, while the RCT findings may serve as a causal and unbiased estimate of program impact for the students who self-selected into the study, we caution against generalizing these results to the population who did not enroll in the experiment. Further, it appears from the differences-in-differences between experimental and nonexperimental and treated and non-treated populations that the experimental results reflect a relatively optimistic view of the program. These results suggest smaller or possibly even negative programmatic effects for those who do not select into the experiment. Taken overall, we believe the results introduce warranted pessimism on the efficacy of the learning community program in its current form.

This study also serves to highlight a few important broader lessons, all of which emanate from the central insight that selection on unobserved characteristics matters. The analysis reveals that students who selected into the study disproportionately possess observed background characteristics that are negatively associated with first-year retention. To many, it might appear reasonable to expect that unobserved differences across the self-selected experimental and non-experimental populations are likely to follow the same pattern as do observed differences, and therefore that observational analyses of program impact would *underestimate* the true effect of the program. Such is not the case here. Our results reveal that the unobserved characteristics of the self-selected control population have a strongly positive and statistically significant effect on first-year college retention. This implies that balance tests based on observed variables are insufficient evidence of the representativeness of the sample or external validity of the study.

Empirical researchers generally do not test for selection on unobserved variables into the estimation sample, either because the data on nonparticipants do not exist or because researchers have not made use of it. Yet selection issues may emerge in many of these contexts and matter greatly for the external validity of results. Information on non-participants in the experiment is critical in testing for such non-random selection. Thus, we suggest, as something of a “platinum standard,” that researchers connect RCTs to more comprehensive data on the larger population of interest, and show concretely whether the results of their study generalize beyond the

experimental sample. This could take the form of within the RCT using exact questions from existing surveys on a random sample of the population of interest, or reserving resources to survey a random sample on the outcome, or ideally linking the RCT to administrative data encompassing the population of interest.

Regarding differences between RCT and observational analyses, absent further testing researchers ought not to conclude that observational approaches are obviously inferior when observational estimators yield different results from those of an RCT design. The context matters greatly. Assuming the observational analysis and RCT rely upon samples with different selection processes (e.g., the population from administrative data, a random sample from the population, a researcher non-randomly selected sample, or a participant self-selected sample), the two approaches estimate different parameters. In which case, the representativeness of the samples determines which estimate provides a closer approximation to the effect for a broader population. The tests that we have introduced provide concrete evidence of where these biases lie and should be incorporated into any similar within-study design.

Finally, researchers should reflect more on what constitutes the population or parameter of interest. Economists and many other social scientists are often interested in parameters pertaining to broad populations. What is the elasticity of labor? What is the effect of health insurance on health or financial stability? Does the neighborhood in which an individual lives affect the course of their lives? Does a learning community help freshmen to persist in college? Each of these regard populations that are broader than those who may be selected for and may select into an experiment. Which parameter is of most interest is context dependent and may be determined by whether we are in the phrasing of Roth (1986) “speaking to theorists” or “whispering in the ear of princes.”

References

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

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

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.

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.*

Bloom, Howard S. 1984. "Accounting for no-shows in experimental evaluation designs." *Evaluation Review* 8, no. 2: 225-246.

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

Calónico, Sebastian, and Jeffrey Smith. 2017. "The Women of the National Supported Work Demonstration." *Journal of Labor Economics* 35, no. S1: S65-S97.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How does your kindergarten classroom affect your earnings? Evidence from Project STAR." *The Quarterly Journal of Economics* 126, no. 4: 1593-1660.

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz. 2016. "The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment." *American Economic Review* 106, no. 4: 855-902.

Chyn, Eric. 2018. “Moved to Opportunity: The Long-Run Effect of Public Housing Demolition on Children.” *American Economic Review*, 108(10): 3028-3056.

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.

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.

Deaton, Angus S. 2009. *Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development*. No. w14690. National Bureau of Economic Research.

Deaton, Angus, and Cartwright, Nancy. 2018. "Understanding and misunderstanding randomized controlled trials." *Social Science & Medicine* 210: 2-21.

Efron, Bradley. 1982. *The jackknife, the bootstrap, and other resampling plans*. Vol. 38. Siam.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and Oregon Health Study Group. 2012. "The Oregon health insurance experiment: evidence from the first year." *The Quarterly Journal of Economics* 127, no. 3: 1057-1106.

Fisher, Ronald A., 1935. *The Design of Experiments* (Edinburgh: Oliver and Boyd).

Folger, John, and Carolyn Breda. 1989. "Evidence from Project STAR about class size and student achievement." *Peabody Journal of Education* 67, no. 1: 17-33.

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.

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

Glazerman, Steven, Daniel Mayer, and Paul Decker. 2006. "Alternative routes to teaching: The impacts of Teach for America on student achievement and other outcomes." *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management* 25, no. 1: 75-96.

Goering, John, Joan Kraft, Judith Feins, Debra McInnis, Mary Joel Holin, and Huda Elhassan. 1999. "Moving to Opportunity for fair housing demonstration program: Current status and initial findings." *Washington, DC: US Department of Housing and Urban Development*.

Gourieroux, Christian, Alain Monfort, and Alain Trognon. 1984. "Pseudo maximum likelihood methods: Theory." *Econometrica: Journal of the Econometric Society*: 681-700.

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. "Attrition bias in experimental and panel data: the Gary income maintenance experiment." *Econometrica: Journal of the Econometric Society* (1979): 455-473.

Heckman, James J., and Sergio Urzua. 2010. "Comparing IV with structural models: What simple IV can and cannot identify." *Journal of Econometrics* 156, no. 1: 27-37.

Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2006. "Understanding instrumental variables in models with essential heterogeneity." *The Review of Economics and Statistics* 88, no. 3: 389-432.

Heckman, James J., and Edward Vytlacil. 2005. "Structural equations, treatment effects, and econometric policy evaluation 1." *Econometrica* 73, no. 3: 669-738.

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.

Imbens, Guido W. 2010. "Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature* 48, no. 2: 399-423.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica: Journal of the Econometric Society*: 467-475.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Ishler, Jennifer L., and M. Lee Upcraft. 2005. "The keys to first-year student persistence." *Challenging and supporting the first-year student: A handbook for improving the first year of college*: 27-46.

Jacob, Brian A. 2004. "Public housing, housing vouchers, and student achievement: Evidence from public housing demolitions in Chicago." *American Economic Review* 94, no. 1: 233-258.

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.

Krueger, Alan B., and Diane M. Whitmore. 2001. "The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR." The Economic Journal 111, no. 468: 1-28.

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

LaLonde, Robert J. 1986. "Evaluating the econometric evaluations of training programs with experimental data." *The American economic review*: 604-620.

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.

Lee, David S. "Training, wages, and sample selection: Estimating sharp bounds on treatment effects." *The Review of Economic Studies* 76, no. 3 (2009): 1071-1102.

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.

Ludwig, Jens, Jeffrey B. Liebman, Jeffrey R. Kling, Greg J. Duncan, Lawrence F. Katz, Ronald C. Kessler, and Lisa Sanbonmatsu. "What can we learn about neighborhood effects from the moving to opportunity experiment?." *American Journal of Sociology* 114, no. 1 (2008): 144-188.

Manpower Demonstration Research Corporation. 1983. *Summary and findings of the national supported work demonstration.* Ballinger Publishing Company.

Meyer, Bruce D., Wallace KC Mok, and James X. Sullivan. 2015. "Household surveys in crisis." *Journal of Economic Perspectives* 29, no. 4: 199-226.

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.

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."

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.

Steiner, Peter M., and Yongnam Kim. 2016. "The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases." *Journal of Causal Inference* 4, no. 2.

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.

Weikart, D. P., Epstein, A. S., Schweinhart, L., Bond, J. T., 1978. The Ypsilanti Preschool Curriculum Demonstration Project: Preschool Years and Longitudinal Results. High/Scope Press, Ypsilanti, MI.

Weikart, David P. "Changing early childhood development through educational intervention." *Preventive Medicine* 27, no. 2 (1998): 233-237.

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.

Young, Alwyn. 2018. "Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results." *The Quarterly Journal of Economics*134, no. 2: 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:

Figure A1: Map of the populations within the data

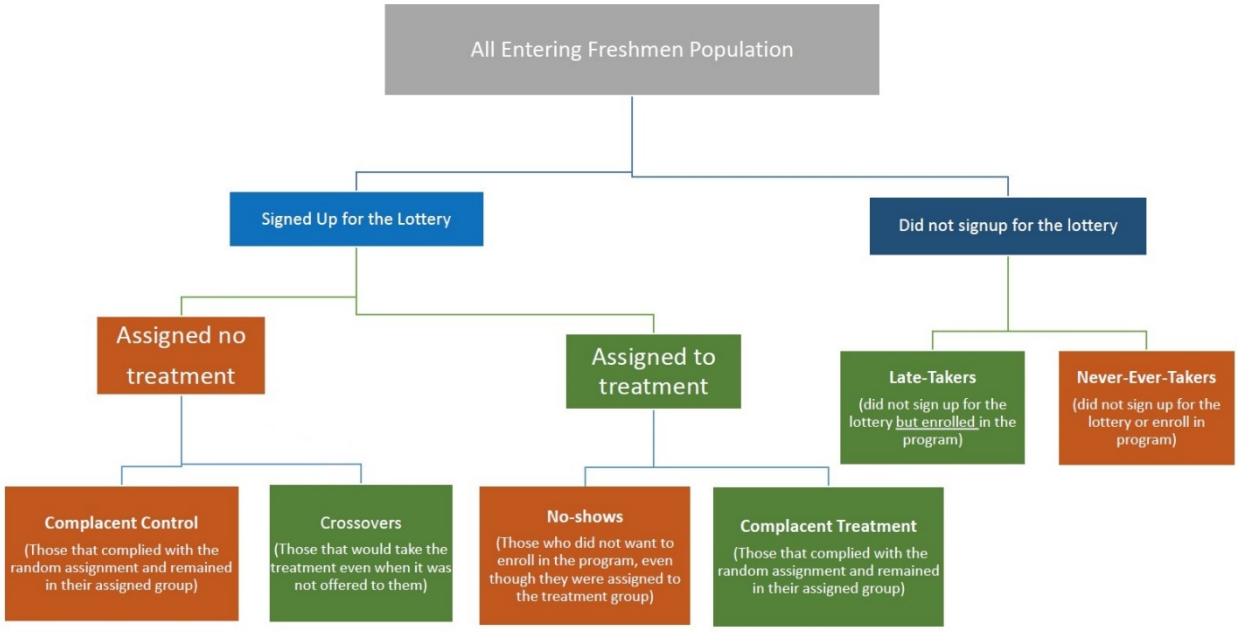


Figure A2: Overlap in the propensity scores by treatment status

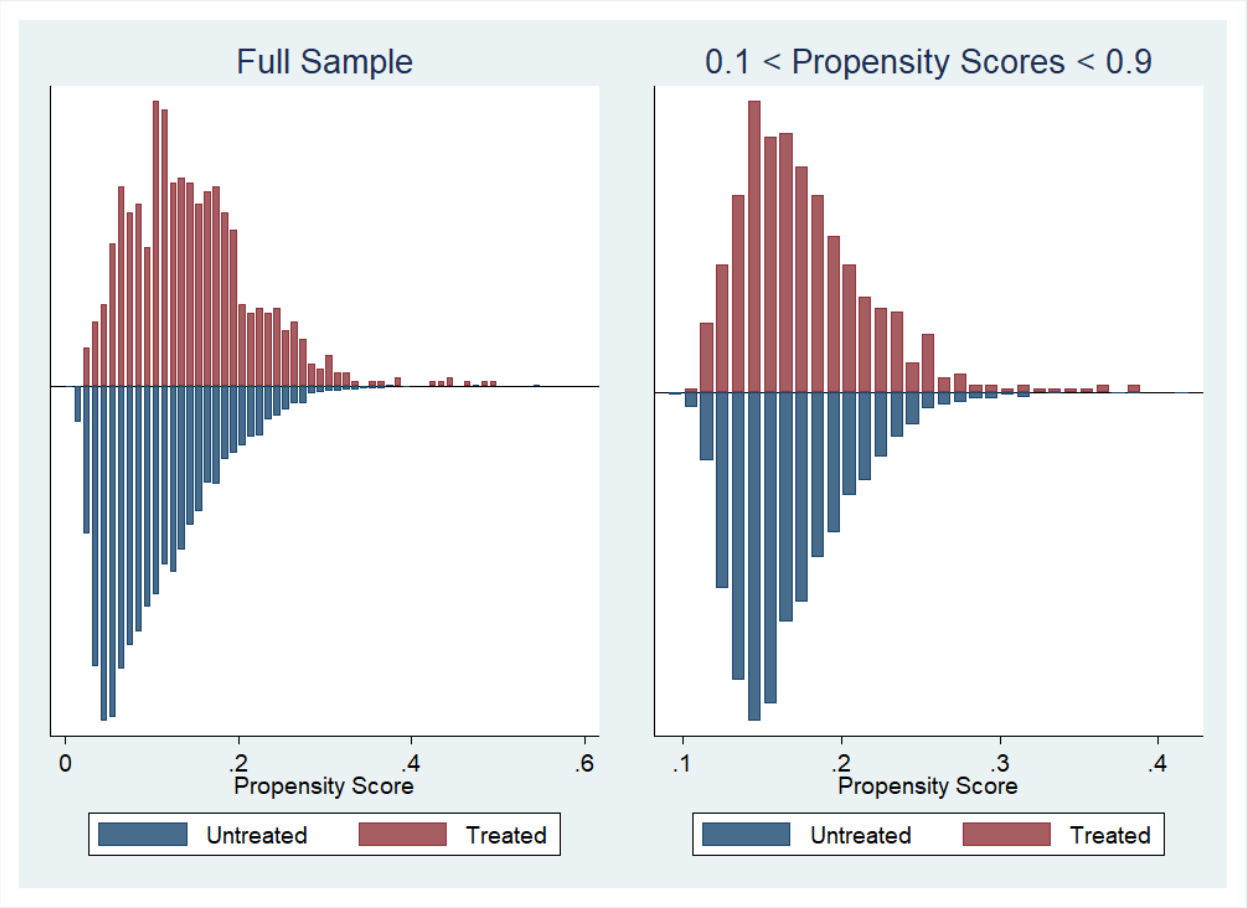


Figure notes: Propensity scores estimated using logit.

Table A1: Observable differences between treatment sample in different populations

	(1) FYLC	(2) FYLC	(3) 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.
*** p<0.01, ** p<0.05, * p<0.1

Table A2: 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 A3: 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.

Table A4: Testing for selection into the lottery, of compliers, and into attrition with interactions

	(1)	(2)	(3)	(4)	(5)	(6)
Won lottery	0.009	0.003			0.009	0.002
	(0.025)	(0.025)			(0.025)	(0.025)
Entered lottery	0.019	0.025	0.038	0.026	0.035	0.026
x FYLC	(0.029)	(0.030)	(0.036)	(0.035)	(0.044)	(0.044)
Won lottery	-0.003	-0.002			-0.003	-0.001
x FYLC	(0.038)	(0.038)			(0.038)	(0.038)
Entered			0.028**	0.038***	0.027**	0.038***
Lottery			(0.011)	(0.011)	(0.013)	(0.013)
FYLC			-0.016	-0.002	-0.016	-0.002
(no lottery)			(0.033)	(0.032)	(0.033)	(0.032)
Observations	1565	1565	8131	8131	8131	8131
Controls	No	Yes	No	Yes	No	Yes
Sample	Lottery	Lottery	Full	Full	Full	Full

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. Columns (3) - (6) present the decomposition of the results from Table 5. The omitted category for these regressions is composed of those never-ever-takers who do not enter the lottery and do not enter the FYLC.

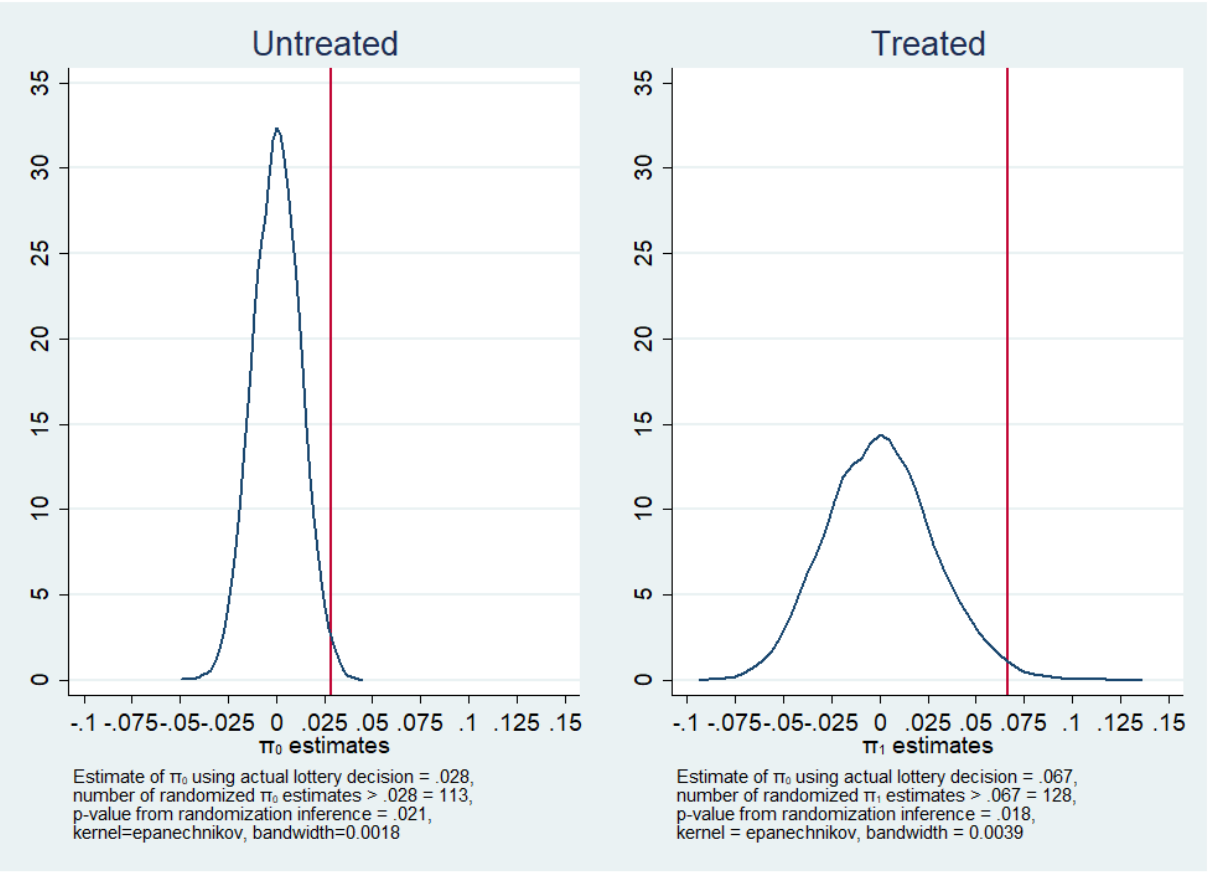
As a robustness exercise, we couple this nonparametric analysis of selection into the experiment with nonparametric randomization inference. In the spirit of Fisher (1935), we test the sharp hypothesis that there are no differences in outcomes between those who enter and those do not enter the experiment within each treatment status. We do this using two different approaches following Young (2018). In each of 10,000 repetitions, we randomly assign each individual with 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 B1 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 B23.

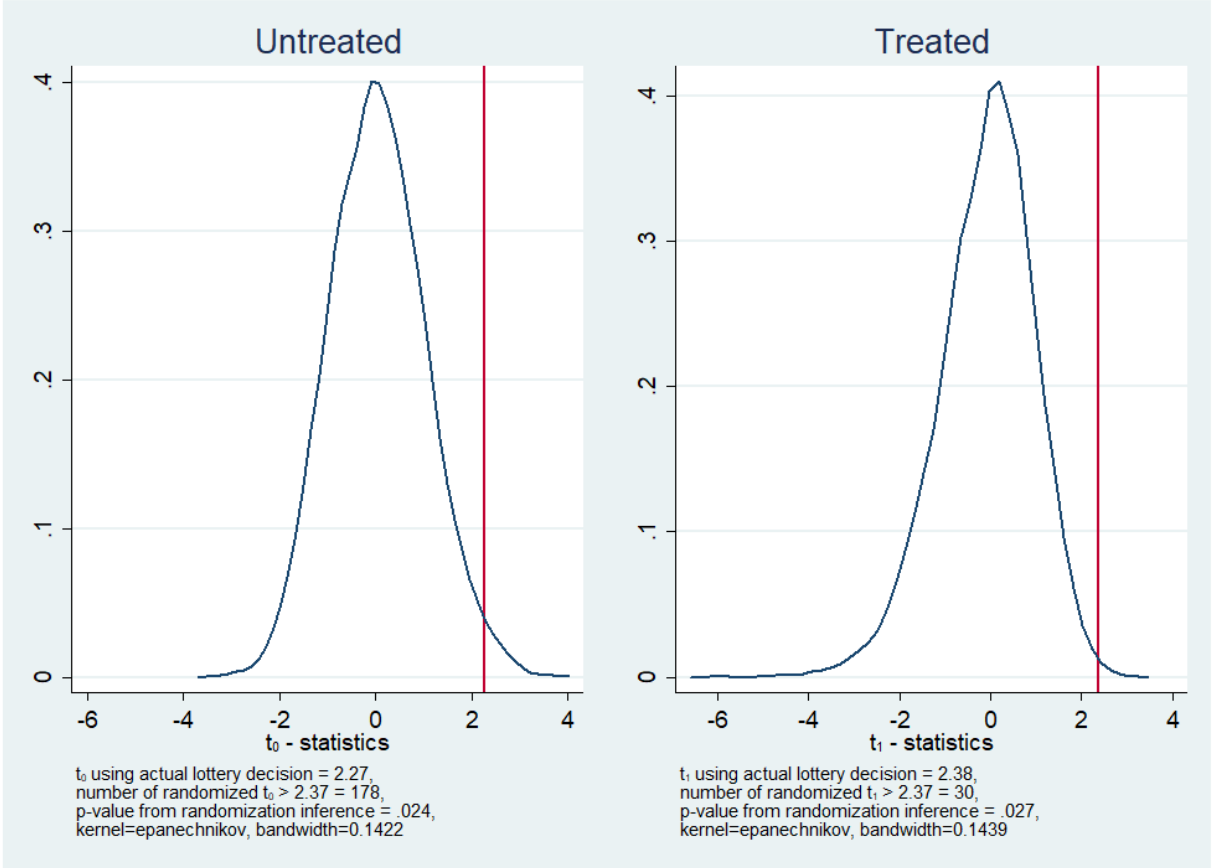
In each case the red line lies on the far-right side, indicating that the realized differences in retention between those who actually do and do not participate in the experiment are unlikely to result from pure chance. As described above, the p-values from our randomization tests correspond closely 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 B3 and B4 and Table B1 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 B1: Distribution of placebo $\hat{\pi}$ where "lottery participation" is randomly assigned



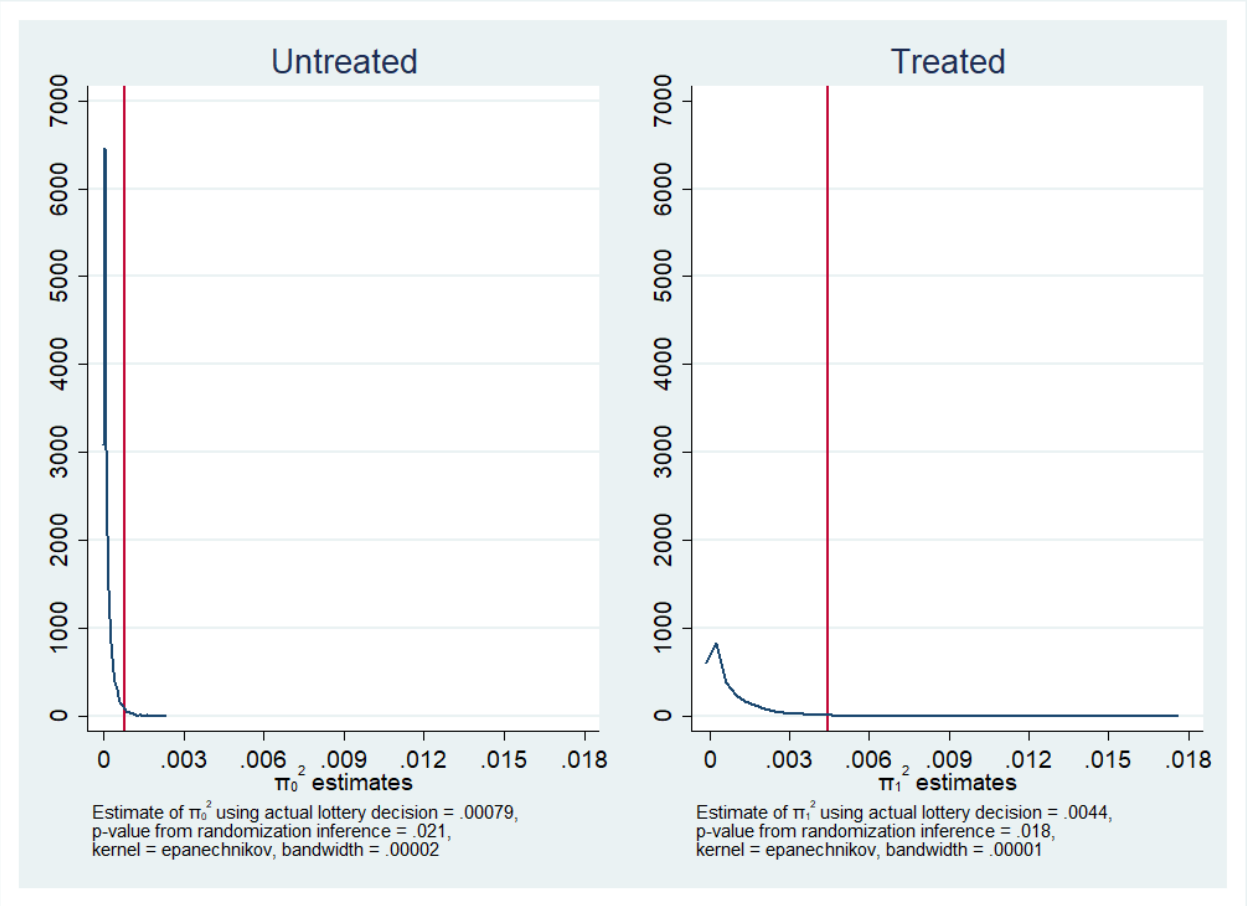
Notes: Binomial random assignment to lottery participation with probabilities of inclusion in the lottery by treatment status set at 0.11 for the untreated and 0.87 for the treated reflecting the shares observed in the data. Distributions constructed from 10,000 repetitions. The red verticle lines denote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

Figure B2: Distribution of placebo t-statistics where "lottery participation" is randomly assigned



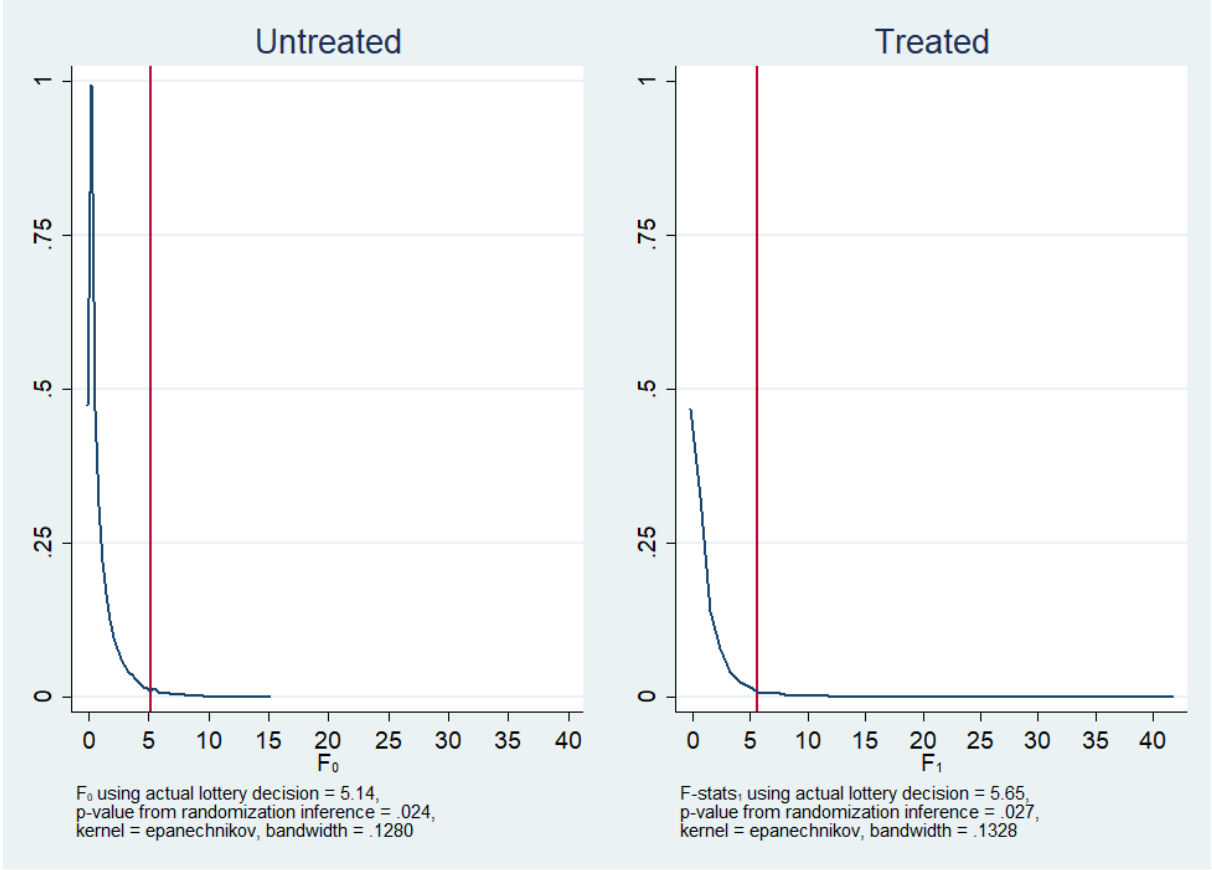
Notes: Binomial random assignment to lottery participation with probabilities of inclusion in the lottery by treatment status set at 0.11 for the untreated and 0.87 for the treated reflecting the shares observed in the data. Distributions constructed from 10,000 repetitions. The red verticle lines denote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

Figure B3: Distribution of squared placebo coefficients



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 B4: 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 B1: 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.