# 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

July, 2020

Abstract:
Voluntary selection into experimental samples is ubiquitous and may lead 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. Intent-to-treat and local-average-treatment-effect estimates reveal no discernable programmatic effects. Our tests reveal that the experimental sample is positively selected on unobserved characteristics suggesting limited external validity. Finally, we compare observational and experimental estimates, considering the internal and external validity of both approaches to reflect on within-study comparisons themselves. Keywords: Post-secondary education, college attrition, experimental design, generalizability. Classification codes: C18, C90, I23.

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

This paper explores 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 college student retention. We offer “intent to treat” (ITT) estimates of the effect of being randomized into treatment. 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 from the self-selected population 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 (Imbens and Angrist, 1994). 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.

We next turn to the generalizability of our estimates to populations of interest broader than those who compose the experimental sample.[1] First, a small share of the students who received treatment 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, utilizing an opt-out design. The external validity of our RCT findings is thus a matter of some importance. Do the results observed in the experiment apply to those who did not volunteer to participate in the program?

External validity may be a concern to researchers in many experimental contexts. In some instances, those who select into an experiment are indeed the exact population of interest. However, in many instances the population of interest is much broader.[2] The composition of the population of interest depends on the question and audience. Policy makers may be exclusively interested in the effect of the policy on those who currently select into it. However, if the selection criteria change or the program is expanded, the effects of the program on these new populations will be of primary interest. Additionally, researchers often utilize RCTs to answer more general questions, beyond the terrain of strict program evaluation. Here, the population of interest is also typically broader, and selection into the study may not be random.

---

[1] Non-compliance with randomization may present its own complications for generalizability. We utilize existing tests from Huber (2013) and Black et al. (2017) to examine this margin, though in our context, we believe selection into the experimental sample to be more significant.

[2] As Deaton and Cartwright (2017) write "the [experimental] 'sample' might be, but rarely is, a random sample from some population of interest."

Moreover, non-representativeness of experimental samples may originate from researchers selecting participants as in Allcott (2015) or from participants' decisions about whether to participate. Accordingly, participant consent requirements may make participant-based nonrandom selection more common. Self-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 the 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 and in responsiveness to treatment may exist and present more persistent problems for estimation.

Accordingly, we propose direct and straightforward tests for selection into the experiment on the basis of unobserved characteristics and heterogeneous responsiveness to treatment.[3] The first compares first year retention rates for those who do not receive treatment, according to whether they participate in the experiment. The second does the same among the treated populations comparing those who receive treatment through the experiment to those who receive treatment without participating in the randomization. When both treated and non-treated individuals are present outside the experiment, differences may be due to selection into treatment outside the experiment or to selection into the treatment itself. To our knowledge, we are the first to provide conditions under which differences in outcomes of these two tests indicate selection into the experiment on unobserved variables and thereby contradict claims of external validity. Most importantly, our monotonicity of potential outcomes assumption holds that on average, sub-populations are either positively or negatively selected on both potential outcomes. This

[3] 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).

assumption seems likely to hold in many instances. Though such a framework is not necessary, it is theoretically justified by dynamic complementarities as presented in Cunha and Heckman (2007).

We then implement the proposed tests for nonrandom selection in our particular RCT setting in order to examine the generalizability of the experimental results to the broader population of interest. We possess information on the non-experimental population as well as those who selected into the experiment. We utilize this information to explore the extent of otherwise unobserved differences between the experimental sample and the broader populations of interest within both treated and untreated populations. As our setting also includes experimental participants who do not comply with their randomized treatment assignment, we first check the ignorability of this noncompliance by following similar tests proposed by Huber (2013) and Black et al. (2017).

While employing these existing tests reveals little evidence of selection into compliance with the randomization within the experiment, we find significant selection into the experiment itself using the proposed methods. 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, our tests reveal that the experimental population possesses differential 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 voluntarily enroll in the experiment. This positive selection on unobserved variables holds for both those who do and do not receive treatment, pointing to selection into the experiment (as opposed to selection into treatment in the observational setting). The magnitude of the selection into the RCT clearly raises concerns regarding the generalizability of the RCT results to these larger populations of interest.

Whether the purpose of the research is to test hypotheses derived from theory, explain

empirical regularities, or evaluate or inform policy making, it is often desirable to extend the causal estimates beyond the population for which the estimates directly apply.[4] Indeed, researchers often utilize RCTs to answer general or theoretical questions which typically pertain to a broader population than the RCT sample itself. For instance, does neighborhood quality affect residents' educational, health, or employment outcomes (as in Ludwig et al., 2008)? "Do workers work more if wages are high" (Fehr and Goette, 2007)? Does early education lead to future educational and economic success (as in Weikart, 1998)? Thus, our tests for external validity may benefit researchers by providing an approach to present concrete evidence of the representativeness of their experiments.

A final contribution of the paper addresses the literature of within-study comparisons launched by LaLonde's (1986) seminal work. In this literature, it is common to use the results from RCTs to benchmark observational approaches. As our tests for external validity implicitly compare the results between an RCT and observationally implemented OLS, we reflect on what we learn from such a comparison considering the populations of interest, the internal validity of observational estimates, and selection into the experiment. In cases where the population of interest is broader than the experimental sample, we conclude that such comparisons often ought not be used to argue for the superiority of one approach over another without further testing.

The paper is organized as follows: First, we review the literature on the central contributions of the paper. Second, we describe in greater detail the learning community program at this institution, the nature of the randomized control trial design, and the data to be used in the analysis. Third, we describe the empirical methodology, followed by the results. Finally, we offer a summary discussion and conclusion.

**Background and Data**

The First Year Learning Community (FYLC) we study began on a small scale and

[4] 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.

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).[5]

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.

While there are many published evaluations of first-year experience programs broadly defined and still more conducted by in-house researchers, the set of studies focused specifically on first-year learning communities is more restricted.[6] Moreover, while some observational studies employ methods beyond simple linear probability models using OLS – for example,

[5] 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.

[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.

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) – in all of these the exogeneity assumptions necessary for causal interpretation of the results are onerous.

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

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

Data for this analysis come from student records on the two freshman cohorts during the years for which the program capacity was increased by virtue of the federal grant. A unique

feature of our analysis is that in addition to retention and demographic information for the self-selected population who applied to be part of the program, we also gather information on the remainder of the freshman class who at the outset expressed no interest in program participation. Having information on the non-experimental population is unfortunately rare in RCT designs. We use this additional information to shed light on the nature of various selection issues which are impossible to explore without it.

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.[7]

As mentioned above, there are three important instances of migration between assigned

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

groups in the data.[8] 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.

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 | 3.46 | 3.53 | -0.07*** |
| | | | (0.02) | | | (0.01) |
| SAT math | 494.25 | 498.65 | 4.40 | 496.57 | 544.40 | -47.83*** |
| | | | (6.15) | | | (3.63) |
| SAT writing | 491.42 | 496.40 | 4.98 | 494.04 | 508.33 | -14.29*** |
| | | | (5.77) | | | (3.29) |
| SAT verbal | 488.00 | 491.14 | 3.14 | 489.65 | 502.39 | -12.73*** |
| | | | (5.88) | | | (3.31) |
| Female | 0.68 | 0.69 | 0.01 | 0.69 | 0.50 | 0.19*** |
| | | | (0.02) | | | (0.01) |
| 1st generation | 0.63 | 0.62 | -0.01 | 0.62 | 0.56 | 0.07*** |
| | | | (0.02) | | | (0.01) |
| Low income | 0.60 | 0.62 | 0.01 | 0.61 | 0.56 | 0.05*** |
| | | | (0.02) | | | (0.01) |
| Lives on Campus | 0.74 | 0.75 | 0.01 | 0.75 | 0.71 | 0.04*** |
| | | | (0.02) | | | (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.

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 many observable dimensions. 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. Furthermore, even without the subsequent expansion of the program,

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

these separate populations highlight the importance of analyzing the generalizability of the LATE to a broader population of interest.

**Empirical Methodology**

We divide our empirical analysis into two sections. First, we utilize the RCT design to identify the intent to treat effect of the FYLC on first-year retention, as well as the local average treatment effect. Second, we compare untreated individuals' outcomes by whether or not they participated in the experiment. We do the same among treated individuals. We provide novel conditions by which differences imply nonrandom selection on unobserved variables into the experiment and external invalidity of the experimental results. Because of the non-compliance with randomization *within* the experiment we also examine evidence of selection into the compliers for whom the RCT estimates apply. More detail about each set of analyses is given below.

***Analysis 1: Estimating treatment effects using the RCT***

For ease of explanation we present a linear correlated random coefficient model of the outcome Y (first-year retention) as a function of the treatment D (FYLC participation), and a vector of observed variables, $\boldsymbol{X}$, comprised of the variables listed in Table 1.

$$Y_i = D_i(\beta + e_i) + \boldsymbol{X_i\gamma} + \varepsilon_i, \quad (1)$$

where $\beta$ is the average treatment effect, $e_i$ represents individual treatment effect heterogeneity, and $\varepsilon_i$ represents an idiosyncratic error. We will later relax this linear parametric model.

Randomization among the experimental group provides two groups of similar size; those assigned to treatment and those assigned to the control group, which should be in expectation identical with respect to both observed and unobserved pre-determined characteristics. Let $Z_i$=1 indicate that the lottery assigned individual *i* to participate in the FYLC. Accordingly, we begin by estimating a linear probability model using OLS among the population who selected into the lottery according to the following specification:

$$Y_i = Z_i\beta_{ITT} + \boldsymbol{X_i\gamma} + \epsilon_i. \quad (2)$$

As causal identification does not hinge on the covariates we conduct the analysis both with and without conditioning on $\boldsymbol{X}$.[9] 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).

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.[10] 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 non-experimental population. We take up these issues in Analysis 2.

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

In this section, we provide tests of selection of the experimental sample on unobserved variables when the non-experimental setting includes both treated and untreated individuals. To our knowledge, we are the first to provide conditions under which the results of these tests indicate nonrandom selection into the experiment and external invalidity of the experiment (as opposed to selection into treatment among the non-experimental population). The possibility of non-random

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

[10] 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.

sample selection of RCTs has received earlier attention.[11] Cole and Stuart (2010) and Andrews and Oster (2018) test for selection based on observed variables. They then model the decision to participate in a study based only on observed characteristics and approximate weights based on these approximations to provide point estimates or (in Andrews and Oster, 2018) estimate bounds on the population average treatment effect. However, we worry that observed data will not fully capture the non-random selection. Thus, 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.

We add our voice to a small but growing literature spanning statistics, biomedical research, political science, and economics, which uses outcome data from observational settings to examine selection into experiments.[12] We discuss herein those with the closest aims or approaches to those of this paper. 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. She 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 (2017) attributes any differences in unobserved characteristics as being the result of the randomization itself, leading to internal invalidity of the RCT. While this work is insightful in showing the use of data from the broader population, 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) and Altidag, Joyce, and Reeder (2015), use a similar test as a placebo test for external validity.

Bartlett et al. (2005), Prentice et. al (2005), Altidag, Joyce, and Reeder (2015), and

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

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

Hartman et al. (2015) each compare outcomes among the treated across experimental and observational settings in order to evaluate the generalizability of experiments. However, unlike Sianesi (2017), in each of these studies treatment status varies in the observational environment. Consequently, none of these studies is able to show whether significant differences in outcomes across observational and experimental settings arise from selection into treatment in the observational setting or selection into the experiment. Hartman et al. (2015) is unique among this literature in providing a theoretical framework explaining the use of these exercises in isolation. They propose a test comparing the outcomes (adjusted for observed covariates) of those who receive treatment within the randomized sample to those who receive treatment otherwise. As with the other three studies, this exercise incorporates heterogeneous responsiveness to treatment into the tests. However, their "strong ignorability of treatment assignment" may fail from selection on unobserved variables either into the experiment or into treatment in the observational setting.[13]

We show that it is only by using both tests on the treated and the untreated populations that we are able to clarify whether there is nonrandom selection on unobserved variables into the experiment and thus whether generalizability of the RCT findings is compromised. While each test may be suggestive on its own, we provide the conditions under which both tests combined provide insight into the representativeness and generalizability of the experiment.

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 the tests. We then adopt the potential outcomes framework from Rubin (1974) in a nonparametric examination of our tests. The nonparametric approach allows us to precisely state what information such tests reveal, and to provide the conditions by which we may interpret the results as pertaining to selection into the experiment on unobserved variables and the external invalidity of the

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

experimental results. In both approaches we will focus first on the untreated, second on the treated, and lastly why both are necessary in conjunction.

Throughout, we will maintain proper randomization within the experiment and monotonicity as described in Imbens and Angrist (1994). As is typical we also maintain 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 last 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. We will require two other assumptions, which we describe in further detail below using the potential outcomes framework.

For the time being we hold that compliance is perfect such that Z and D are in perfect agreement within the experiment. We discuss complications from noncompliance at the end of this section. We add to this familiar framework, L, as an indicator for participation in the experiment. Much of the earlier treatment effects literature 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 "non-randomized-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.

We start with a simple t-test on the estimated coefficient on $L$ in the regression of $Y$ on $\boldsymbol{X}$ and $L$ with this restricted sample:

$$E(Y_i|D_i = 0, L, \boldsymbol{X}) = L_i\pi_0 + \boldsymbol{X_i\gamma_0}. \quad (3)$$

If the treatment status is homogeneous among the non-experimental population, then performing a standard t-test comparing $\widehat{\pi_0}$ directly tests whether there is selection into the experiment on unobserved variables with no further assumptions. This is comparable to the Sianesi (2017) approach. While it may be reasonable to infer that difference in unobserved outcome levels suggest

a lack of external validity (as in Galliani, McEwan, and Quistorff, 2017), differences in treatment effects do not directly enter the tests as neither group received treatment. Adjusting for covariates with the inclusion of the vector $\boldsymbol{X}$ is not necessary and may introduce some finite sample bias (see Lin, 2013), but it focuses attention on unobserved selection and brings our approach in line with Prentice et al. (2005) and Hartman et al. (2015). Further, the sign and magnitude of $\widehat{\pi_0}$ demonstrates the extent and direction of the selection bias. If the observational setting contains both treated and untreated individuals, this difference may arise from selection into the experiment or selection out of treatment in the observational setting.

Assuming that some who did not enter the lottery nonetheless receive treatment, we 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 those who do not enter the experiment differ from those who do. Here, the tests build in differences on the basis of unobserved characteristics and heterogeneous effects. While significant differences indicate selection, again in isolation it may be due to selection into treatment in the observational setting or selection into the experiment.

However, if we observe selection that is the same direction into the experiment among both the treated and untreated populations, we maintain that together the two tests indicate selection into the experiment itself. The intuition here is that those who select into treatment in the observational setting cannot simultaneously be negatively and positively selected.[14]

Such an examination is demanding on the data, as examining selection on unobserved variables into RCTs requires outcome data from a representative sample of experimental-nonparticipants. In our application we benefit from access to the contemporaneous universe of freshmen at the institution. However, individuals who receive treatment and are not randomized may not be present in all settings. Indeed, in some settings the RCT sample is the population of interest and these tests do not apply, so long as the selection processes and context remain constant.

[14] This intuition requires the monotonic selection on potential outcomes assumption described in detail below.

However, there are many examples where populations of interest are broader and where these non-experimental-takers are present.[15] These data may be present in any randomized evaluation of an existing program and instances where treatment is adopted following a randomized trial. While they are absent from many existing experimental designs, this could be because the usefulness to experimental design of a representative sample of treated individuals is not generally understood. We highlight the use of such data and remark that samples used for analysis are chosen by researchers through their research designs. We clarify what we learn from such tests in the section below.

*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 with Y as 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. 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 for simplicity and because we are most interested in responsiveness to treatment rather than selection on unobserved covariates.

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

Even under perfect compliance with the randomization (which we continue to maintain for the moment), testing the generalizability of RCT results requires data from the non-

[15] The compliers who were moved by housing demolitions in Jacobs (2004) and Chyn (2018) are essentially non-randomized-takers to the Moving-to-Opportunity compliers from Goering et al. (1999) and Chetty et al. (2016). Non-randomized-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. They are common also in medicine as medications are adopted after a randomized drug trial.

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. One reasonable candidate may be what we term 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 $E(Y_1|L=0,D=1) \leq E(Y_1|L=0,D=0)$.[16]

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 selection on responsiveness to treatment. This assumption does rule out instances where among the non-experimental population, differences in responsiveness to treatment among the treated and untreated are larger in magnitude and opposite signed as the differences in unobserved levels of the outcome. 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 with the responsiveness to treatment. It 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. It does not need to hold for each individual, but only in expectation. Furthermore, positive correlation between before treatment potential levels of the outcome and treatment effects from the human capital investment is theoretically justified under a dynamic complementarity model as described in Cunha and Heckman (2007).

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

*Selection on unobserved variables and external invalidity*

Assuming weakly monotonic selection into treatment in the non-experimental 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:

$$E(Y|L=1,D=0)-E(Y|L=0,D=0)= \quad (4)$$
$$(1-P)^{-1}\{E(Y_0|L=1)-E(Y_0|L=0)+P[E(Y_0|L=0,D=1)-E(Y_0|L=1)]\},$$
$$E(Y|L=1,D=1)-E(Y|L=0,D=1)=$$
$$P^{-1}\{E(Y_1|L=1)-E(Y_1|L=0)+(1-P)[E(Y_1|L=0,D=0)-E(Y_1|L=1)]\}. \quad (5)$$

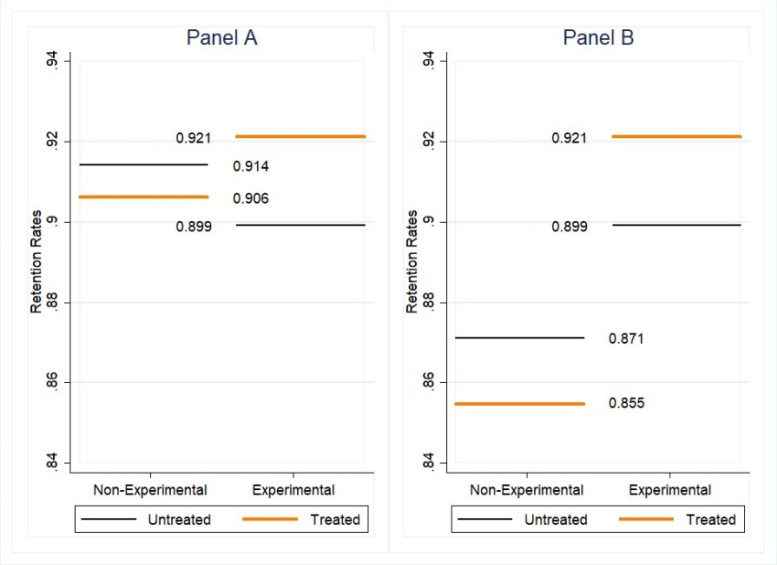
Equation (4) compares the expected outcomes among those who did not receive treatment by whether they participated in the lottery. The first difference on the right-hand side of equation (4) directly examines whether there is selection into the lottery on the basis of potential outcome in the absence of treatment. The latter difference could be nonzero either from selection into the lottery or selection into treatment in the non-randomized population.

Equation (5) 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 is selection into the experiment.

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 as opposed to 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, 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 (4), 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 (5) 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 non-experimental populations in the outcome, and 2) differences that are consistent in sign between the treated and untreated groups.

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 in the observational setting.

*Assumptions for isolating differences in treatment effects*

Turning to the representativeness of the experimental results, we first may adopt the ancillary assumptions from Kowalski's (2016, 2018) examination of external validity *within* experimental samples. If we assume the potential outcomes are monotonically related to the probability of treatment, we can interpret these same results as indicative of external invalidity, given that probability of treatment rises with participation in the experiment. As discussed in detail in Kowalski (2016, 2018) and Brinch et al. (2017), such an assumption follows by maintaining a Roy model of economic decision making. However, this assumption is strong as it pertains to the entire range of propensities for treatment and leads us to reject external validity even when the differences in treated and untreated average outcomes is identical between experimental participants.

Secondly, we may use a possibly weaker assumption, that selection on levels into the experiment on average is common among the treated and untreated. In which case, the difference between equations (5) and (4) is indicative of differences in treatment effects. We can recover this difference from the following regression equation:

$$Y_i = L_i\pi_{10} + D_i\pi_{01} + L_i \times D_i\pi_{11} + e_i. \quad (6)$$

While the magnitude also matters, under this assumption, $\widehat{\pi_{10}}$ and $\widehat{\pi_{11}}$ having the same sign and a standard t-test on $\widehat{\pi_{11}}$ provides evidence of the external invalidity of experimental estimates.

*Noncompliance within 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 observe whether it holds, we assess its plausibility by conducting the tests previously described in the existing literature (Huber, 2013; Black et al., 2017).

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 adjusted and non-adjusted regressions of retention on randomized assignment using only the untreated lottery participants. 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. Here, we need not worry about selection into assignment as proper randomization insures that it is exogenous. 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 incorporates differences in 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* |
| *Non-randomized-takers* | $E(Y_i\|\boldsymbol{x}, D=1, L=0, Z=0) = \boldsymbol{x\gamma} + \beta + \bar{e}_l + \bar{\varepsilon}_l$ | *Non-randomized-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.

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 tests for external validity if the differences in outcomes within the experiment by treatment status near zero.

Yet another way we 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.

**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 ITT estimates. Were we to take these point estimates literally and maintain the stable unit treatment value assumption, expanding the program to all compliers would cut first-year attrition at the institution by a mere 3%. As the compliers compose 12.4% of the freshman class, if we could generalize these point estimates to the 90% of the freshman class who eventually received treatment this would be a substantial 24% reduction in attrition. However, once again these estimates are imprecisely estimated and we cannot reject zero overall treatment effects, even for the compliers.

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 + Non-randomized-takers | Treated + Crossovers + Non-randomized-takers | Crossovers + Non-randomized-takers | Crossovers + Non-randomized-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 non-randomized-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 non-randomized-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 non-randomized-takers.[17]

In columns (3) and (4), we compare two final treated groups – the crossovers and non-randomized-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

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

groups is 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 problematic selection into the population of 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 $E[Y_1|D=0,L=0] \geq E[Y_1| L=1]$.

We present the results from a differences in differences analysis in Table 5, which serve to summarize the information above and provide statistical inference regarding whether the selection into the experiment differs between those who did and did not receive treatment. Under the alternate assumption that that on average selection on levels into the experiment is common among the treated and untreated, this latter difference in selection indicates whether the treatment effects revealed in the experiment is external valid to the rest of the freshman class. The coefficient on the interaction between Lottery and FYLC in columns (3) and (4) provide this differences-in-differences estimate ignoring noncompliance. The point estimate reveals a 0.026 to 0.038 larger difference between the experimental and non-experimental retention rates among the treated than is the same difference among the untreated. While this large point estimate suggests a lack of external validity, it is not statistically significant at conventional levels. The same holds when we add indicators to accommodate noncompliance within the experiment in

columns (5) and (6).

The coefficient on FYLC provides the difference in outcomes between the treated and untreated in the non-experimental portion of the population, mimicking an observational analysis in the absence of randomization on these individuals. Point estimates imply zero to slight negative effects of treatment though in no case are they statistically significant. Naturally with these estimates, we worry that selection on unobserved variables into treatment may bias the estimated effects. However, we observe from the previous analysis that there the RCT results do not directly apply to this population. Also, there is significant positive selection into the experiment both with and without treatment. Under the assumption that potential outcomes are monotonic with respect to the probability of treatment, this selection implies a lack of external validity. Given this selection into the experiment, it may be difficult to argue that the RCT estimates provide more accurate treatment effects for this non-experimental population.

Table 5: 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.

***Comments regarding within-study comparisons in light of selection into RCTs***

In seeking to understand selection into the experiment and the generalizability of RCT results to a broader population, we have generated observational results as well as those from an experiment. In light of our tests for selection into the experiment on unobserved variables, we ask what can be learned from a comparison of the observational and experimental results about competing research designs in the "within-study design" tradition following LaLonde's (1986) seminal work? In LaLonde (1986) and the many that have come after, results from observational approaches -- such as OLS, propensity score matching (PSM), and regression discontinuity -- are benchmarked against experimental results. Reviews of these studies in Glazerman, Levy, and Myers (2003) Bloom, Michalopolis, and Hill (2005), and Cook, Shadish, and Wong (2008) have found that in most cases there are significant differences between experimental and observational results, and these differences have served to elevate RCT relative to observational approaches.

Cook, Shadish, and Wong (2008) comment that there is heterogeneity among within-study comparisons and give seven criteria for evaluating such comparisons. One criterion is that "The experiment and observational study should estimate the same causal quantity," such as ITTs or ATEs. We add that self-selection into experiments alters the causal quantity being estimated. This is because a program may have different effects on the type of people who sign up for the experiment than it does on people who do not. If the population of interest is the experimental participants, the RCT provides a valid estimate of the policy relevant parameter, worthy of use for benchmarking. In that case, considering individuals from an observational setting outside the population of interest alters the estimand of the observational approach. It would be unsurprising were the observational approach to fail to estimate a parameter other than its estimand. On the other hand, if the population of interest is broader than the RCT participants, without further testing we do not know whether the differences in estimates arise from internal validity failings of the observational approach or external validity failings of the RCT.

To illustrate this point, let us use the OLS estimates on the entire freshman class (which we hold here is the population of interest) for comparison against the RCT estimates. The results

from columns 3 through 6 of Table 5 may be interpreted as weighted decompositions of such OLS regression estimates in which retention is regressed upon treatment, ignoring the randomization. We acknowledge that this design is not an exemplar of observational approaches, but use it for ease of explanation.

Note that $plim\widehat{\beta_{OLS}} = ATE + Bias_{OLS}$ and $plim\widehat{\beta_{RCT}} = LATE$, assuming proper randomization. The local aspect of this LATE pertains to participation in the randomization, but it becomes more localized with noncompliance. The upshot here is that the two estimators may converge to different parameters, even if the controls included in OLS are sufficient to capture selection into treatment. We can relate the two parameters according to the following, maintaining perfect compliance with randomization within the experiment:[18]

$$ATE = LATE \times P(L = 1) + E(b_i|L = 0) \times \big(1 - P(L = 1)\big), \quad (7)$$

where $E(b_i|L = 0)$ is the average effect among those who do not participate in the experiment. It is typical in this literature to suspect bias in the observational approach. In comparing the two estimates, we observe the following:

$$plim\widehat{\beta_{OLS}} - plim\widehat{\beta_{RCT}} = [1 - P(L = 1)][E(b_i|L = 0) - LATE] + Bias_{OLS}\,. \quad (8)$$

Equation (8) nicely demonstrates that the comparison in results provides a mixture of the possible external invalidity of the RCT ($E(e_i|compliers = 0) - LATE$) and the possible bias in the OLS estimate. Thus, differences in estimates alone cannot point to general failings of either approach.

Table 6 contains these cumulative effects as estimated by OLS, QMLE Logit, and propensity score matching (PSM). We note that by adding these additional observations from the non-experimental population the estimand mechanically changes, as it now pertains to a larger population. We perform all analyses both on the full sample as well as the sample in which estimated propensity scores are between 0.1 and 0.9, out of respect for the overlap assumption

[18] Allowing for noncompliance does not present significant complication.

and in accordance to the rule of thumb provided by Crump, et al. (2009).

Contrary to the findings from the RCT design, the Table 6 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 when we restrict attention to the observations in which the overlap is thick, the estimated effects are even larger with coefficient estimates all over 5 percentage points with p-values ranging from 0.03 to less than 0.001. However, we only know that the observational approach is biased in giving the ATE for the broad population because; 1) selection into treatment largely transpires through selection into the RCT, and 2) we know from the results in Table 4 that there is positive selection into the RCT on unobserved characteristics. In other words, the observational method fails due to selection into the RCT.

Table 6: 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.

If instead, we use for comparison the observational estimate on only those who did not enlist in the experiment, we have no such insight. The difference in point estimates may well be due to selection into the experiment on treatment effects or selection into treatment in the observational setting or sampling error. Thus, when populations of interest are broader than experimental samples, 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 clearly indicating the population of interest and applying tests for selection into the RCT to within-study comparisons, we can see whether the RCT should be used as a benchmark or whether the bias indeed lies in observational estimators or whether such comparisons are uninformative in the 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 non-experimental 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. We believe that researchers should reflect more on what constitutes the population or parameter of interest. In some instances, this is the exact population for which experimental results hold. However, when selection criteria change or researchers seek to answer general questions or extrapolate to other settings, the population of interest is much broader. In those instances, participation decision by researchers or subjects may compromise the generalizability of experimental results.

Our analysis reveals that students who selected into the study disproportionately possess *observed* background characteristics that are negatively associated with first-year retention. However, as we have shown, the unobserved characteristics of the experimental participants 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, 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 a potentially larger population of interest, or reserving resources to survey a random sample on the outcome, or ideally linking the RCT to administrative data encompassing that population.

Finally, 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. When the population of interest is broader than the experiment, the two approaches estimate different parameters. In which case, differences may indicate either bias within the observational approach or external invalidity of the experimental approach. As the tests that we have introduced may provide evidence for where these biases lie, within-study comparisons provide additional possible applications for these test.

## References

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

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

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

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

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

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

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

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.

Bloom, Howard S., Charles Michalopoulos, and Carolyn J. Hill. 2005. "Using Experiments to Assess Nonexperimental Comparison-Group Methods for Measuring Program Effects." In H. S. Bloom (Ed.), Learning more from social experiments (pp. 173-235). New York: Russell Sage Foundation.

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.

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

Cook, Thomas D., William R. Shadish, and Vivian C. Wong. 2008. "Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons." *Journal of Policy Analysis and Management* 27, no. 4: 724-750.

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.

Cunha, Flavio, and James Heckman. 2007. "The technology of skill formation." *American Economic Review* 97, no. 2: 31-47.

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.

Fehr, Ernst, and Lorenz Goette. 2007. "Do workers work more if wages are high? Evidence from a randomized field experiment." American Economic Review 97, no. 1: 298-317.

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.

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

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

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:* 25, no. 1: 75-96.

Glazerman, Steven, Dan M. Levy, and David Myers. 2003. "Nonexperimental versus experimental estimates of earnings impacts." *The Annals of the American Academy of Political and Social Science* 589, no. 1: 63-93.

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.

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

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

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

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.

Prentice, Ross L. , Robert Langer, Marcia L. Stefanick, Barbara V. Howard, Mary Pettinger, Garnet Anderson, David Barad, J. David Curb, Jane Kotchen, Lewis Kuller, Marian Limacher, Jean Wactawski-Wende, (2005). "Women's Health Initiative Investigators, Combined Postmenopausal Hormone Therapy and Cardiovascular Disease: Toward Resolving the Discrepancy between Observational Studies and the Women's Health Initiative Clinical Trial," *American Journal of Epidemiology*, Volume 162, Issue 5, 1 September, Pages 404–414, https://doi.org/10.1093/aje/kwi223

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.

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

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

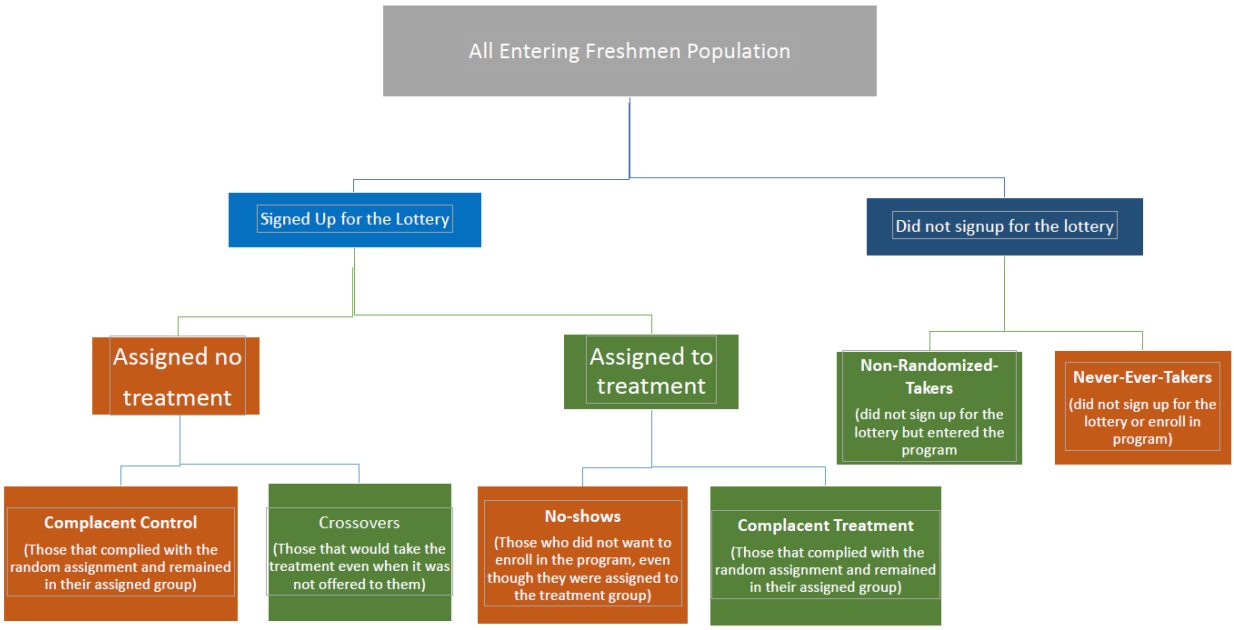


Figure A2: Overlap in the propensity scores by treatment status

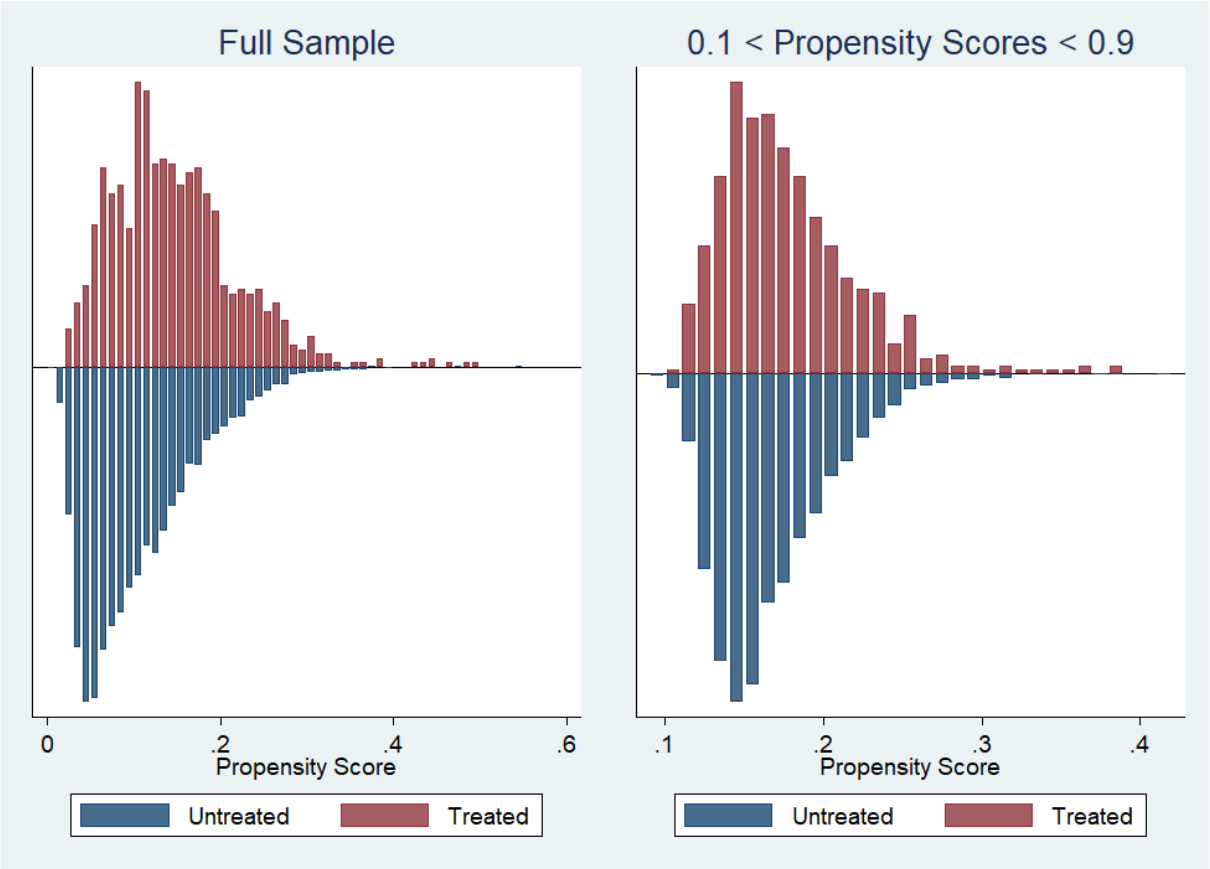



Figure notes: Propensity scores estimated using logit.

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.

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

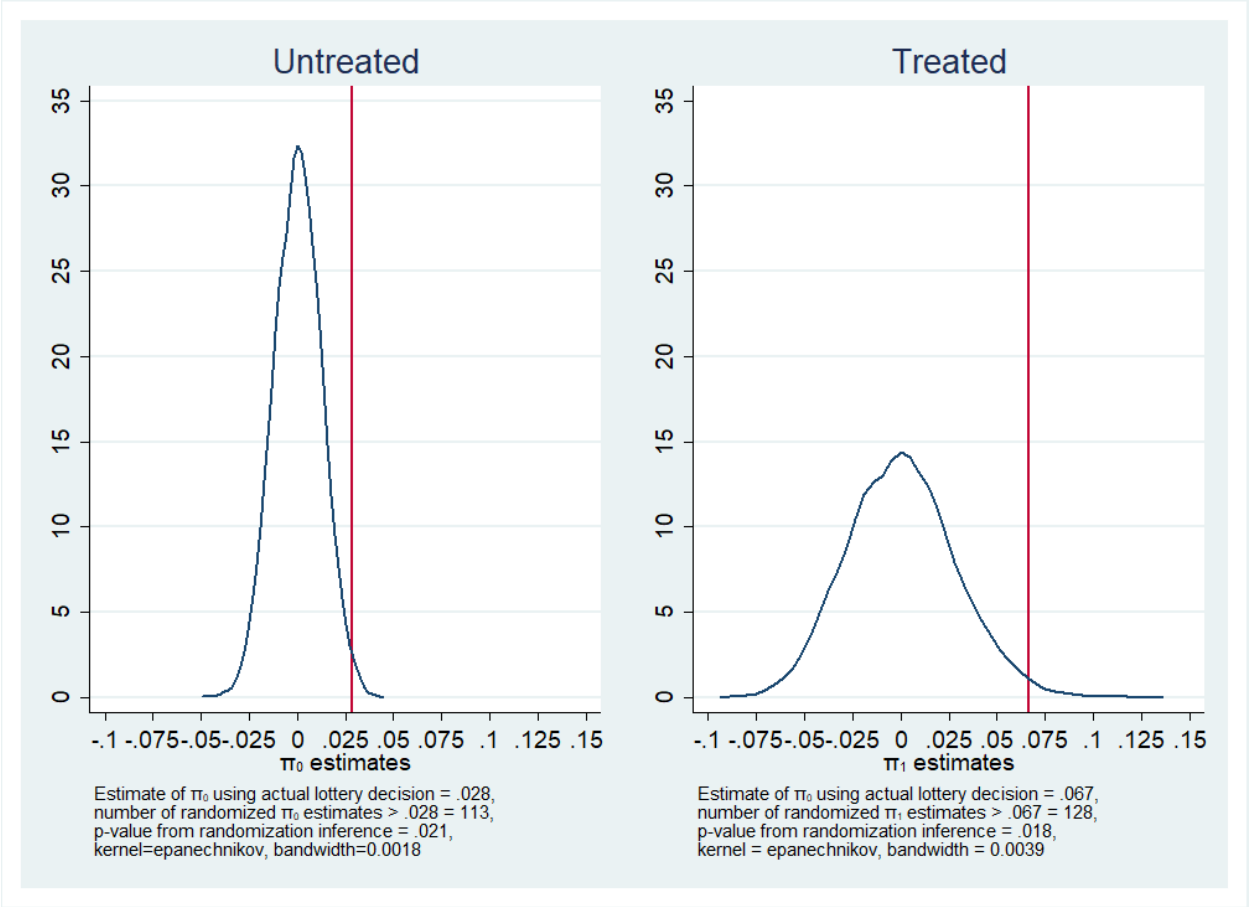
assigned groups. We then compare the differences in retention observed under the actual lottery participation decisions to the distribution of placebo differences we observe under random assignment of "lottery participation." The share of placebo differences whose magnitudes are more extreme than the 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

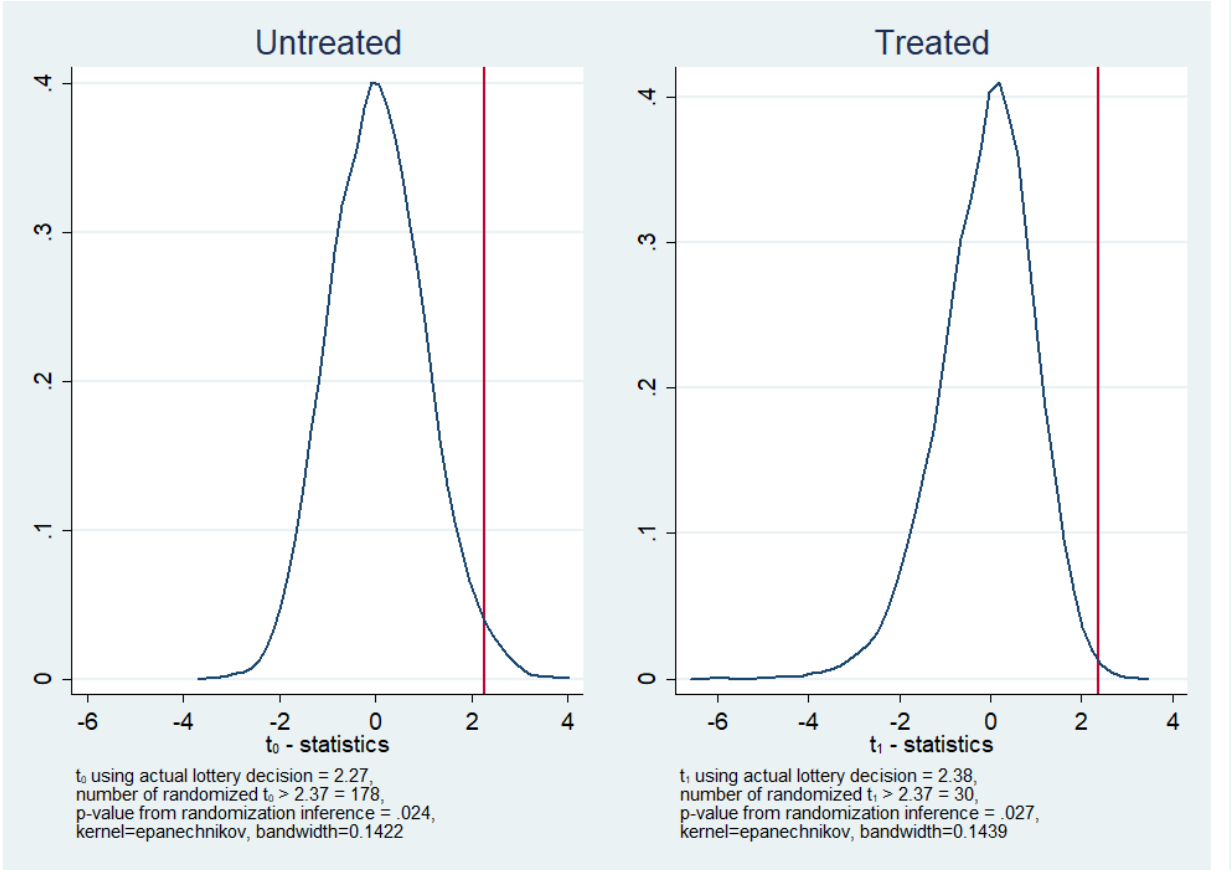
the placebo differences when lottery participation is randomly assigned. In each case, randomization inference provides similar or smaller p-values than those constructed from the Huber-White robust standard errors.

Figure B1: Distribution of placebo $\hat{\pi}$ where "lottery participation" is randomly assigned



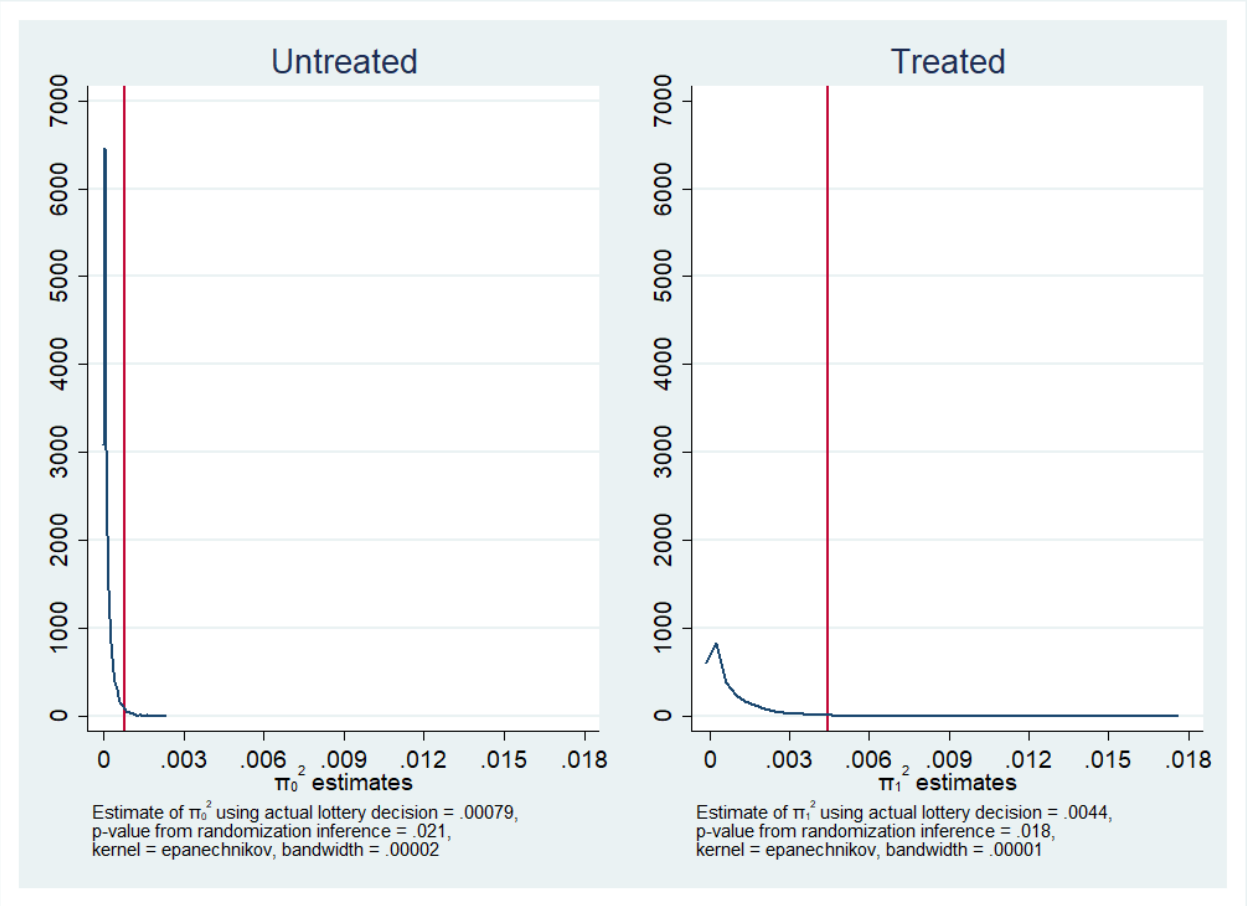

Notes: Binomial random assignment to lottery participation with probabilities of inclusion in the lottery by treatment status set at 0.11 for the untreated and 0.87 for the treated reflecting the shares observed in the data. Distributions constructed from 10,000 repetitions. The red verticle lines denote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



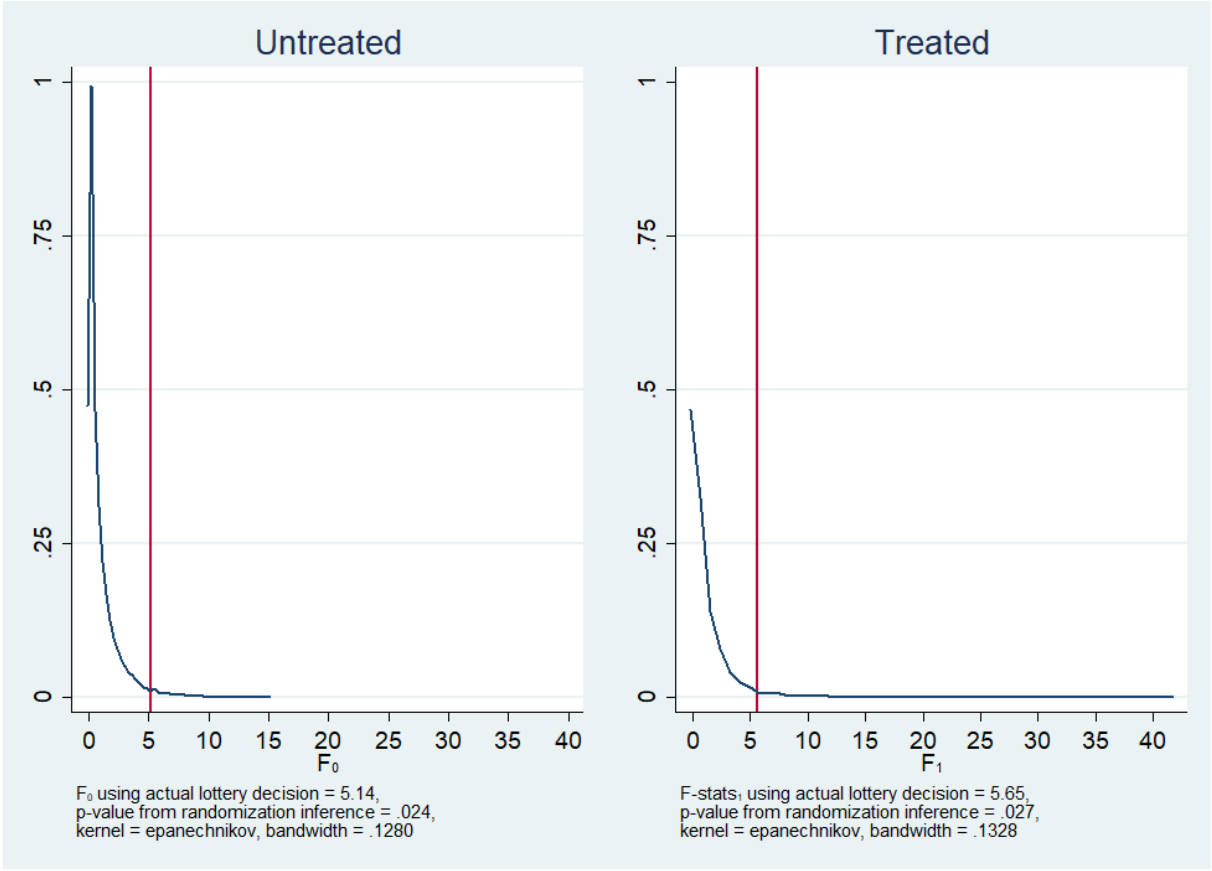

Notes: Binomial random assignment to lottery participation with probabilities of inclusion in the lottery by treatment status set at 0.11 for the untreated and 0.87 for the treated reflecting the shares observed in the data. Distributions constructed from 10,000 repetitions. The red verticle lines denote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