Do Learning Communities Increase First Year College Retention?

Testing the External Validity of Randomized Control Trials

Tarek Azzam, Michael D. Bates, and David Fairris

Abstract:

In this paper, we (1) utilize a randomized control trial (RCT) to estimate the impact of a learning community on first year college retention; (2) introduce simple tests for the external validity of RCT results, and apply these tests to our data; and (3) compare observational estimates to those from the RCT, considering the internal and external validity of both approaches. Intent-to-treat and local average treatment estimates reveal no discernable programmatic effects, whereas observational estimates are positive and statistically significant. The experimental sample, though negatively selected on observed characteristics, is positively selected on unobserved characteristics implying limited external validity of the RCT.

Azzam: Department of Psychology, Division of Behavioral & Organizational Sciences, Claremont Graduate University, (email: tarek.azzam@cgu.edu) Bates: Department of Economics, University of California, Riverside, (email: mbates@ucr.edu); Fairris: Department of Economics, University of California, Riverside, (email: dfairris@ucr.edu). 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. 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.

Introduction

The goal of much of the empirical work in economics is to estimate a credibly causal (internally valid) effect, which also applies to the entire population of interest (externally valid). In an effort to obtain credible causal estimates, researchers typically exploit naturally occurring exogenous variation or employ randomized control trials (RCTs). However, even when the internal validity of estimates is credible, the same estimates often hold only for a localized subpopulation and may lack external validity (Imbens and Angrist, 1994). Whether the purpose of the research is to test hypotheses derived from theory, explain empirical regularities, or evaluate or inform policy making, it is often desirable to extend the causal estimates beyond the population for which the estimates directly apply.[1] In RCTs, the experimental sample sometimes does, but often does not, correspond to the population of interest.[2] As a result, the generalizability of the RCT results to the remaining population is pertinent to whether a given theory generally holds, the degree to which one mechanism explains a broader phenomenon, or whether a particular policy should be expanded or scaled back.

Non-representative experimental samples may originate through either researcher or participant selection processes. In order to address the external validity of their work, many who design and implement RCTs attempt to randomly sample from the population or show the degree to which their experimental sample is representative of the broader population of interest. Despite researchers' best efforts to achieve a representative experimental sample, participants are often self-selected, even if only in granting consent. Many of the most influential experiments within economics rest on voluntary selection into the study. Individuals randomized in the National Supported Work Demonstration used in LaLonde's (1986) evaluation of observational methods, the Moving to Opportunity housing voucher experiment (Goering et al., 1999), and the Oregon health insurance experiment (Finklestein et al., 2012) are all self-selected into the experimental

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

[2] The composition of the population of interest depends on the question and audience. Policy makers may be only interested in the effect of the policy on those who currently select into it. However, in many instances the population of interest is much broader. Were the program expanded or the selection criterion changed, the effects on these new entrants may also be of interest. Further, researchers often utilize RCTs to answer general questions, which also typically pertain to a broader population than the RCT sample.

sample to some degree.[3] This 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.

Further, 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. In this paper, we build off of Huber (2013) and Black et al. (2017), both of which examine selection on the basis of unobserved heterogeneity into compliance within the sample. Kowalski (2018) explicitly ties such tests to the external validity of estimates within the sample. We contribute to this literature by providing to our knowledge the first tests for external validity of experimental results to the broader population from which the experimental sample originates. We do this on the basis of unobserved heterogeneous characteristics, including heterogeneous responsiveness to treatment. These straight-forward tests both provide evidence for the extent to which results from RCTs are generalizable to a larger population of interest, beyond the experimental sample, and provide estimates of the direction and magnitude of selection on unobserved characteristics.

We explore these matters in the context of an analysis of the efficacy of a freshmen year learning community in increasing first year college retention at a large four-year public research university. The United States currently lags behind several developed economies in the percentage of its population holding four-year degrees. While in 1995 the United States lead the world in the share of the population with a bachelor's degree or higher, by 2016 it had fallen to tenth (OECD 1995, 2017). The share of the population aged 25-34 in the United States with tertiary degrees is 47.5%, which lags significantly behind international competitors such as South Korea, Canada, and Japan with rates ranging from 60-70%. A number of efforts, both public and private, have attempted to close that gap.[4] Fostering increased college-going and college-

[3] In the National Supported Work Demonstration the majority of participants were drawn from those who had signed up, but were not enrolled in other government programs, but remaining slots were filled by "walk-in" enrollees (MDRC, 1980).

[4] Public initiatives were prominent in the Obama administration. For examples from the private sector, see the Lumina Foundation's Strategic Plan to increase the percentage of college graduates from roughly 40% to 60% by the year 2025. https://www.luminafoundation.org/lumina-goal

completion is important for counteracting trends in growing wage inequality and in preparing the work force of the future. Higher education scholars are clear that success in the first year of college is key to college completion since first year retention rates are typically far below those of succeeding years (Isher & Upcraft, 2005).

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 experience programs to ease the transition into college and to support students academically and socially as they adjust to the college experience. These first year experience programs may take many forms, such as freshman seminars aimed at teaching study skills and time management, supplemental or remedial instruction in core subjects, and peer mentoring and tutoring. Learning communities are another heavily utilized first year experience commonly 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 a learning community program at a four-year higher education research university increases first year college student retention. We offer first “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 (i.e., those who “won” the lottery) ended up not taking the

program, and some students who were not assigned to the program from the self-selected population (i.e., those who "lost" the lottery) 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 the randomization. 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 college retention at a four-year institution. The ITT and LATE estimates of program impact reveal no statistically significant effect on first year retention.

Next, we consider the issue of generalizability of our results. To do so, we must first define the population of interest. Ultimately, the composition of the population of interest depends on the purpose of the research. Were the purpose of the empirical work to evaluate theory, the population of interest would be composed of the entire population for whom the theory is theorized to apply. In contrast, were it a pure policy evaluation, policy makers may be only interested in the effect of the policy on those who currently select into it. There is also a middle ground. Programs occasionally expand or change the selection criteria, and the effects on these new entrants may also be of interest.

To illustrate this point, consider again the case of housing vouchers. Jacobs (2004) and Chyn (2018) argue that self-selection into the famous Moving to Opportunity experiment may have mitigated the estimated effects of neighborhoods reported in Goering et al. (1999), Orr et al. (2003), Sanbonmatsu et al. (2011), and Chetty et al. (2016). Those who selected into study may have been less susceptible to the effects of receiving a voucher than was the broader population. The involuntary receipt of housing vouchers may be somewhat unique to the context studied in Jacobs (2004) and Chyn (2018). Policy makers may well care more about the effects of vouchers on those who might apply to get them.[5] However, the MTO experiment is also of great interest to social scientists because it identifies the effect of neighborhood quality for those who enter the study. Whether neighborhood quality matters to the lives of residents is a population-level question, for which the self-selection into the study may matter. While rightfully influential, the

[5] Even if those who self-select into the program are the population of interest, the experimental sample may still fail to reflect the entire population of interest if randomization causes would-be participants to opt out, the experiment is not up to scale, or the exact recruitment and selection processes are subject to change.

question remains as to whether we can generalize the results from such RCTs to the subsets of the populations from which the experimental samples are drawn who do not appear in the experiment.

In our context, the purpose of the research was to evaluate the FYLC policy at the implementing institution. While this purpose initially narrowed the population of interest, ultimately, there were several different populations of interest, at various stages in the history of the program's evolution illustrating the potential broadening of the ex post population of interest. Initially, the intent of the experiment was to uncover the effect of the FYLC on the experimental population in a program with voluntary enrollment. Thus, we first examine whether we can generalize the LATE to the rest of the students who enrolled in the experiment. We follow Black et al. (2017) to test the external validity of the RCT LATE estimate within the experimental sample. We find that those who enter the experiment, but do not comply with the randomization either by selecting out of the learning community despite winning the lottery or by entering the program despite losing the lottery, are not statistically different from those who do comply with the randomization. These results support generalizing the RCT results to the average effect of treatment on those who select into the study.

However, the population of interest moved beyond the experimental sample in two stages. The first stemmed from a deviation between the plan and the practice. A small share of the students who received treatment were late arrivals, and so never participated in the randomization process. Naturally, we should be interested in the effects of the program on these students as well. Second, the experimental results are specific to the particulars of the recruitment strategy, which may often evolve. In our context, in later years the institution extended the program to nearly 90% of the population of freshmen in the college in which the program originated. [6] As a result, the population of interest for policy evolved, making the external validity of the experimental results to a larger population of interest crucially important.

[6] There exist many instances in the evolution of learning communities in which institutions of higher education initially experimented with these programs on a population of voluntary participants, only to later mandate them for targeted populations – from developmental students with weak academic skills to honors students with superior academic skills and even for the entire freshman class as a whole (see, for example, Matthews, Smith, and MacGregor 2012).

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

Our results on this latter set of external validity issues reveal that those students who express a desire to enroll in the program are, in many observed respects, from more vulnerable segments of the student population – they tend, for example, to have lower high-school GPAs, lower SAT scores, and come from less-advantaged backgrounds. However, in conducting the tests for selection into the experimental sample, we find that the experimental population differs significantly from the remaining population who do not enter the experiment on unobserved characteristics both unconditionally and conditional on observed covariates. In particular, their unobserved characteristics – presumably, things like grit, determination, focus, and commitment – make them even more likely to succeed in college, as measured by first year retention rates, than their peers who express no interest in the first year learning community. Consequently, the RCT results cannot be generalized to these larger populations of interest. Our results thus represent a cautionary tale for evaluative exercises in the context of voluntary selection that do not adequately address the problem of possible selection bias on unobserved characteristics.

Lastly following LaLonde's (1986) seminal work, we use the data to perform a "within-study design" comparing the experimental results to those from standard observational approaches, which have been used both by institutional researchers and academics to estimate the effects of FYLCs. Moreover, we analyze these two sets of results considering the internal and external validity of both experimental and the standard observational approaches in order to reflect on what we learn from such within-study designs in general.

Using OLS, a quasi-maximum-likelihood nonlinear estimation, and propensity score matching, we estimate the effect of FYLCs on the first year college retention both unconditionally and conditional on the variables available to institutions for admission decisions. We find stark

differences in estimated impacts between the two approaches, with the observational methods revealing large and statistically significant benefits of the program and the RCT uncovering little discernable impact.

What explains these disparate findings? The observational analysis (on a particular population or a random sample of it) provides an estimate of the effect of treatment on that population if the assignment of treatment is exogenous conditional on the set of covariates. The RCT, on the other hand, provides an internally valid estimate (assuming proper randomization) of treatment on the sample that was selected and participated in the experiment. Only in the presence of random selection into both the experiment and treatment would we expect the results of the different approaches to be the same. Absent this, differences alone between the two sets of results are uninformative as to which results (if any) should be privileged in arriving at a population-level parameter. Without further tests, we do not know whether selection into treatment or into the experiment is problematic, nor the relative magnitudes of the selection issues. The tests that we propose provide more clarity on such selection. Here (as in LaLonde, 1986), treatment is administered largely through the experiment.[7] As a result, the nonrandom selection into the RCT, which made the RCT externally invalid, causes the observational analysis to be biased.

The paper is organized as follows: First, we review the literature on the central contributions of the paper – evaluations of the impact of learning communities on first year college retention and the issue of "external validity" in RCT estimates of program impact. 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. The final section offers a summary discussion and conclusion.

Literature Review

Many evaluation studies of the broad range of first year experience programs are conducted by "in-house" institutional researchers. These studies commonly take the form of survey data on student and faculty participants, reporting on features such as satisfaction with the

[7] In LaLonde (1986), the only members within the data who receive treatment first enlisted in the experimental study.

experience and subjective responses of the extent to which the program achieved a set of pre-established goals. An early review of program evaluation studies containing more rigorous research designs is contained in Barefoot, Warnock, Dickinson, Richardson, and Roberts (1998). The empirical designs of some of these studies are simple comparisons of outcome means (e.g., retention or freshman year GPA) across participant and non-participant groups, while others add covariates to control for differences in observed background characteristics. Few employ empirical strategies designed to handle possible nonrandom selection on unobserved covariates. Pascarella and Terenzini (2005) offer a somewhat more recent review of first year experience evaluations and call for a randomized control trial (RCT) design, which can address the problem of selection bias.

Numerous observational studies have been published since the appearance of Pascarella and Terenzini's review article (e.g., Porter and Swing, 2006, and Jamelske, 2009), some utilizing more advanced techniques, such as propensity score matching (Clark and Cundiff, 2011), instrumental variables (Pike, Hansen, and Lin, 2011), and Heckman's two-step procedure (Hotchkiss, Moore, and Pitts, 2006). In each, the exogeneity assumptions necessary for causal interpretation of the results are problematic. A handful of studies of first-year experience programs contain more convincing internal validity through RCT designs to avoid selection bias in their estimates of the causal impact of first year experience programs. For instance, Bettinger and Baker (2014) find that peer mentoring via phone, text messaging, and social networks has a statistically significantly 5-6 percentage point increase on retention of nontraditional students in a wide range of American universities. In contrast, Paloyo, Rogin, and Siminski (2016) find a small and statistically insignificant effect of supplemental instruction in certain introductory courses on grades in those courses in a large Australian university. Angrist, Lang, and Oreopoulis (2009) find sizable positive impacts of academic support services such as peer advising, mentoring, and supplemental instruction at a large Canadian university, but only for female students.

Focusing on the impact of learning communities more specifically, there are three RCT studies that estimate their impact on various programmatic outcomes. Two of these studies estimate the impact of remedial learning communities on retention rates in two-year community college settings (Scrivener et al. (2008) and Visher et al. (2012). Both find small positive effects on

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

The third RCT evaluation of learning communities provides the closest study to the one at hand. Russell (2017) examines the effects of experimental study groups at the Massachusetts Institute of Technology, an elite four-year institution highly regarded in the sciences, technology, engineering, and mathematics (STEM) fields. This learning community brings groups of students together in linked introductory courses with smaller class sizes, enhanced mentoring by both upper-division majors and faculty, and dedicated study spaces to foster the formation of academic and support study groups. 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. Women in the program are statistically more academically successful, as measured by GPA and total credits, and underrepresented minority students are though noisily estimated twice as likely to major in higher-paying STEM fields as a result of the program. 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.

In this paper, we utilize an experimental design to evaluate the impact on retention of a learning community program with voluntary enrollment, taking place at a large four-year public research university with a diverse student body. The existing literature to date has not offered a rigorous causal estimate of the impact of a learning community on first year retention for a four-year institution of higher education.

In the treatment effects literature, every effort is made to estimate a true causal estimate of program impact that is distinct from estimates which are plagued by bias and thus internally

invalid, or which hold only for some local population and therefore lack the external validity to generalize to the broader population of interest. One such parameter, the average treatment effect (ATE) can be defined using the potential outcomes framework of Rubin (1974). Let Y_{1i} be the outcome individual *i* would realize were she given treatment (D_i=1), and Y_{0i} be the outcome individual *i* would realize were she not given treatment (D_i=0). The ATE is then $E[Y_{1i} - Y_{0i}]$ over the population. The fact that both potential states of the world are not realized simultaneously for the same individual, requires researchers to assume that for some subsample the assignment into treatment is otherwise independent of the outcomes. The exact assumptions required for internal validity depend upon the parameters to be estimated and the empirical design employed, whether it be a randomized control trial (RCT) or an observational design such instrumental variables (IV).

Imbens and Angrist (1994) show that researchers can identify average treatment effects on a subsample of the population with minimal assumptions – namely, (1) an instrument (*Z*) exists such that it causally affects assignment into treatment, and is otherwise unrelated to the potential outcomes (Y_1 and Y_0) and (2) the instrument monotonically affects assignment into treatment, which may be expressed in the binary case, without loss of generality, as $D_{1i} \geq D_{0i}$, for all *i*, where D_1 is the treatment assignment that would be realized were the instrument to take a value of 1. Under these assumptions, Imbens and Angrist show that researchers identify a local average treatment effect (LATE), that is the ATE for the subsample whose assignment into treatment was determined by the instrument. They term this subsample the compliers, as opposed to always-takers or never-takers who respectively would or would not enter the treatment regardless of the value the instrument takes. Monotonicity assumes that there are no so-called defiers, who are responsive to the instrument in the opposite direction as the compliers.

With an RCT, we need only assume that the randomization was truly random in order to conduct an intent to treat analysis (ITT), by which researchers identify the average effect of trying to apply treatment to some sample involved in the RCT (Angrist, Imbens, and Rubin, 1996). There may be limits to the usefulness of ITT analyses. The presence of no-shows from those randomly assigned to treatment and substitution into treatment of those who were not randomly assigned to treatment can obscure the true effect of treatment. However, by using the randomization as an instrumental variable, we can recover an estimate of the effect of treatment itself (Bloom, 1984).

Doing so, though, requires the same assumptions as shown in Imbens and Angrist (1994) and provides an estimate of a similar parameter. Thus, when we use randomization as an IV, we obtain an estimate of the average effect of treatment for those who participated in the randomization and who received treatment because of the randomization. Thus, it is a LATE rather than an estimate of "the treatment on the treated," which is a term often misapplied to this parameter.

There are multiple selection processes which may lead the LATE for the subsample for which we have a causal estimate to differ from the ATE of the entire population of interest. First, the compliers for whom we have estimated the LATE may differ from those who dropout of treatment and those who substitute into treatment. In which case, it would be unreasonable to expect the LATE for the compliers to carry over to the rest of the sample who received treatment. Researchers often compare the observed characteristics of those who leave and those who substitute into treatment – in violation of their initially-assigned status – with those who remain in their assigned control or treatment groups as rough evidence for whether such non-compliance with the randomization is worrisome. Huber (2013) provides new tests for nonrandom noncompliance on the basis of unobserved characteristics, and Black et al. (2015) extends similar tests to more general settings. Failure of these tests to identify differences in unobserved characteristics in the no-shows or crossovers, provides reassurance that the RCT is externally valid and that the LATE may generalize to other populations of the experimental sample.

As noted above, the population may but often does not include the entire population of interest. Deaton and Cartwright (2017) comment that "frequently, [the experimental sample] is selected in some way, for example to those willing to participate, or is simply a convenience sample that is available to those conducting the trial." To some extent this is true of the study at hand. However, the issue of nonrandom selection into RCTs is not unique to this context. The ethics of conducting RCTs on human subjects usually requires participants' consent, making virtually all social experiments in which the population of interest is broader than the experimental sample susceptible to such selection.

Andrews and Oster (2018) and Ghanem, Hirshleifer, and Ortiz-Becerra (2018) give careful attention to the issue of possibly non-random sample selection of RCTs. Andrews and Oster (2018) model the decision to participate in a study and approximate weights using observed variables to adjust RCT results to provide an estimate of the ATE and provide bounds on the ATE. These estimates and bounds require assumptions on the relationship between the observed and unobserved covariates, and their relationship to treatment effect heterogeneity, which may not hold in many cases.

Ghanem, Hirshleifer, and Ortiz-Becerra (2018) provide distributional tests for non-random attrition on the basis of baseline outcome data. These rigorous tests provide concrete evidence of whether attrition threatens the validity of the RCT results without further data. However, attrition is but one way in which an experimental sample may become unrepresentative. Further, with only baseline data on those who attritt, these tests can address only time-invariant unobserved heterogeneity and are unable to incorporate within tests differences in responsiveness to treatment.

We directly test for selection into the experiment on the basis of unobserved characteristics and heterogeneous responsiveness to treatment. We do so by comparing outcomes for those who do not receive treatment, by whether they participate in the experiment, both conditional on observed covariates and unconditionally. We do the same among the treated populations comparing those who receive treatment by randomized assignment to those who received treatment without participating in the randomization. We interpret the results of those tests as pertaining to the external validity of the research design.

Comparison of outcomes between the experimental control and the untreated population has been conducted in Lise, Seitz, and Smith (2015) and Sianesi (2017). Lise, Seitz, and Smith (2015) adopt a within-study design to test the performance of their search and matching model against the results of the Canadian Self Sufficiency Project, which uses randomization to evaluate the effectiveness of incentivizing welfare beneficiaries to seek employment. They do so by first calibrating their model of the population using the control sample from the experiment. They then introduce the treatment into the model and compare the predicted outcomes from the

model to the experimental results. In so doing, they hold the experiment as the high standard against which the model is evaluated.

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. Under the assumption of homogeneous average responsiveness to treatment across those who select into or out of the study (CIA-β), Sianesi attributes any differences in unobserved characteristics as being the result of the randomization itself. She then applies these tests to examine the selection on unobserved variables into and primarily out of the Employment Retention and Advancement (ERA) experiment in the United Kingdom. Sianesi finds substantial selection on unobserved variables into the experiment in the ERA context, similar to what we find in this study. However, under CIA-β, she credits such selection as evidence of the randomization causing differing results. Whereas, we view selection on unobserved characteristics to be evidence against the external validity of the RCT estimates.

It seems difficult to maintain that responsiveness to treatment would be the same (CIA-β) across populations that differ significantly on unobserved characteristics. Our additional tests compare the outcomes of those who receive treatment within the randomized sample to those who receive treatment otherwise, thus incorporating heterogeneous responsiveness to treatment into the tests. When such "essential heterogeneity"[8] is integrated into the tests, maintaining homogenous slopes despite significant differences on total unobserved heterogeneity seems unreasonable. Consequently, rather than assuming homogeneous responsiveness to treatment we view these tests as directly testing the external validity of the experimental LATE estimates. These tests are very simple to perform, but do require data about individuals not randomly assigned. We provide more detail about the context of our RCT and the methodological details of our tests below. The benefit to researchers is that we illustrate a way for researchers using RCTs to provide more concrete evidence of the external validity of their experimental results.

Our final exercise follows LaLonde's seminal work (1986), which gave rise to a "within-study design" literature. Such studies largely depict social experiments as the standard against which

[8] "Essential heterogeneity" is the term Heckman et al. (2006) coins to describe such heterogeneous responsiveness.

the performance of observational approaches are judged. However, we argue that one cannot necessarily assign a hegemony of estimation design to every case, particularly if a population parameter is of paramount interest. There are at least two reasons for this conclusion. First, as suggested by Calonico and Smith (2017), within-study designs are context dependent; the sample for comparison and the variables available to researchers may matter a great deal. For example, the fact that the observed variables are inadequate to capture differences across populations in one context does not imply that the same holds true in all contexts. In other instances, the assignment of the variable of interest conditional on observed covariates may not lead to selection bias either due to exogenous natural assignment of "treatment" or to the richness of the set of the observed variables.

We add to this discussion evidence of a second point; namely that RCTs and observational analyses often estimate different parameters. The parameters identified in each study depend upon both the strategy researchers use and the sampling they employ. Consider instances (such as the one explored in this paper) where the observational sample is randomly selected or includes the broader population from which the RCT sample is drawn through the participation decision of both researchers and participants. Were treatment exogenously dispersed throughout each sample (conditional on covariates), OLS would deliver an estimate of the average effect in the population or on the selected population depending on which sample was used. Were some form of instrumentation needed to focus attention on only the exogenous variation in treatment (either due to naturally occurring endogeneity or noncompliance with the randomization of the experiment), the analysis of each sample would produce a different LATE. It is not clear which would be more representative of the average population effect. Accordingly, we demonstrate the necessity of testing for the external validity of the RCT, as the results of these tests determine whether the RCT results are indeed the standard against which we should assess observational approaches.

Background and Data

The First Year Learning Community (FYLC)[9] began on a small scale and included approximately 200 students from a population of roughly 4,000 incoming freshman. During its

[9] Not the program's actual name.

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.

During the initial few years of the FYLC implementation, the program was evaluated using standard observational analyses, controlling for differences across those who enrolled in the program and those who did not in various background characteristics, such as gender, high-school GPA, SAT scores, first generation college status, and socio-economic status. The results of this design suggested that students in the FYLC program were retained at a much higher rate (and statistically significantly so) than students in the comparison group (Fairris, Castro and Son, 2010). However, students voluntarily enrolled in the FYLC on a first-come, first-served basis, and the evaluation may consequently suffer from bias due to nonrandom selection on unobserved characteristics. This provided the impetus for an analysis of impact utilizing an RCT design.

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).[10] The staff was largely successful in accomplishing this goal.

In prior years, the program had been highlighted in presentations to matriculating students and their parents at consecutive sessions of the summer freshman orientation program, at which incoming students enroll for fall classes. Whenever enrollment reached program capacity at those enrollment sessions, the program was no longer advertised in presentations at subsequent orientation sessions to either students or parents. 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

[10] 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.

assignment. Under the former regime, the enrollment decision took place within a day of students and parents hearing about the program during freshman orientation.

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.

Table 1: Student Background Characteristics

	Assigned Control	Assigned Treatment	Difference	Lottery Sample	Non-lottery Sample	Difference
High-school	3.46	3.46	0.01	3.46	3.53	-0.07
GPA			(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	0.74	0.75	0.01	0.75	0.71	0.04

Campus			(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 the background variables is statistically significantly different across the treated and control populations. The same should be true for unobserved characteristics as well (Shadish, Cook & Campbell, 2002). 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. We turn to a careful analysis of how observed background characteristics translate into retention prospects below.

As mentioned above, there are three important instances of migration between assigned groups in the data. In Table A1 in the appendix, we show how these migrants differ from the rest of their assigned group, by regressing treatment on student covariates separately within each of the three exclusive subpopulations: 1) those who were assigned treatment; 2) those who were assigned control status; 3) those who did not enroll in the lottery.

Of the 824 students initially assigned to the treatment group, 170 (or 21%) did not attend any of the program courses or services. They are not a random draw from the assigned treatment group. These *no-shows* have statistically significantly slightly higher SAT math scores, slightly lower verbal scores, and were substantially less likely to come from low income families than those who remained in the program.

There is also contamination in the control sample in this randomized control trial. An analysis of course enrollment records indicated that 108 students (15%) assigned to the control group

enrolled in FYLC courses (presumably with permission of the program director and as a partial replacement for those no-shows from the assigned treated group) even though they technically should not have been allowed to do so. These *crossovers* possess statistically significantly slightly lower SAT math scores, higher verbal scores, and were more likely to be female and to live on campus than those who remained in the control population.

Finally, 117 of 6,566 students (2%) who did not initially express interest in enlisting in the program and were accordingly not entered into the lottery eventually entered the program. Though the observed differences are much smaller in magnitude, these *late-takers* possess lower high-school GPAs, lower math SAT scores, high verbal SAT scores, are more likely to be female and first generation college students, and less likely to be from a poor family (defined as earning less than $30,000 per year) than the rest of the freshman class in the non-lottery population.

We present a figure depicting these various subpopulations in Figure A1 of the appendix. None of these various migrations in violation of initial assignment bias the "intent to treat" estimates of program impact, though they do present complications in estimating the effects of treatment itself. However, their presence also provides opportunities for exploring the extent to which our estimated LATE can be generalized to the entire population of interest. We explore these issues as well in the analysis below.

Empirical Methodology

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

Analysis 1

Randomization among the experimental group provides two groups of similar size; those who won and those who lost the lottery. These two groups should be in expectation identical with

respect to both observed and unobserved pre-determined characteristics. Accordingly, we may estimate the causal "intent to treat" effects of the program using standard approaches.

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

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

where *Retention*$_i$ indicates whether student *i* remained in school the following year, *won*$_i$ indicates whether individual *i* entered and won the lottery, and $\boldsymbol{X_i}$ is a rich vector of student background characteristics discussed in the "Data" section above. As causal identification does not hinge on the covariates we conduct the analysis both with and without conditioning on $\boldsymbol{X}$. We prefer to include these controls because doing so generally provides more efficient estimates that remain consistent.[11]

As stated above, the main dependent variable used for this analysis is first year retention (equaling 1 if the student persists into the second year and 0 otherwise). Consequently, we may wish to adopt a functional form that respects this binary form. However, we do not know the exact functional form of the error term, and may wish our inference to be robust to heteroscedasticity. Standard maximum likelihood approaches require assuming that the functional form is properly specified including that the errors are independent and identically distributed for consistency. Using the quasi-maximum likelihood estimation (QMLE) framework, we can allow some features of the density function to be misspecified, but still consistently identify the conditional mean with appropriate inference, as long as we correctly specify the distributional family (Gourieroux, Monfort, and Trognon, 1984). The log likelihoods of many commonly specified distributions (such as normal, exponential, Bernoulli, and Poisson) all belong to the linear exponential family. Similar to Papke and Wooldridge (1996), we use the logit QMLE to estimate the non-linear model below:

$$E[Retention|treatment, \boldsymbol{X}] = \frac{e^{(\alpha + won_i\beta_{ITT} + \boldsymbol{X_i\gamma})}}{1 + e^{(\alpha + won_i\beta_{ITT} + \boldsymbol{X_i\gamma})}}. \quad (2)$$

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

We then average over the estimated partial effects for each observation to obtain an estimate of the average intent to treat effect, which is easily comparable to the estimates from OLS estimation. Again, we perform estimation both including and excluding the vector of control variables (***X***).

Were compliance with the lottery perfect, the average intent to treat estimate would also provide an estimate of the average effect of treatment for the experimental sample. However, 170 individuals entered and won the lottery, yet never joined the FYLC, and 108 students lost the lottery, but were still able to make their way into the program. 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 2-stage least square with the lottery as an instrumental variable for enrollment in the FYLC. Thus, we model FYLC according to the following:

$$FYLC_i = \alpha + won_i\delta_1 + \boldsymbol{X_i\delta_2} + e_i, \tag{3}$$

where $FYLC_i$ indicates whether student *i* enrolled in the learning community. We then use the fitted values from this regression to estimate the average effect of FYLC on retention.

$$Retention_i = \alpha + \widehat{FYLC_\iota}\beta_{TOT} + \boldsymbol{X_i\gamma_1} + \epsilon_i. \tag{4}$$

With non-linear estimation including the fitted values does not yield consistent estimates. However, we can treat the endogeneity in *FYLC* by also including the residuals from estimating equation 3 (Vytlacil, 2002; Wooldridge, 2014). We use logit QMLE to estimate the following:

$$E[Retention|treatment, \boldsymbol{X}] = \frac{e^{(\alpha+won_i\beta_{TOT}+\boldsymbol{X_i\gamma_1}+\widehat{\boldsymbol{v_i}}\boldsymbol{\gamma_2})}}{1+e^{(\alpha+won_i\beta_{TOT}+\boldsymbol{X_i\gamma_1}+\widehat{\boldsymbol{v_i}}\boldsymbol{\gamma_2})}}, \tag{5}$$

where $\widehat{\boldsymbol{v_\iota}}$ are the residuals from estimating equation **(**3) with OLS. The t-statistic on $\boldsymbol{\gamma_2}$ provides a convenient test for whether the noncompliance with the lottery introduces selection bias, necessitating the instrumental variables approach. 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 with 500 replications to estimate the standard errors.

While this procedure provides us with internally valid causal estimates of the effect of treatment, without further assumptions these estimates hold only for those who actually received treatment because they won the lottery. We may wonder whether these estimates generalize to the average treatment effect among the whole experimental sample. Our second set of analyses will directly test whether such selection in and out of treatment is ignorable for identifying broader treatment effects, after conditioning on observed covariates.

Still a more interesting parameter may be the average treatment effect for a larger population of interest—for example, what we expect to happen to the first year retention rate if all were automatically enrolled in FYLC. Broadening these results to the average treatment effect requires that selection into the lottery is also ignorable.[12] In our second set of analyses, we will also gain insight into whether those who self-select into the lottery differ on unobserved characteristics compared to those who do not enter the lottery.

Analysis 2

In this piece of analysis, we examine the extent to which our RCT results may generalize to broader populations. In so doing, we first apply the tests proposed in Black et al. (2017) which are closely related to those introduced by Huber (2013) to judge whether we may reliably generalize our experimental results to the rest of the experimental sample. We then introduce tests for whether the experimental sample is representative of the entire population of the incoming freshmen class.

To formalize these tests, let D indicate treatment (FYLC participation), Y be the outcome (first year retention), and Z denote the binary assignment (whether or not an individual wins the lottery for FYLC participation). We add to this familiar framework L as an indicator for participation in the experiment. Much of the earlier treatment effects literature as well as both Huber (2013) and Black et al. (2017) considers only the population for which L=1. However, we are ultimately also interested in generalizability to the entire population, and so we must add two additional groups to the typical division of the sample among compliers, always-takers, and never-takers. Namely, we

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

add the "late-takers" who take-up the treatment despite not entering the lottery, and the "never-ever-takers" who do not enter the lottery and do not take the treatment. Again, we maintain the monotonicity assumption that there are no defiers. Accordingly, we summarize the groups that comprise our sample, and write the expected outcome for each subsample conditional on $\boldsymbol{X}=\boldsymbol{x}$ in Table 2.

Table 2 Sample composition

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

We write the model in familiar linear form with unobserved heterogeneous intercepts as well as heterogeneous effects of treatment:

$$Y_i = \boldsymbol{X_i\gamma} + D_i b_i + \varepsilon_i, \tag{6}$$

where $\boldsymbol{X_i}$ denotes a vector of observed characteristics. Here, ε_i represents the unobserved heterogeneous intercept, while $b_i = \beta + e_i$ represents the heterogeneous responsiveness to treatment, which are centered on the ATE, β. Note that the model implicitly assumes that neither Z nor L directly affects the outcome variable, however selection into treatment may depend on both.

Following Black et al. (2017), in order to test for nonrandom selection into compliance on the basis of unobserved heterogeneity we can test whether the mean heterogeneous fixed errors and heterogeneous effects differ across populations using side-by-side comparisons. For instance, the difference between controls and no-shows may be expressed as the following:

$$E(Y_i|\boldsymbol{x}, D = 0, L = 1, Z = 0) - E(Y_i|\boldsymbol{x}, D = 0, L = 1, Z = 1) = \bar{\varepsilon}_c - \bar{\varepsilon}_{ns}. \quad (7)$$

As does Black et al. (2017), we can test whether this difference is zero in the following conditional mean function for the sample that enters the lottery but does not take up treatment:

$$E(Y_i|D_i = 0, L_i = 1) = Z_i\pi_{01} + \boldsymbol{X_i\gamma_{01}}. \quad (8)$$

As both Huber (2013) and Black et al. (2017) note, because the no-shows are composed only of never-takers and the control group of never-takers and compliers, this test ultimately assesses whether the compliers differ systematically on the basis of unobserved characteristics from never-takers. Thus, if we reject the null the hypothesis that $\pi_{01} = 0$, then selection on unobserved characteristics may be problematic, and the LATE results from the RCT are unlikely to generalize to the remaining experimental sample. We repeat the exercise among those in the experimental sample who receive treatment.

Accordingly, we can test whether always-takers differ systematically from compliers within the lottery by estimating the following for the sample that enters the lottery and takes treatment:

$$E(Y_i|D_i = 1, L_i = 1) = Z_i\pi_{11} + \boldsymbol{X_i\gamma_{11}}. \quad (9)$$

Performing a standard t-test on $\hat{\pi}_{11}$ 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 test also factors in possible heterogeneous treatment effects.[13]

To introduce our tests for selection into the experimental sample on the basis of unobserved heterogeneity, we first split the population into just four groups: those who entered the lottery and received treatment; those who entered the lottery and did not receive treatment; those who did not enter the lottery and did not receive treatment; and those who did not enter the lottery but did receive treatment. This last group – the late-takers – may not be present in all settings, but they are certainly not unique to our experiment. The compliers who were moved by housing demolitions

[13] Under the assumption that $\bar{\varepsilon}_c - \bar{\varepsilon}_a = \bar{\varepsilon}_t - \bar{\varepsilon}_s$ (which may only hold in special cases), testing the difference between $\widehat{\pi_{10}}$ and $\widehat{\pi_{00}}$ provides a direct test for whether there the effects of differ between the compliers and always-takers in the population.

in Jacobs (2004) and Chyn (2018) are essentially late-takers to Moving-to-Opportunity compliers from Goering et al. (1999), Orr et al. (2003), Sanbonmatsu et al. (2011) and Chetty et al. (2016). Late-takers are also present in the data underlying the evaluation of the efficacy of Teach for America in Glazerman, Mayer, and Decker (2006) and in the large-scale class-size experiment of Tennessee STAR analyzed in Folger and Breda (1989), Krueger (1999), Krueger and Whitmore (2001), and Chetty et al. (2013) among many others. The late-takers are not necessary to test for selection on unobserved variables, but their existence provides an additional test for selection into the experiment on the basis of unobserved variables, including responsiveness to treatment. Let $E(\varepsilon_i|D=0,L=0)=\bar{\varepsilon}_{00}$, $E(\varepsilon_i|D=0,L=1)=\bar{\varepsilon}_{01}$, $E(\varepsilon_i|D=1,L=0)=\bar{\varepsilon}_{10}$, and $E(\varepsilon_i|D=1,L=1)=\bar{\varepsilon}_{11}$.[14] Likewise, let $E(e_i|D=1,L=0)=\bar{e}_{10}$ and $E(e_i|D=1,L=1)=\bar{e}_{11}$.[15] Accordingly, the difference in outcomes conditional on $\boldsymbol{X}=\boldsymbol{x}$ within treatment status is given by the following:

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

$$E(Y_i|\boldsymbol{x},D=0,L=1)-E(Y_i|\boldsymbol{x},D=0,L=0)=\bar{\varepsilon}_{01}-\bar{\varepsilon}_{00} \quad (11)$$

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

$$E(Y_i|D_i=0)=L_i\pi_0+\boldsymbol{X_i\gamma_0}. \quad (12)$$

So long as the treatment and non-treatment do not differ by participation in the lottery and for any setting of covariates there is a chance to see each state of treatment, a substantially or significantly non-zero $\widehat{\pi_0}$ provides evidence of selection on unobserved characteristics into the experiment, making the claim of external validity of the experimental results to the non-experimental population difficult to accept. The intuition is simple, since neither received treatment, any differences in outcomes must be due to differences in selection. Further the sign and magnitude of $\widehat{\pi_0}$ demonstrates the extent and direction of the selection bias.

[14] As the model is full saturated, we may write $\bar{\varepsilon}_{11}=1-\bar{\varepsilon}_{10}-\bar{\varepsilon}_{01}-\bar{\varepsilon}_{00}$.

[15] We may add $E(e_i|D=0,L=0)=\bar{e}_{00}$ and $E(e_i|D=0,L=1)=\bar{e}_{01}$ such that $\bar{e}_{11}=1-\bar{e}_{10}-\bar{e}_{01}-\bar{e}_{00}$, where $\bar{e}_{00}$ and $\bar{e}_{01}$ are the average differential effects those who did not receive treatment whether or not they were in the experiment would have experienced were they to have received treatment.

Granted that some who did not enter the lottery make their way into treatment, we may conduct an additional test on the remaining sample, restricted to those who do receive treatment:

$$E(Y_i|D_i = 1) = L_i\pi_1 + \boldsymbol{X_i\gamma_1}. \quad (13)$$

A simple t-test on $\widehat{\pi_1}$ with this restricted sample provides a summative test of whether $\bar{e}_{11} + \bar{\varepsilon}_{11} - \bar{e}_{10} - \bar{\varepsilon}_{10}$ equals zero. In so doing, we test whether those who do not enter the experiment differ on the basis of unobserved characteristics and heterogeneous effects from those who do enter the experiment.[16]

In order to show what these tests reveal and the assumptions upon which our interpretation of the results rely, we revisit the potential outcomes framework where Y is the observed outcome, Y_1 is the outcome that would be manifested under treatment, and Y_0 is the outcome that would be manifested without treatment. As before, L=1 denotes participation in the lottery, Z=1 denotes being selected for treatment by the lottery, and D=1 indicates receipt of treatment. Let P (P= E(D|L=0)) stand for the share of those who do not participate in the lottery, but do receive treatment.

In simple settings, interpretation of the test requires no additional assumptions. Specifically, when compliance with the randomization is perfect and treatment status is homogeneous among the non-experimental population, we maintain, that the randomization was carried out properly. That is $E(Y_1|L=1,Z=1) = E(Y_1|L=1,Z=0)$ and $E(Y_0|L=1,Z=1) = E(Y_0|L=1,Z=0)$. Second, we maintain that being selected for the control (or treatment) has no effect on the outcome independent of treatment status, such that $E(Y|L=1,D=0,Z=0) = E(Y_0|L=1,D=0) = E(Y_0|L=1) = E(Y|L=1,D=0)$.[17] Both of these assumptions are standard to interpreting experimental results.

Interpretation of the results from our tests in more complicated settings require assumptions beyond the standard assumptions previously mentioned and the standard monotonicity assumption required for estimating the LATE.[18] Noncompliance leads us to add the first additional assumption; namely, that noncompliance with the randomization is "ignorable, i.e.,

[16] Under the assumption that $\bar{\varepsilon}_{01} - \bar{\varepsilon}_{00} = \bar{\varepsilon}_{11} - \bar{\varepsilon}_{10}$, testing the difference between $\widehat{\pi_1}$ and $\widehat{\pi_0}$ provides a direct test for whether the effects differ between the experimental sample and the remaining population.

[17] Equivalently, $E(Y|L=1,D=1,Z=1) = E(Y_1|L=1,D=1) = E(Y_1|L=1) = E(Y|L=1,D=1)$.

[18] With noncompliance, in order to interpret the experimental results as estimating a LATE, we must maintain the standard monotonicity assumption introduced by Imbens and Angrist (1994) that there are no defiers, who are responsive to the randomization in the opposite direction as the compliers.

not jointly related to treatment and the outcome."[19] Were noncompliance problematic for generalizing within the experimental sample, it may be uninteresting to pursue the question of generalization to an even broader population. Further, we obtain evidence pertaining to whether noncompliance is ignorable by following Huber (2013) or Black, et al. (2017). Combining this assumption with the standard assumptions previously listed implies the following: $E(Y_1|L=1) = E(Y_1|L=1,D=1) = E(Y_1|L=1,D=0) = E(Y|L=1,D=1)$ and $E(Y_0|L=1) = E(Y_0|L=1,D=0) = E(Y_0|L=1,D=1) = E(Y|L=1,D=0)$.

Lastly, due to the presence of both individuals who do and do not take treatment in the non-experimental population, we utilize a second additional assumption; namely monotonic selection by potential outcome. That is if $E(Y_0|L=0,D=1) >> E(Y_0|L=0,D=0]$, then $E(Y_1|L=0,D=1) \geq E(Y_1|L=0,D=0)$.[20] While this assumption seems reasonable, it does rule out instances where among the non-experimental population, differences in responsiveness to treatment are larger in magnitude and opposite signed as the differences in levels of the potential outcome between those who select into or out of treatment.

First, we examine the simple case in which compliance with the randomization is perfect within the experiment and the entire population that does not enter the experiment also does not enter treatment. In this simple case, we only need our first two assumptions. We would like to test whether $E(Y_1|L=1) = E(Y_1|L=0)$ and $E(Y_0|L=1) = E(Y_0|L=0)$, but with our data we are left testing $E(Y|L=1,D=1)$ against $E(Y|L=0,D=1)$ and $E(Y|L=1,D=0)$ against $E(Y|L=0,D=0)$. In comparing $E(Y|L=1,D=0)$ to $E(Y|L=0,D=0]$, we directly test whether there is selection into the experiment on the potential level of the outcome under no treatment. Likewise, consider the case in which the entire non-experimental population receives treatment and the control group within the experiment experiences a withholding of treatment. By the same logic, comparing $E(Y|L=1,D=1)$ to $E(Y|L=0,D=1)$ provides a direct test of selection into the experiment on the potential level of outcome and responsiveness to treatment.

Next, we consider a slightly more complicated case in keeping with the current study, where there is not perfect compliance nor is treatment status homogenous among the non-experimental

[19] For consistency, we adopt this phrasing is from Huber (2013).

[20] Equivalently, we could state the assumption as $E(Y_0|L=0,D=0) >> E(Y_0|L=0,D=1]$, then $E(Y_1|L=0,D=0) \geq E(Y_1|L=0,D=1)$.

population. Under our third assumption, we focus on instances where noncompliance is ignorable. Under these three assumptions, we may write the differences in realized outcomes stratified by realized treatment status as the following:

$$E(Y|L=1,D=0) - E(Y|L=0,D=0) = \\ E(Y_0|L=1) - E(Y_0|L=0) + P[E(Y_0|L=0,D=1) - E(Y_0|L=1)] = 0, \quad (14)$$

$$E(Y|L=1,D=1) - E(Y|L=0,D=1) = \\ E(Y_1|L=1) - E(Y_1|L=0) + P[E(Y_1|L=0,D=0) - E(Y_1|L=1)] = 0. \quad (15)$$

The first unconditional test we conduct under these assumptions compares the expected outcomes of those who did not receive treatment by whether they participated in the lottery. The first difference on the right hand side of equation (14) directly examines whether there is selection into the lottery on the basis of potential outcomes 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. The second test we conduct compares the expected outcomes of those who did receive treatment by whether they participated in the lottery. In equation (15), the first difference directly examines whether there is selection into the lottery on the basis of potential outcomes in the event that both populations were to receive treatment. Again, the latter difference could be nonzero either from selection into the lottery or selection into treatment in the non-randomized population.

Taken together, the two tests may demonstrate how problematic selection into the experiment is. Suppose we observe meaningfully positive differences between the lottery and non-lottery populations within both states of treatment according to the following:

$$E(Y|L=1,D=0) - E(Y|L=0,D=0) \gg 0, \quad (16)$$

and

$$E(Y|L=1,D=1) - E(Y|L=0,D=1) \gg 0. \quad (17)$$

The inequality in equation (16) could be due to those who select into the lottery having higher potential outcomes on average than those who do not self-select, or to those who select into

treatment, but not the lottery, having higher than average potential outcomes than the remaining non-lottery population. Similarly, the inequality in equation (17) could be due to those who select into the lottery having on average higher potential outcomes than those who do not, or those who do not select into treatment nor the lottery having on average higher potential outcomes than those who do select into treatment but did not enter the lottery. However, among those who do not enter the lottery, we cannot simultaneously maintain that those who chose to receive or not to receive treatment are positively selected on their propensity to persist in college.

The fact equations (16) and (17) depend on different potential outcomes (Y_0 and Y_1 respectively) creates a complication that necessitates the monotonic selection on potential outcomes assumption. With this assumption, we hold that differences in responsiveness to treatment are not so large and in the opposite direction as to reverse the sign of selection on the differences in levels of the potential outcome between those who select into or out of treatment. Under this assumption, both tests agreeing on the direction of selection implies that the lottery population is non-randomly selected from the larger population. Therefore, generalizing the results to the larger population would be unreasonable. Naturally, the similar reasoning would hold, if both were substantially negative. Disagreement between the tests suggests strong selection into treatment among the non-lottery population. Having both differences qualitatively close to zero is reassuring regarding the representativeness of the experimental population.

We follow this nonparametric approach to the question by using a nonparametric test for whether the difference between retention rates for those who do and do not chose to enter the lottery are meaningful. Following Efron's (1982) nonparametric percentile method, in each of 10,000 repetitions, we resample the data with replacement and randomly assign each draw 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. We then find the placebo differences (π_0 being the average difference in retention by lottery participation for those who do not receive treatment and π_1 serving as the same for the treated). Next, we compare the differences in retention observed under the actually lottery participation decisions to the distribution of placebo differences we observe under random assignment of "lottery participation." The percentiles of the distribution of these differences around the median of these differences may be sensibly be interpreted as the according confidence intervals.

As shown in Young (2016) using a similar approach, we may also construct nonparametric p-values for our previously estimated mean differences, π_0 and π_1. We can do so using either comparisons in the coefficients or comparisons in the t-statistics. For the coefficients, the p-value becomes essentially the share placebo coefficient estimates whose absolute value (or square) is greater than the absolute value of the difference using actual lottery assignment.

Similarly, we can calculate the p-value of our original estimate using the t-statistics from each repetition. Accordingly, Young (2016) shows that with M additional draws the p-value of the difference is given by the following:

$$\text{Sampling randomization p-value} = \frac{1}{M+1}\{\sum_{m=1}^{M} I_m(> t_a^2) + U[1 + \sum_{m=1}^{M} I_m(= t_a^2)]\}, \quad (18)$$

where *U* is a random variable drawn from the uniform distribution, $t_a^2 = \left[\frac{\hat{\pi}_a}{se(\hat{\pi}_a)}\right]^2$ is the squared t-statistic from the actual lottery participation decisions and $I_m(> t_a^2)$ and $I_m(= t_a^2)$ are indicator functions for whether the placebo squared t-statistic is larger than t_a^2. These approach avoids possible finite sample bias and applies minimal assumptions or structure to the data, while providing valid and transparent inference.

Another way to approach the issue of external validity 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.

Accordingly, we narrow our test for whether selection differs across those who select into the experiment and those who do not by estimating the following on the no-shows and never-ever-takers:

$$E(Y_i|D_i = 0, T = n, e) = L_i\pi_0^0 + \boldsymbol{X_i\gamma_0^0}. \quad (19)$$

Performing a t-test on our estimate of π_0^0 provides evidence on whether the never-takers are representative of those who do not take the treatment and never enlisted in the experiment.

We can duplicate this analysis on the sample that receives treatment to build differences in the heterogeneous effects into the analysis:

$$E(Y_i|D_i = 1, T = d, l) = L_i\pi_1^0 + \boldsymbol{X_i\gamma_1^0}. \quad (20)$$

Performing a t-test on our estimate of π_1^0 provides evidence on whether the always-takers are representative and whether we may expect the experimental results to generalize to those who never entered the experiment.

Analysis 3

In this section, we conduct a conventional observational analysis of program impact on the treated population. We conduct this analysis with two purposes in mind. First, we compare the observational results with the experimental estimates to explore issues of bias in conventional observational designs where the population of interest may be only those students who voluntarily enroll in the experiment. The observational design is still commonly used by in-house institutional researchers and appears in much of the earlier-published program evaluation studies of first year learning communities, in the context of both voluntary and mandated enrolment.

Secondly, we conduct this within-study design to reflect on the difference in the observational and experimental results in the context of our tests for external validity, where the population of interest extends beyond those who self-selected into the experiment, and the aim of the observational design is to uncover the population average effect of the program rather than the effect of the program for those who selected into the experiment and were moved into the program by way of randomization. Differences in results between the two approaches may originate from either lack of internal validity of the observational approaches or lack of external validity of the RCT or both. We use the results from analysis 2 to provide evidence for the cause of any divergence in results from these two approaches.

If the treated and untreated populations are alike conditional on our set of observed characteristics, observational analyses will produce unbiased causal estimates of the average

program impact over both the self-selected and larger populations of interest. However, if the two groups differ with regard to unobserved characteristics which cannot be controlled for in the analysis and which affect retention prospects – observational methods will lead to biased and generally uninformative estimates.

We first estimate the effect of enrollment in the FYLC on first year retention using unconditional OLS regressions, covariate adjusted OLS regressions, and logit QMLE analysis. This analysis is similar to the analysis used to identify our average intent to treat estimates. The first key difference is that in these analyses we use the full sample of freshman entrants including both the self-selected lottery entrants and those who initially did not apply for the FYLC lottery. The second key difference is that for these analyses, treatment is measured by an indicator for enrollment in the FYLC instead of by an indicator for winning the lottery.

We supplement this analysis by also adding more sophisticated propensity score matching techniques, which are used by Clark and Cundiff (2011), for example, to evaluate the efficacy of a FYLC without random assignment. Accordingly, we estimate both the average treatment effect on the treated by averaging over the difference between the retention of each treated student and the retention of the student in the remaining population who is most similar to the treated student, but did not receive treatment. We also report estimates of the average treatment effect for comparability. Such techniques require "modelling" of the propensity to enter treatment (FYLC) based on observed characteristics. We adopt the standard practice of using logit to estimate these propensity scores. As we are estimating these propensity scores, conducting appropriate inference requires that we account for possible estimation error. We consequently bootstrap the standard errors to account for this issue.

Validity of this and similar techniques require two assumptions; overlap and ignorability (also known as the conditional independence assumption). In our context, the ignorability assumption requires that we possess sufficient information in the control variables such that there would be no expected difference in retention between those who receive treatment and those who do not in the absence of treatment. In our notation, we must maintain the following:

$$E(Y_i|\boldsymbol{x}, D = 1) - E(Y_i|\boldsymbol{x}, D = 0) = 0. \quad (21)$$

Researchers typically cannot directly test whether ignorability is violated. However, the tests we provide in the proceeding section allows us to do just that.

Figure 1: Overlap in the propensity scores by treatment status

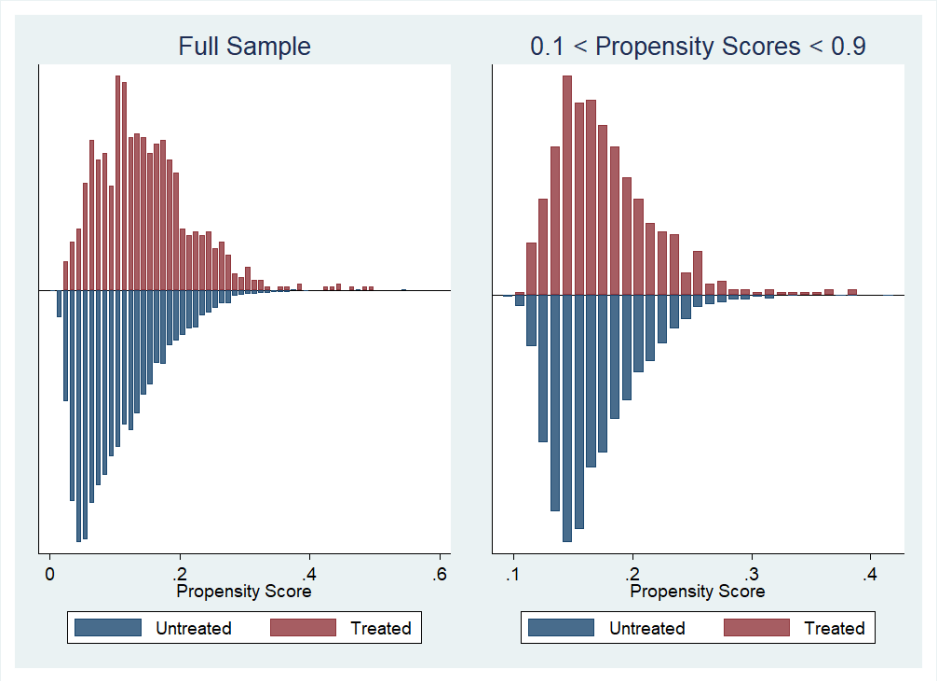


Figure notes: Propensity scores estimated using logit.

The overlap assumption requires that, for any setting of observed characteristics, there is a chance the individual could be in either the treatment or control group. We can examine the overlap assumption through the estimated propensity scores. Figure 1 presents histograms of these estimated propensity scores split by treatment status. Crump, Hotz, Imbens, and Mitnik (2009) provide a rule of thumb that observations with propensity scores above 0.9 and below 0.1 should be discarded. We accordingly perform all analyses both on the full sample as well as this trimmed subsample.

Results

Analysis 1

Table 3: RCT estimates

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

	(1) Retention	(2) Retention	(3) Retention	(4) Retention
Won lottery	0.019	0.018	0.019	0.018
	(0.015)	(0.015)	(0.015)	(0.014)

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

FYLC	0.029	0.027	0.030	0.027
	(0.022)	(0.022)	(0.025)	(0.022)
Residuals			-0.012	-0.004
			(0.032)	(0.029)

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

Won lottery	0.648	0.648	0.648	0.648
	(0.019)	(0.019)	(0.019)	(0.019)
Observations	1565	1565	1565	1565
Retention Mean	0.910	0.910	0.910	0.910
Controls	No	Yes	No	Yes
Model	LPM	LPM	QML	QML

The first two column report results from linear models whereas columns (3) and (4) report estimates from nonlinear estimation. Logit was used in QML estimation. The control function residuals used with QML in panel B were estimated using OLS. Columns (1) and (3) are unconditional estimates whereas columns (2) and (4) include baseline covariates. Robust standard errors in parentheses. Bootstrap standard errors with 500 replications were used for inference in QML control function estimation.

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 exactly the same whether estimated by OLS or logit QML. The quantitative magnitude – a roughly two percentage point increase in the retention probability – is not insubstantial, but the estimates have large standard errors and none are close to being statistically different from zero at any conventional threshold.

Panel B gives the LATE estimates, while Panel C provides the first stage estimates, which reveal that the randomization provides a strong instrumental variable in explaining variation in FYLC participation. The estimated impacts of the program in the second-stage regression analysis increase in quantitative magnitude – by roughly one percentage point – compared to the intent to

treat estimates, but once again these estimates are imprecisely estimated and thus statistically insignificantly different from zero.

The control function residuals in columns 3 and 4 of Panel B preview some of the analysis presented in Analysis 2 below. The coefficient estimates are 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 estimated LATE may generalize to the rest of the experimental population.

Analysis 2

Columns (1) and (2) of Panel A compare the retention probabilities of no-shows and the control 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.

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. Thus, the panel A results suggest that the two migrations within the experimental population do not present problems in estimating the program impact in the RCT design. The ITT and LATE estimates are of inconsequential difference, and the estimated program impact can be comfortably generalized to apply to migrants as well as to assignment compliers.

Table 4: Testing for selection into and within the lottery

	(1)	(2)	(3)	(4)
Panel A: Test for nonrandom attrition and noncompliance with the lottery				
Won	0.009	0.000	0.005	0.005

	(0.025)	(0.026)	(0.029)	(0.029)
Observations	803	803	762	762
Controls	No	Yes	No	Yes
Sample	Control + No-shows	Control + No-shows	Treated + Crossovers	Treated + Crossovers
Treatment status	Untreated	Untreated	Treated	Treated
Panel B: Test for selection into the experiment among the untreated				
Lottery	0.028	0.039	0.035	0.040
	(0.011)	(0.011)	(0.023)	(0.023)
Observations	7252	7252	6619	6619
Controls	No	Yes	No	Yes
Sample	Control + No-shows + Never-ever-takers	Control + No-shows + Never-ever-takers	No-shows + Never-ever-takers	No-shows + Never-ever-takers
Treatment status	Untreated	Untreated	Untreated	Untreated
Panel C: Test for selection into the experiment among the treated				
Lottery	0.067	0.063	0.062	0.062
	(0.034)	(0.033)	(0.042)	(0.045)
Observations	879	879	225	225
Controls	No	Yes	No	Yes
Sample	Treated + Crossovers + Late-takers	Treated + Crossovers + Late -takers	Crossovers + Late -takers	Crossovers + Late -takers
Treatment status	Treated	Treated	Treated	Treated

All results are from OLS regressions. Robust standard errors in parentheses.

In Panels B and C of Table 4, we explore issues of non-random selection into the lottery itself among the untreated and treated populations, respectively. Columns (1) and (2) of Panel B explore the extent to which those who selected into the lottery, but were untreated, differ regarding the probability of retention from the "never-ever takers" (i.e., those who did not select into the lottery and did not later become treated as late-takers). Column (1) provides the unconditional estimates, such that the reported coefficient provides the nonparametric difference in the mean outcomes of the untreated by whether or not they participated in the lottery. Column (2) conditions on predetermined observed student characteristics. While this approach may introduce finite sample bias, it is also generally more efficient (though not noticeably in this case) and focuses attention on the differences on unobserved variables. The fact that the coefficient on Lottery is statistically and economically significantly positive in both specifications indicates that those who enter the lottery are more likely to persist in college

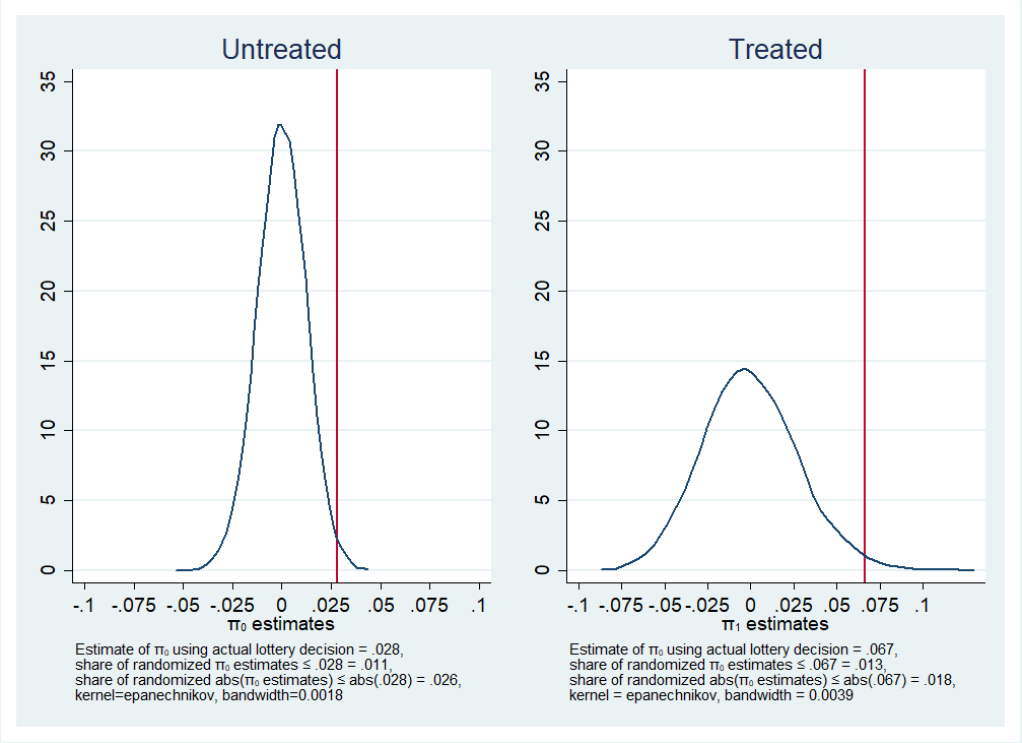
regardless of the program, as neither population in these regressions took part in the FYLC. The fact that the magnitude of the coefficient grows from 0.028 (p-value = 0.013) to 0.039 (p-value = 0.001) with the addition of covariates indicates that lottery participants are negatively selected on observed characteristics – something we presented as a preliminary hypothesis in the comparison of background characteristics across these two populations in the "data" section above. However, the positive selection into the lottery based on unobserved characteristics is more pronounced than the negative selection on observed variables.

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

In Panel C we turn to selection into the lottery among the treated populations. Columns (1) and (2) compare retention probabilities for the treated population that selected into the lottery and those late-takers who expressed no interest in the program initially, but later changed their minds and were admitted into the FYLC. We find that, among the treated, those who entered the lottery are roughly 6 percentage points (p-value = 0.062) more likely to persist than those who came into the program as late crossovers. In columns (3) and (4), a comparison is made between two final treated groups – the crossovers and late-crossovers – both of whom were treated and migrated from initially assigned or chosen positions in order to receive treatment. Once again, the central distinguishing feature of these two groups is the initial decision to participate in the lottery. While the results reveal no statistically significant difference in retention probabilities

across these two groups, owing to either observed or unobserved variables, the quantitative magnitude of the difference owing to unobserved characteristics is very large (equivalent to the estimate in the first two columns). It is possible that the much-reduced sample size may explain the somewhat larger standard errors that render the difference in conditional retention probabilities statistically insignificantly different from zero. Though the results from Panel C are not statistically significantly different from those in Panel B, the fact that the estimated coefficient on the indicator for lottery entrance in Panel C is nearly double that from Panel B is suggestive of differences in heterogeneous effects of the program between populations.

Figure 2: Distribution of placebo $\hat{\pi}$ where "lottery participation" is randomly assigned

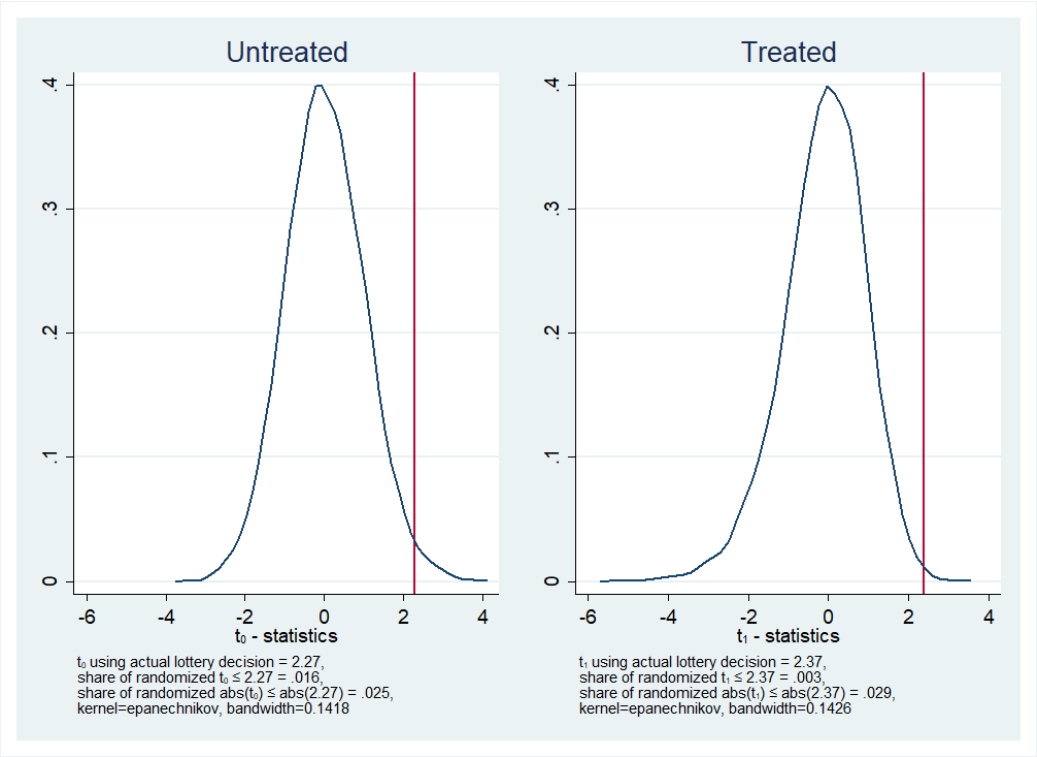


Notes: Binomial random assignment to lottery participation with probabilities of inclusion in the lottery by treatment status set at 0.11 for the untreated and 0.87 for the treated reflecting the shares observed in the data. Distributions constructed from 10,000 repetitions. The red verticle lines denote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

In Figures 2 and 3 we show the distributions of placebo coefficients and corresponding t-statistics from our randomization tests. Figure 2 presents the estimated differences in retention among the untreated (on the left) and among the treated (on the right) when "lottery participation" is randomly assigned is randomly assigned in each of 10,000 repetitions. We show the estimated difference in retention using the actual lottery participation using a red verticle

line. Figure 3 repeats the exercise using the t-statistics on each difference to incorporate the precision of each estimate into the simulation. 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 participation in the experiment are unlikely to result from pure chance.

Figure 3: 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.

We show the formalized p-values from our randomization tests as well the selected distributional values of the placebo distributions in table 5. Rows one and three of table 5 present the nonparametric unconditional differences in retention rates between the experimental and non-experimental populations stratified by treatment status with the accompanying p-values from randomization testing based only on the coefficent estimates. For comparison, the second and third rows show the first, fifth, tenth , fiftieth, nintieth, nity-fifth, and ninty-nineth percentile of

the placebo differences when lottery participation is randomly assigned.[21] Following Young (2016) we repeat the exercise using the t-statistics presented in table 4 compared against the distribution of placebo t-statistics. For all tests and for both the treated and untreated populations, the p-values from the randomization tests fall between 0.015 and 0.03.

Table 5: Nonparametric randomiaztion 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	0.018								
Placebo $\widehat{\pi_0}$			-0.000	-0.030	-0.021	-0.016	-0.000	0.015	0.020	0.029
Treated										
Actual $\widehat{\pi_1}$	0.067	0.026								
Placebo $\widehat{\pi_1}$			-0.000	-0.061	-0.045	-0.035	-0.001	0.035	0.048	0.070
t-statistics										
Untreated										
Actual t_0	2.27	0.025								
Placebo t_0			0.021	-2.239	-1.581	-1.238	-0.007	1.304	1.693	2.509
Treated										
Actual t_1	2.37	0.029								
Placebo t_1			-0.121	-3.019	-1.962	-1.478	-0.047	1.127	1.466	2.020

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 A2 and A3.

As both the treated and untreated populations reveal that those who participate in the experiment are more likely to persist in college than those who do not, we conclude that there is evidence of positive selection on unobserved characteristics into the lottery. The RCT results cannot be reasonably generalized to those who do not self-select into the lottery, regardless of ultimate treatment status. As a result, it seems there is little external validity of this RCT to the remaining population.

[21] The p-vales are constructed using the square of the differences and t-statistics.

Analysis 3

If the evaluation of program impact had not relied on random assignment, but rather had utilized an observational research design, how would the estimated program impact have differed? We generate an observational estimate of program impact for a capacity constrained setting, where the treated, including crossovers and late-takers, are compared to non-participants that include both the control, no-shows, and never-ever-takers.

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.

The results are in Table 6. Contrary to the findings from the RCT design, the estimated coefficient on the treatment variable in the observational analysis is positive and statistically significant regardless of specification or procedure invoked. Moreover, the estimated quantitative impact is large – a roughly 5 percentage point gain in retention probability by virtue of participation in the FYLC. However, because selection into treatment largely transpires through selection into the RCT, we know from the results in Panel B, column (2), of Table 4 that this estimate is biased due to self-selection on unobserved characteristics. For observationally

equivalent students, those who self-select into the lottery have statistically significantly different and quantitatively higher retention rates independent of program participation.

Based on the observed differences among the treated and control populations and the way in which retention probabilities are negatively correlated with those differences, analysts employing such observational analyses might be tempted to hypothesize that the observational results are *underestimates* of true program impact. Indeed, such reasoning underlies the bounds of ATEs proposed in Andrews and Oster (2018). Note that as we restrict the sample to that for which there is more overlap on observed covariates, the observational estimates universally grow. While students who select into the experiment may be vulnerable with regard to observed correlates regarding first year retention, this vulnerability is combined with unobserved characteristics – such as commitment, grit, deep academic engagement, or a healthy work ethic – that more than make up for their observational vulnerabilities.

As a final exercise, and by way of summarizing the empirical findings, we decompose the findings from Table 6 into selection into each of the six populations according to the test for nonrandom noncompliance outlined in Huber (2013). In the first two columns of Table 7, we focus exclusively on the experimental sample as a direct application of Huber (2013). The omitted category in columns (1) and (2) are the control group who entered the lottery, lost the lottery, and did not enroll in the FLYC. 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. Thus, the coefficient on Won indicates the increased retention from just winning the lottery, whereas Won lottery x FYLC reveals the increased retention from winning and also entering the FLYC. The coefficient on Entered lottery x FYLC reveals the marginal change in retention among those who did not win the lottery but managed to make it into the program.

Columns (3) and (4) present the decomposition of the results from Table 6. Here, the full sample is used and we add an indicator for those who entered the lottery as well as an indicator for not entering the lottery but entering the FYLC. 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.

In the aggregate, these results echo those previously presented. Columns (1) and (2) reveal that we can identify no evidence of a causal effect of the FYLC on retention for those who select into the experiment. Further, there are no substantive differences, regarding retention probabilities and thus likely program impact, between the compliers, never-takers, and always-takers. Columns (3) and (4) show that the only statistically meaningful difference in retention prospects is between those who do and do not enter the lottery (p-values of 0.038 and 0.003 respectively). The observed analysis leads to a biased estimate of program impact when the population of interest is those who self-selected into the study.

Table 7: Testing for selection into the lottery, of compliers, and into attrition with interactions

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

All results are from OLS regressions. Robust standard errors in parentheses.

The results also provide evidence that calls into question the generalizability of the LATE estimate of program impact from the RCT design for the non-experimental population. While we observe no detectable selection into or out of treatment within the experimental sample, we do observe selection into the experiment itself on otherwise unobserved propensities to persist in college and possibly in responsiveness to treatment. Thus, it would be unreasonable to generalize the RCT results to those who entered the program late without participating in the lottery or to those who might join as the scale expands.

How do we accordingly assess the experimental and observational approaches? In this instance, we have no means by which to directly assess selection into treatment for those who did not enter the experiment. However, the majority of our treated population received treatment by selecting into the study. As a result, the nonrandom selection that compromises the external validity of the experimental results also contaminates the internal validity of the observational estimates whether the population of interest is the self-selected or a larger segment such as the entire freshman class. Were the experimental sample representative, the two approaches would provide more similar estimates. It would be unreasonable to claim that either estimate – the experimental or the observational – captures the population average effect of the FYLC on first year college retention. Thus, the take away from the difference in observational and experimental results is not the superiority of the experimental, except when the population of interest is limited to the self-selected, but rather, these differences are symptomatic of persistent problems in uncovering the population parameter. It is only by applying tests for external validity that we can determine whether the RCT delivers a valid estimate of this elusive parameter.

Conclusions

This paper introduces new tests for the external validity of RCTs beyond the experimental population. It does so by first utilizing an RCT design to estimate the impact of a learning community on first year college retention for those who select into the study at a large four-year research university. It is the first of its kind of which we are aware. 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.

There were significant migrations from the assigned populations in the experimental sample. In conducting tests from Black et al. (2015) for whether the assignment compliers differ 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 that the "local average treatment effect" estimate of program impact may be safely generalized 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 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 treated population of self-selected students, these results are in no way generalizable to the population who did not enroll in the experiment.

These results serve to highlight a few important broader lessons, all of which emanate from the central insight that selection on unobserved characteristics matters. First in the spirit of Calónico and Smith (2017), our observational analysis makes clear that our seemingly rich set of covariates is insufficient for maintaining the conditional independence assumption. These observational results suggest, incorrectly, that the program led to a large statistically significant increase in first year retention for treated students. Even more interestingly, given the nature of the selection process on observed variables – where students who selected into the study disproportionately possess observed background characteristics that are negatively associated with first year retention – researchers employing observational techniques might be inclined to hypothesize, under a presumption that unobserved differences across the self-selected experimental and non-experimental populations are likely to follow the same pattern as do observed differences, that observational analyses of program impact are underestimates of the true effect of the program. Such is not the case here. As we have seen, our results reveal that the unobserved characteristics of the self-selected control population have a strongly positive effect on retention.

Nonrandom selection affects the assessment of external validity as well as internal validity. In much of the existing research employing an RCT design, testing for selection into the experimental sample has not been conducted, either because the data on nonparticipants do not exist or because researchers have not made use of it. And yet selection issues of various sorts emerge in many of these contexts. Our results suggest that non-random selection based on unobserved characteristics including responsiveness to treatment may matter greatly for the generalizability of the RCT results. Information on non-participants is critical in testing for such non-random selection. Thus, we recommend that when conducting RCTs, researchers should

collect or link their studies to more comprehensive data on the population of interest and show concretely whether the results of their study generalize beyond the experimental sample.

Without further testing researchers ought not to conclude that observational approaches fail when observational estimators yield different results than an RCT. The context matters greatly. Assuming the observational analysis and RCT rely upon samples with different selection processes (e.g., the population from administrative data, a random sample from the population, a researcher non-randomly selected sample, or a participant self-selected sample), the two approaches estimate different parameters. Even if both approaches are internally valid, they may arrive at different results. In which case, the representativeness of the samples determines which estimate provides a closer approximation to the effect in the broader population. The tests that we have introduced provide concrete evidence of where these biases lie and should be incorporated into any similar within-study design.

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

References

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

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of causal effects using instrumental variables." *Journal of the American statistical Association* 91, no. 434: 444-455.

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.

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.

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

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

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

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

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

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

Fairris, D., Castro, M., Son, J. 2010. "Quantitative impacts of learning communities: An analysis of the impact of CHASS Connect on retention and academic performance towards degree completion." *Proceedings, American Institute for Higher Education*: 344-360.

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.

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.

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

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

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

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

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.

Jamelske, Eric. 2009. "Measuring the impact of a university first-year experience program on student GPA and retention." *Higher Education* 57, no. 3: 373-391.

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

Krueger, Alan B. 1999. "Experimental estimates of education production functions." *The quarterly journal of economics* 114, no. 2: 497-532.

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.

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.

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

Matthews, Roberta S., Barbara Leigh Smith, and Jean MacGregor. 2012. "The Evolution of Learning Communities: A Retrospective," in *Discipline-Centered Learning Communities: Creating Connections Among Students, Faculty, and Curricula: New Directions for Teaching and Learning, Number 132,* Kimberly Buch and Kenneth E. Barron, editors. Wiley Periodicals, Inc.

Organization for Economic Cooperation and Development. 1995. *Education at a Glance: OECD Indicators.* OECD: Paris, France

Organization for Economic Cooperation and Development. 2017. *Education at a Glance: OECD Indicators 5.* OECD: Paris, France

Orr, Larry, Judith Feins, Robin Jacob, Eric Beecroft, Lisa Sanbonmatsu, Lawrence F. Katz, Jeffrey B. Liebman, and Jeffrey R. Kling. 2003. "Moving to opportunity: Interim impacts evaluation."

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.

Papke, Leslie E., and Jeffrey M. Wooldridge. 1996. "Econometric methods for fractional response variables with an application to 401 (k) plan participation rates." *Journal of applied econometrics* 11, no. 6: 619-632.

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.

Porter, Stephen R., and Randy L. Swing. 2006. "Understanding how first-year seminars affect persistence." *Research in Higher Education* 47, no. 1: 89-109.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology* 66, no. 5: 688.

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."

Sanbonmatsu, Lisa, Lawrence F. Katz, Jens Ludwig, Lisa A. Gennetian, Greg J. Duncan, Ronald C. Kessler, Emma K. Adam, Thomas McDade, and Stacy T. Lindau. 2011. "Moving to opportunity for fair housing demonstration program: Final impacts evaluation."

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

Shadish, W. R., Cook, T. D., & Campbell, D. T. 2002. Experimental and quasi-experimental designs for generalized causal inference.

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.

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

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.

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.

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

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.

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

Young, Alwyn. 2016. "Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results." *Working Paper*.

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

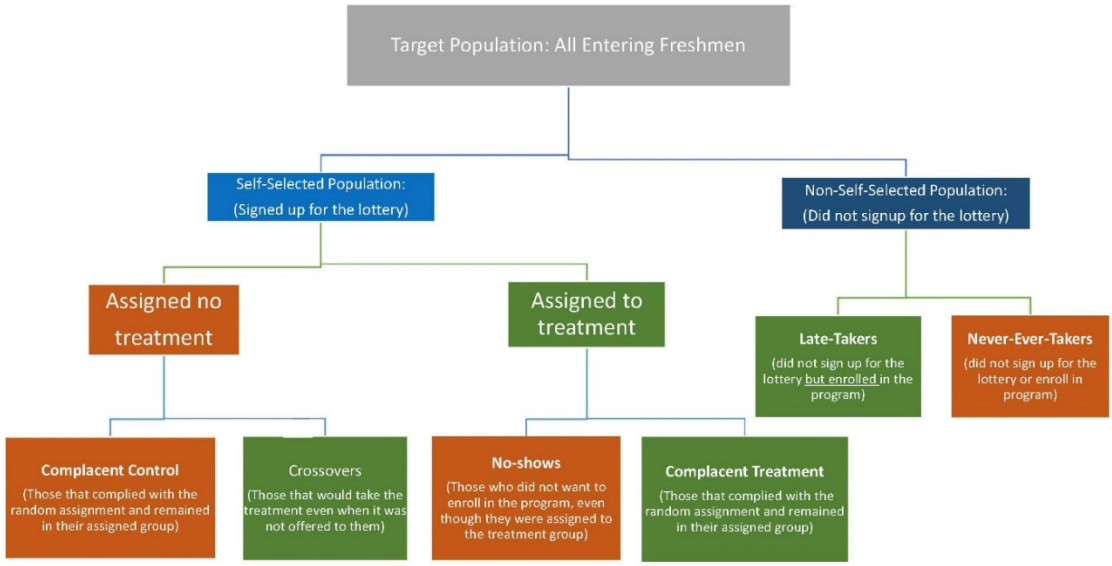
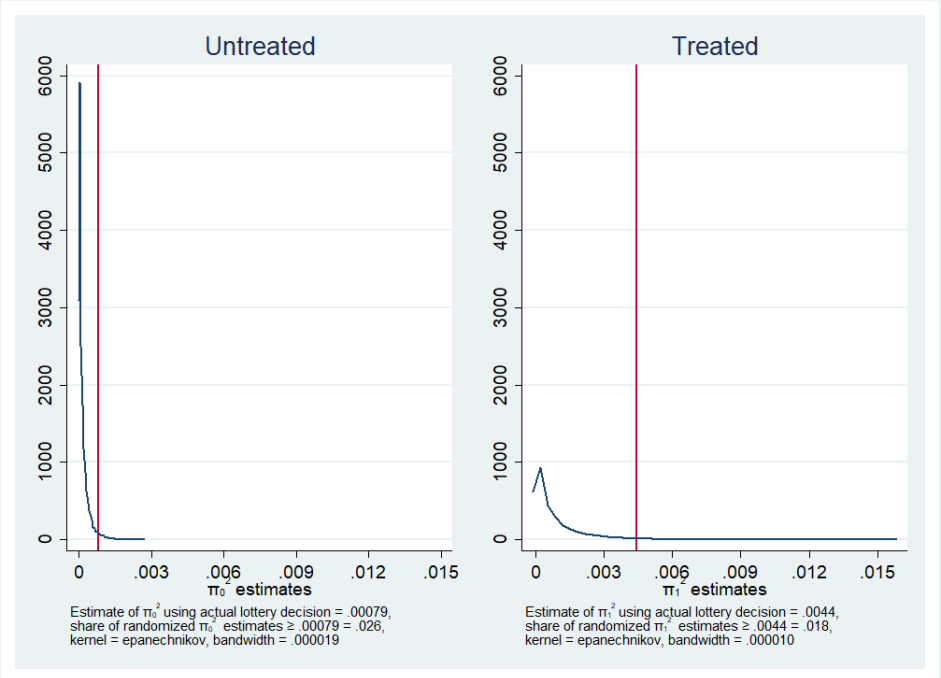


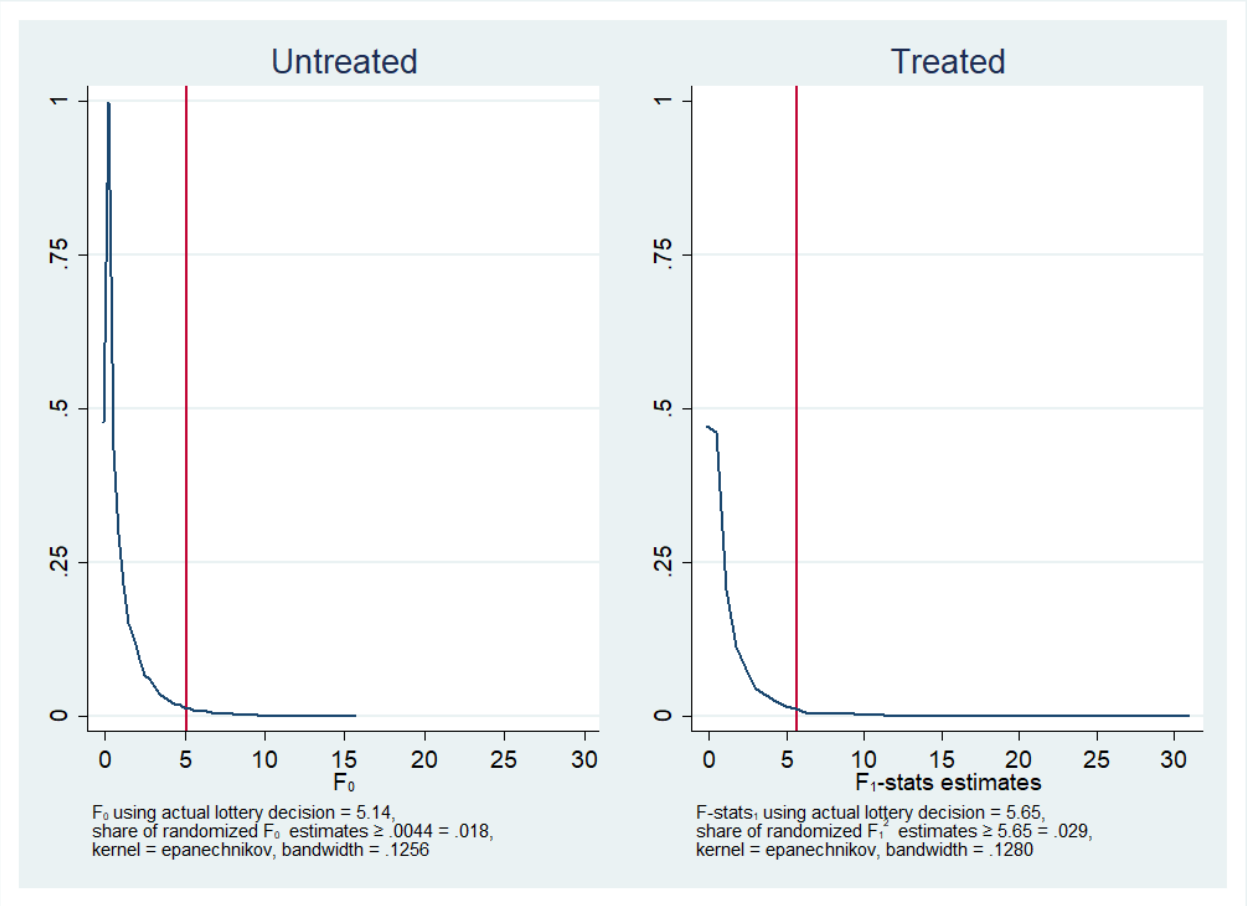
Figure A2: Distribution of squared placebo coefficients



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 A3: Distribution of squared placebo t-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 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)	(0.044)	(0.053)	(0.053)

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

Won lottery	0.648	0.648	0.648	0.648
	(0.019)	(0.019)	(0.019)	(0.019)
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. $p<0.01$, $p<0.05$, $p<0.1$