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

Tarek Azzam, Michael D. Bates, and David Fairris

NBER Education Meetings
April 12, 2019

Why do we conduct empirical research?

Roth (1986) categories motivations for empirical work in economics into the following:

- Speak to theorists: testing predictions from theory
- Searching for facts: uncovering or explaining phenomena
- Whisper in the ear of princes: informing best practices

Why do we conduct empirical research?

Roth (1986) categories motivations for empirical work in economics into the following:

- Speak to theorists: testing predictions from theory
- Searching for facts: uncovering or explaining phenomena
- Whisper in the ear of princes: informing best practices

Why do we conduct empirical research?

Roth (1986) categories motivations for empirical work in economics into the following:

- Speak to theorists: testing predictions from theory
- Searching for facts: uncovering or explaining phenomena
- Whisper in the ear of princes: informing best practices

Internal Validity

- Researchers expend much effort to find an estimate that is **internally valid**: credibly causal and free from bias or inconsistency.
- In seeking internally valid estimates, we find they often hold only for a localized population.
- This issue was highlighted nicely by Imbens and Angrist's influential work on the LATE.

Internal Validity

- Researchers expend much effort to find an estimate that is **internally valid**: credibly causal and free from bias or inconsistency.
- In seeking internally valid estimates, we find they often hold only for a localized population.
- This issue was highlighted nicely by Imbens and Angrist's influential work on the LATE.

Internal Validity

- Researchers expend much effort to find an estimate that is **internally valid**: credibly causal and free from bias or inconsistency.
- In seeking internally valid estimates, we find they often hold only for a localized population.
- This issue was highlighted nicely by Imbens and Angrist's influential work on the LATE.

External Validity

Each category often demands an estimate that is **externally valid**: generalizable to a broader population than that for which the estimates apply.

- Speak to theorists
- Searching for facts
- Whisper in the ear of princes

Threats to external validity may come from noncompliance within the experiment.

Or the experimental sample may differ from the broader population.

Exemplary examples

- Perry Preschool: researchers selected children from poor households in nearby neighborhoods where the kids' IQs ranged from 70 - 85.
- Oregon Health Insurance Experiment: Participants selected from medicaid wait list into which applicants selected into during a 5 week window.
- Moving to Opportunity: Housing vouchers randomly assigned among those who volunteered for the program.
- More broadly:
  - Many drug trials exclude populations (female participants particularly those of child bearing years.)
  - IRB requires almost all participants to consent.

Exemplary examples

- Perry Preschool: researchers selected children from poor households in nearby neighborhoods where the kids' IQs ranged from 70 - 85.
- Oregon Health Insurance Experiment: Participants selected from medicaid wait list into which applicants selected into during a 5 week window.
- Moving to Opportunity: Housing vouchers randomly assigned among those who volunteered for the program.
- More broadly:
  - Many drug trials exclude populations (female participants particularly those of child bearing years.)
  - IRB requires almost all participants to consent.

Exemplary examples

- Perry Preschool: researchers selected children from poor households in nearby neighborhoods where the kids' IQs ranged from 70 - 85.
- Oregon Health Insurance Experiment: Participants selected from medicaid wait list into which applicants selected into during a 5 week window.
- Moving to Opportunity: Housing vouchers randomly assigned among those who volunteered for the program.
- More broadly:
  - Many drug trials exclude populations (female participants particularly those of child bearing years.)
  - IRB requires almost all participants to consent.

Exemplary examples

- Perry Preschool: researchers selected children from poor households in nearby neighborhoods where the kids' IQs ranged from 70 - 85.
- Oregon Health Insurance Experiment: Participants selected from medicaid wait list into which applicants selected into during a 5 week window.
- Moving to Opportunity: Housing vouchers randomly assigned among those who volunteered for the program.
- More broadly:
  - Many drug trials exclude populations (female participants particularly those of child bearing years.)
  - IRB requires almost all participants to consent.

Recent literature examining external validity

Heterogeneity in the literature in addressing external validity: Many show balance tests based on observed variables, though others provide no formal examination of the issue.

- Andrews and Oster (2019) model the decision to participate in a study with observed covariates, and approximate weights using observed variables to provide bounds on the ATE.
- In estimating the demand for charter schools, Walters (2018) also estimates the unobserved preference for charter schools and relates these estimates to effect heterogeneity.
- Ghanem, Hirshleifer, and Ortiz-Becerra (2018) develop tests for nonrandom attrition on baseline values of the outcome.

Related literature testing on unobserved variables

- Huber (2013), Black et al. (2017) and Kowalski (2018) provide tests for selection on unobserved variables into the compliers from within the study sample.
- Sianesi (2017) proposes non-parametric tests (similar to some of those we propose for the untreated) to assess the bias from randomization itself under the assumption of homogeneous responsiveness to treatment.

What we do

1. Provide straightforward tests for nonrandom selection into RCTs on unobserved characteristics and responsiveness to treatment, and provide assumptions for interpretation pertaining to external validity.
2. Evaluate the efficacy of a learning community program in increasing first year college student retention.
3. Test for nonrandom selection to examine the external validity of our RCT results.
4. Perform a "within-study design," comparing the experimental results to those from standard observational approaches.
5. Reflect on within study designs considering the internal and external validity of both approaches.

What we do

1. Provide straightforward tests for nonrandom selection into RCTs on unobserved characteristics and responsiveness to treatment, and provide assumptions for interpretation pertaining to external validity.
2. Evaluate the efficacy of a learning community program in increasing first year college student retention.
3. Test for nonrandom selection to examine the external validity of our RCT results.
4. Perform a "within-study design," comparing the experimental results to those from standard observational approaches.
5. Reflect on within study designs considering the internal and external validity of both approaches.

What we do

1. Provide straightforward tests for nonrandom selection into RCTs on unobserved characteristics and responsiveness to treatment, and provide assumptions for interpretation pertaining to external validity.
2. Evaluate the efficacy of a learning community program in increasing first year college student retention.
3. Test for nonrandom selection to examine the external validity of our RCT results.
4. Perform a "within-study design," comparing the experimental results to those from standard observational approaches.
5. Reflect on within study designs considering the internal and external validity of both approaches.

What we do

1. Provide straightforward tests for nonrandom selection into RCTs on unobserved characteristics and responsiveness to treatment, and provide assumptions for interpretation pertaining to external validity.
2. Evaluate the efficacy of a learning community program in increasing first year college student retention.
3. Test for nonrandom selection to examine the external validity of our RCT results.
4. Perform a "within-study design," comparing the experimental results to those from standard observational approaches.
5. Reflect on within study designs considering the internal and external validity of both approaches.

What we do

1. Provide straightforward tests for nonrandom selection into RCTs on unobserved characteristics and responsiveness to treatment, and provide assumptions for interpretation pertaining to external validity.
2. Evaluate the efficacy of a learning community program in increasing first year college student retention.
3. Test for nonrandom selection to examine the external validity of our RCT results.
4. Perform a "within-study design," comparing the experimental results to those from standard observational approaches.
5. Reflect on within study designs considering the internal and external validity of both approaches.

College Retention

npr KPCC

SIGN IN SHOP DONATE

NEWS ARTS & LIFE MUSIC SHOWS & PODCASTS SEARCH

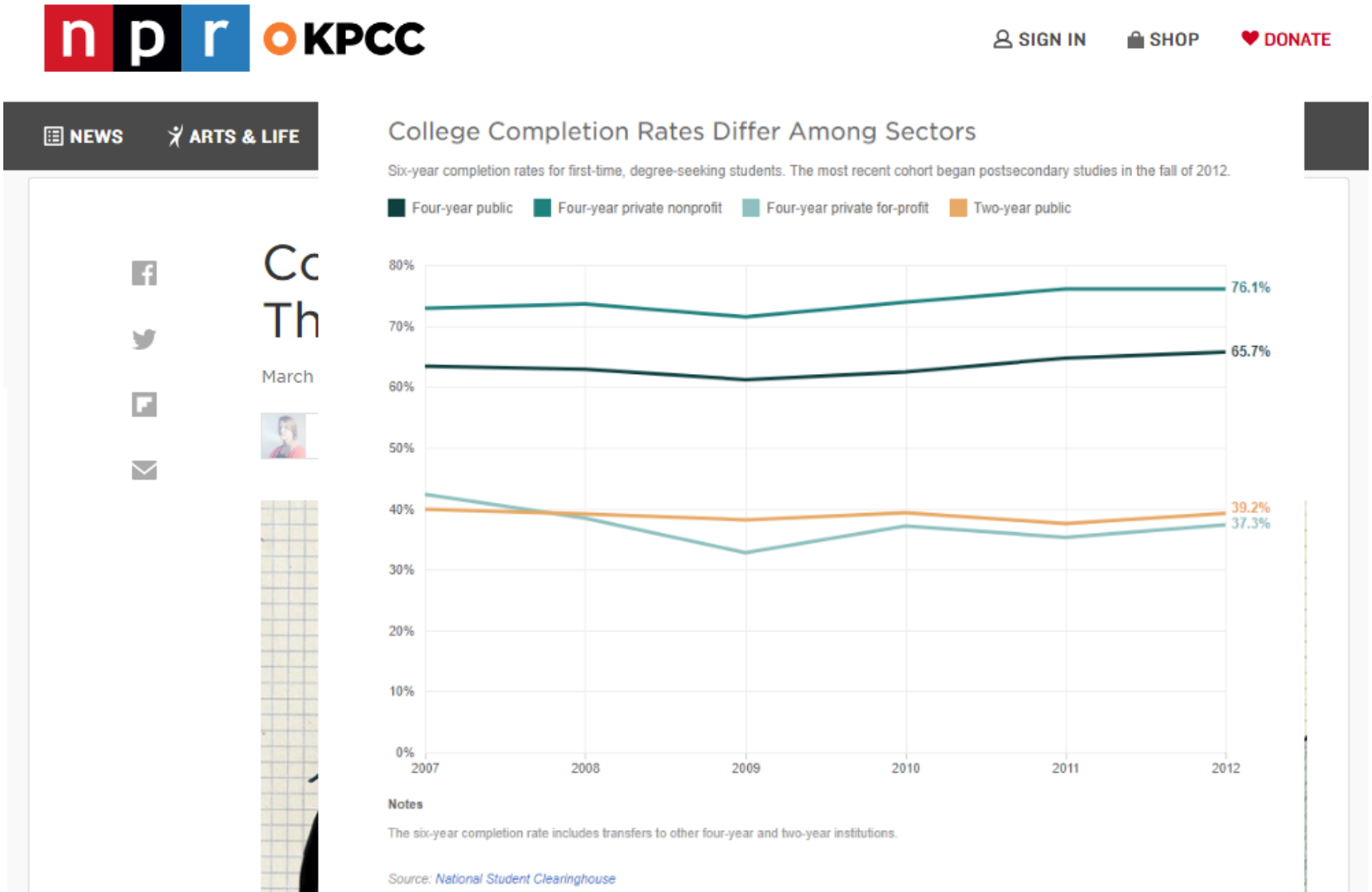
College Completion Rates Are Up, But The Numbers Will Still Surprise You

March 13, 2019 · 7:01 AM ET

ELISSA NADWORNY



College Retention



First Year Learning Communities (FYLCs)

- More than half of the attrition from college occurs before the second year.
- Educational researchers have identified First Year Learning Communities (FYLCs) as central to combating attrition from college.
- 91 percent of institutions report having some sort of FYLC on their campus.

Related Literature: First Year Learning Communities (FYLCs)

- Large education literature such as Clark and Cundiff (2011), which uses propensity score matching as evidence of large effects of FYLCs.
- Scrivener et al. (2008) and Weiss (2012) find positive effects of FYLCs in community colleges.
- Russell (2017) finds positive effects of FLYCs on the GPA of female students at MIT, though no overall effects.

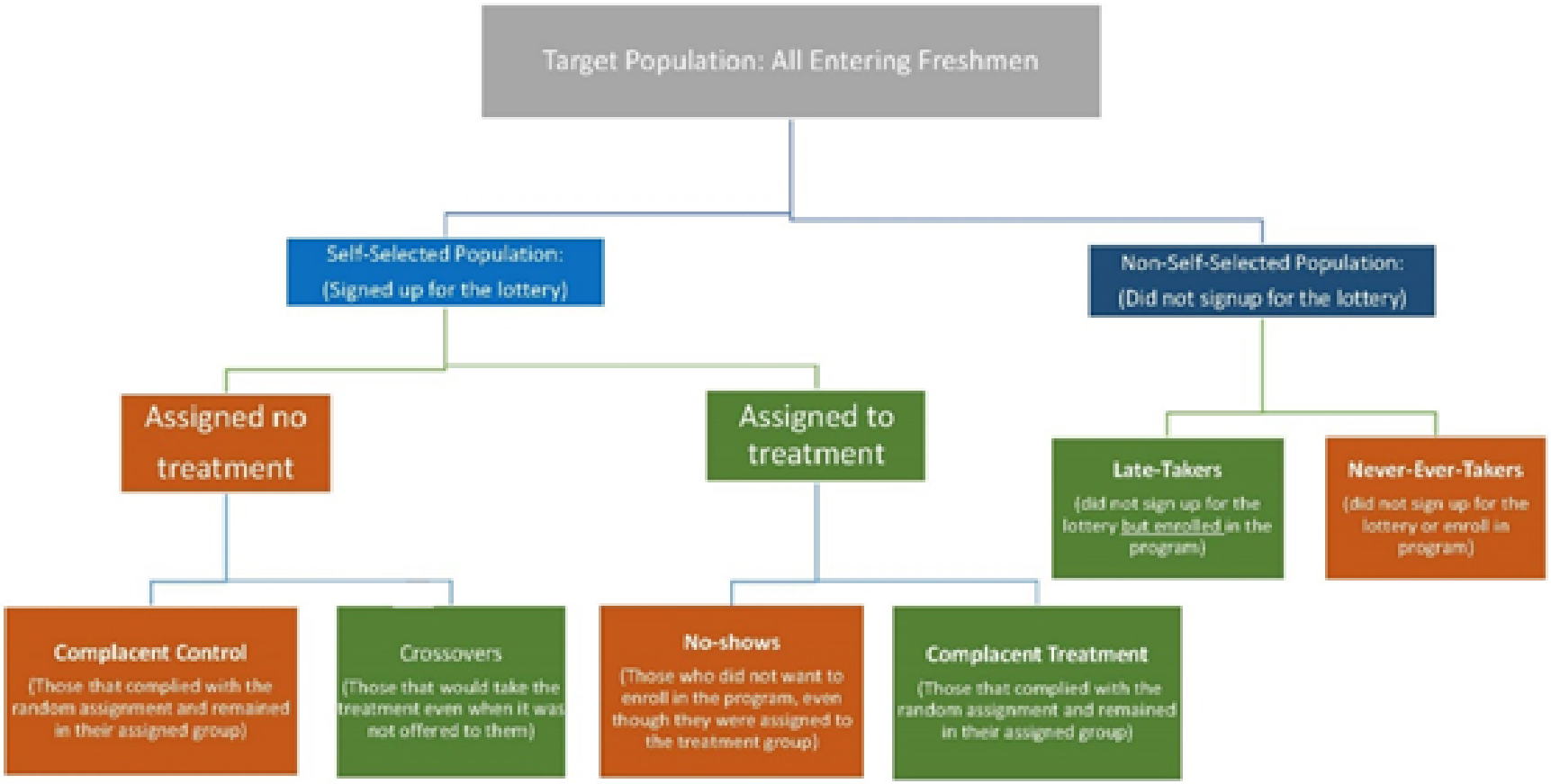
▸ Preview Results

Describing the Treatment

- Students enroll in a 3 quarter writing-intensive, theme-based sequence with the same students.
- Students were required to attend organized study sessions.
- Enrolled in peer mentoring.
- Engaged in additional extra-curricular activities.
- In subsequent years the program has been scaled up to 90% of entering freshman in the college.

▸ Population of Interest

Recruitment, Randomization, and Compliance



Participants are negatively selected on observed characteristics

Table 1: Student Background Characteristics

	Assigned Control	Assigned Treatment	Difference	Lottery Sample	Non-lottery Sample	Difference
High-school GPA	3.46	3.46	0.01	3.46	3.53	-0.07***
			(0.02)			(0.01)
SAT math	494.25	498.65	4.4	496.57	544.4	-47.83***
			(6.15)			(3.63)
SAT writing	491.42	496.4	4.98	494.04	508.33	-14.29***
			(5.77)			(3.29)
SAT verbal	488	491.14	3.14	489.65	502.39	-12.73***
			(5.88)			(3.31)
Female	0.68	0.69	0.01	0.69	0.5	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	0.6	0.62	0.01	0.61	0.56	0.05***
			(0.02)			(0.01)
Lives on Campus	0.74	0.75	0.01	0.75	0.71	0.04***
			(0.02)			(0.01)
N	741	824	1565	1565	6566	8131

Low income is defined as family income below $30,000. Robust standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Model

Consider the following model:

$$y_i = \mathbf{X_i}\gamma + b_i D_i + \epsilon_i,$$

- y_i is an indicator for persisting in college.
- D_i stands for treatment status.
- $b_i = \beta + e_i$ where e_i is the heterogeneous responsiveness to treatment.
- ϵ_i represents students unobserved propensity to persist.
- Typically we want to identify β.
- $\mathbf{X_i}$ is a vector of observed covariates.

What do we get in an RCT?

For our experimental sample we can estimate:

$$y_i = \mathbf{X_i}\gamma + Z_i\delta^* + \epsilon_i,$$

- Z_i is winning the lottery.
- But, we have noncompliance.
- δ^* is the effect of winning the lottery among those who enter the lottery (often called the **"intent to treat"**).
- Estimated with OLS and Logit.
- This isn't usually what we want.

RCT LATE

Or we can estimate:

$$y_i = \mathbf{X_i}\gamma + b_{LATE}\hat{D}_i + \epsilon_i,$$

- $\hat{D}_i$ is the predicted treatment status from the lottery.
- $D_i = \mathbf{X_i}\gamma_\mathbf{1} + \gamma_2 Z_i + u_i,$
- b_{LATE} is the effect of treatment for those whose treatment status is determined by the lottery.

Populations within the data

Name	Conditional outcome	Composition
Complacent Treatment	$E(Y_i \mid \mathbf{x}, D=1, L=1, Z=1) = \mathbf{x}\boldsymbol{\gamma} + \beta + \overline{e_t} + \overline{\epsilon_t}$	Compliers and always-takers
Complacent Control	$E(Y_i \mid \mathbf{x}, D=0, L=1, Z=0) = \mathbf{x}\boldsymbol{\gamma} + \overline{\epsilon_c}$	Compliers and never-takers
No-shows	$E(Y_i \mid \mathbf{x}, D=0, L=1, Z=1) = \mathbf{x}\boldsymbol{\gamma} + \overline{\epsilon_n s}$	Never-takers
Crossovers	$E(Y_i \mid \mathbf{x}, D=1, L=1, Z=0) = \mathbf{x}\boldsymbol{\gamma} + \beta + \overline{e_c o} + \overline{\epsilon_c o}$	Always-takers
Never-ever-takers	$E(Y_i \mid \mathbf{x}, D=0, L=0, Z=0) = \mathbf{x}\boldsymbol{\gamma} + \overline{\epsilon_n}$	Never-ever-takers
Late-takers	$E(Y_i \mid \mathbf{x}, D=1, L=0, Z=0) = \mathbf{x}\boldsymbol{\gamma} + \beta + \overline{e_l} + \overline{\epsilon_l}$	Late-takers

Testing for problematic noncompliance within the experiment

Following Huber (2013) and Black et al. (2017), Among the untreated, test H_0 that $\hat{\pi_{01}} = 0$ in the following:

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

Among the treated, test H_0 that $\hat{\pi_{11}} = 0$ in the following:

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

Rejection of the null indicates problematic selection on unobserved covariates into compliance, Kowalski (2018) shows under what conditions failure to reject of the nulls indicates a lack of external validity.

Intuitive parametric tests for selection

Among the Untreated

$$E(y_i|\mathbf{x}, D_i = 0) = \mathbf{X_i}\boldsymbol{\gamma_0} + L_i\pi_0$$

A substantially or significantly non-zero $\hat{\pi_0}$ provides evidence of selection on unobserved characteristics into the experiment.

Among the Treated

$$E(y_i|\mathbf{x}, D_i = 1) = \mathbf{X_i}\boldsymbol{\gamma_1} + L_i\pi_1$$

A substantially or significantly non-zero $\hat{\pi_1}$ provides evidence of selection on unobserved characteristics and/or responsiveness to treatment, making a claim of external validity difficult to accept.

What we have and what we would like to know

What do these tests tell us about selection into the experiment?

What do these tests tells us about external validity?

- External validity of an experiment implies: $E(Y_1|L=1) - E(Y_0|L=1) = E(Y_1|L=0) - E(Y_0|L=0)$.
- We would like to know whether $E(Y_0|L=1) = E(Y_0|L=0)$ and $E(Y_1|L=1) = E(Y_1|L=0)$.
- Our data only contains $E(Y|L=1, Z=1, D=1)$, $E(Y|L=1, Z=0, D=1)$, $E(Y|L=1, Z=1, D=0)$, $E(Y|L=1, Z=0, D=0)$, $E(Y|L=0, D=1)$, *and* $E(Y|L=0, D=0)$.

Selection into the experiment: simple case

Simple Case: Suppose compliance within the experiment is perfect and non-experimental treatment assignment is homogeneous.

- Maintaining **proper randomization** and **no independent effect of random assignment** provides
  $E(Y|L=1,D=0,Z=0) = E(Y_0|L=1,D=0,Z=0) = E(Y_0|L=1,D=0) = E(Y_0|L=1) = E(Y|L=1,D=0)$,
  and the similar among the treated.
- Under **homogeneous treatment assignment** among **non-experimental** group,
  $E(Y|L=0,D=0) = E(Y_0|L=0)$ or
  $E(Y|L=0,D=1) = E(Y_1|L=0)$

Selection into the experiment: simple case

Simple Case: Suppose compliance within the experiment is perfect and non-experimental treatment assignment is homogeneous.

- Maintaining **proper randomization** and **no independent effect of random assignment** provides
  $E(Y|L=1,D=0,Z=0) = E(Y_0|L=1,D=0,Z=0) = E(Y_0|L=1,D=0) = E(Y_0|L=1) = E(Y|L=1,D=0)$,
  and the similar among the treated.
- Under **homogeneous treatment assignment** among **non-experimental** group,
  $E(Y|L=0,D=0) = E(Y_0|L=0)$ or
  $E(Y|L=0,D=1) = E(Y_1|L=0)$

Selection into the experiment: simple case

Simple Case: Suppose compliance within the experiment is perfect and non-experimental treatment assignment is homogeneous.

- Maintaining **proper randomization** and **no independent effect of random assignment** provides
  $E(Y|L=1,D=0,Z=0) = E(Y_0|L=1,D=0,Z=0) = E(Y_0|L=1,D=0) = E(Y_0|L=1) = E(Y|L=1,D=0)$,
  and the similar among the treated.
- Under **homogeneous treatment assignment** among **non-experimental** group,
  $E(Y|L=0,D=0) = E(Y_0|L=0)$ or
  $E(Y|L=0,D=1) = E(Y_1|L=0)$

Selection into the experiment: simple case

Simple Case: Suppose compliance within the experiment is perfect and non-experimental treatment assignment is homogeneous.

- Thus, testing $E(Y|L=1,D=0)$ vs. $E(Y|L=0,D=0)$ directly **tests selection into the experiment** on the **potential level of the outcome** under no treatment.
- Testing $E(Y|L=1,D=1)$ vs. $E(Y|L=0,D=1)$ directly tests selection into the experiment on the **potential level of outcome and responsiveness to treatment**.
- In either case, external validity is not directly tested, but presuming external validity in the presence of significant evidence of selection would I believe be unreasonable.

Selection into the experiment: simple case

Simple Case: Suppose compliance within the experiment is perfect and non-experimental treatment assignment is homogeneous.

- Thus, testing $E(Y|L=1, D=0)$ vs. $E(Y|L=0, D=0)$ directly **tests selection into the experiment** on the **potential level of the outcome** under no treatment.
- Testing $E(Y|L=1, D=1)$ vs. $E(Y|L=0, D=1)$ directly tests selection into the experiment on the **potential level of outcome and responsiveness to treatment**.
- In either case, external validity is not directly tested, but presuming external validity in the presence of significant evidence of selection would I believe be unreasonable.

Selection into the experiment: simple case

Simple Case: Suppose compliance within the experiment is perfect and non-experimental treatment assignment is homogeneous.

- Thus, testing $E(Y|L=1, D=0)$ vs. $E(Y|L=0, D=0)$ directly **tests selection into the experiment** on the **potential level of the outcome** under no treatment.
- Testing $E(Y|L=1, D=1)$ vs. $E(Y|L=0, D=1)$ directly tests selection into the experiment on the **potential level of outcome and responsiveness to treatment**.
- In either case, external validity is not directly tested, but presuming external validity in the presence of significant evidence of selection would I believe be unreasonable.

Selection into the experiment: simple case

Simple Case: Suppose compliance within the experiment is perfect and non-experimental treatment assignment is homogeneous.

- Thus, testing $E(Y|L=1, D=0)$ vs. $E(Y|L=0, D=0)$ directly **tests selection into the experiment** on the **potential level of the outcome** under no treatment.
- Testing $E(Y|L=1, D=1)$ vs. $E(Y|L=0, D=1)$ directly tests selection into the experiment on the **potential level of outcome and responsiveness to treatment**.
- In either case, external validity is not directly tested, but presuming external validity in the presence of significant evidence of selection would I believe be unreasonable.

Assumptions for external validity in the case at hand

Noncompliance within the experiment and heterogeneous treatment assignment in the non-experimental population require two additional assumptions:

1. Noncompliance with the randomization is ignorable, i.e., not jointly related to treatment and the outcome:
   $E(Y|L=1,D=0) = E(Y_0|L=1)$ and
   $E(Y|L=1,D=1) = E(Y_1|L=1)$
2. Monotonic selection by potential outcome: if
   $E(Y_0|L=0,D=1) >> E(Y_0|L=0,D=0)$, then
   $E(Y_1|L=0,D=1) \geq E(Y_1|L=0,D=0)$.

Selection into the experiment: case at hand

Let $P = E(D|L = 0)$ and suppose the following:

$$E(Y|L = 1, D = 0) - E(Y|L = 0, D = 0) >> 0, \qquad (1)$$

and

$$E(Y|L = 1, D = 1) - E(Y|L = 0, D = 1) >> 0. \qquad (2)$$

What do equations 1) and 2) tell us?

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

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

Testing unobserved heterogeneity into experimental sample: nonparametric approach

We compare mean outcomes between those who entered and did not enter the experiment within each treatment status.

Among **Untreated**, null hypothesis 1 :
$E(Y|L = 1, D = 0) - E(Y|L = 0, D = 0) = 0$

Among **Treated**, null hypothesis 2:
$E(Y|L = 1, D = 1) - E(Y|L = 0, D = 1) = 0$

Following Efron (1982), we use randomization inference, whereby we randomly assign lottery participation and obtain the p-values from the share of differences in coefficients (or t-statistics of those differences) that are more extreme than those realized in the true lottery participation decisions.

▸ Parametric approach ▸ Pairwise tests ▸ Within study design

RCT Results

Table 2: RCT estimates

Panel A: Intent to treat effects on retention				
	(1)	(2)	(3)	(4)
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 retention				
FYLC	0.029	0.027	0.03	0.027
	(0.022)	(0.022)	(0.025)	(0.022)
Residuals			-0.012	-0.004
			(0.032)	(0.029)
Observations	1565	1565	1565	1565
Retention Mean	0.91	0.91	0.91	0.91
Controls	No	Yes	No	Yes
Model	LPM	LPM	QML	QML

Robust standard errors in parentheses. Logit used in QML estimation. *** p¡0.01, ** p¡0.05, * p¡0.1

▸ 1st stage ▸ GPA results

No evidence of selection within the experimental sample

Table 3: Testing selection within the experimental sample

	(1)	(2)	(3)	(4)
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

Robust standard errors in parentheses. OLS used in all specifications. *** p¡0.01, ** p¡0.05, * p¡0.1

Selection into the experiment (untreated)

Table 4: Testing selection into the experiment among the untreated

	(1)	(2)	(3)	(4)
Lottery	0.028**	0.039***	0.035	0.040*
	(0.011)	(0.011)	(0.023)	(0.023)
Observations	7258	7258	6625	6625
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

Robust standard errors in parentheses. OLS used in all specifications. *** p¡0.01, ** p¡0.05, * p¡0.1

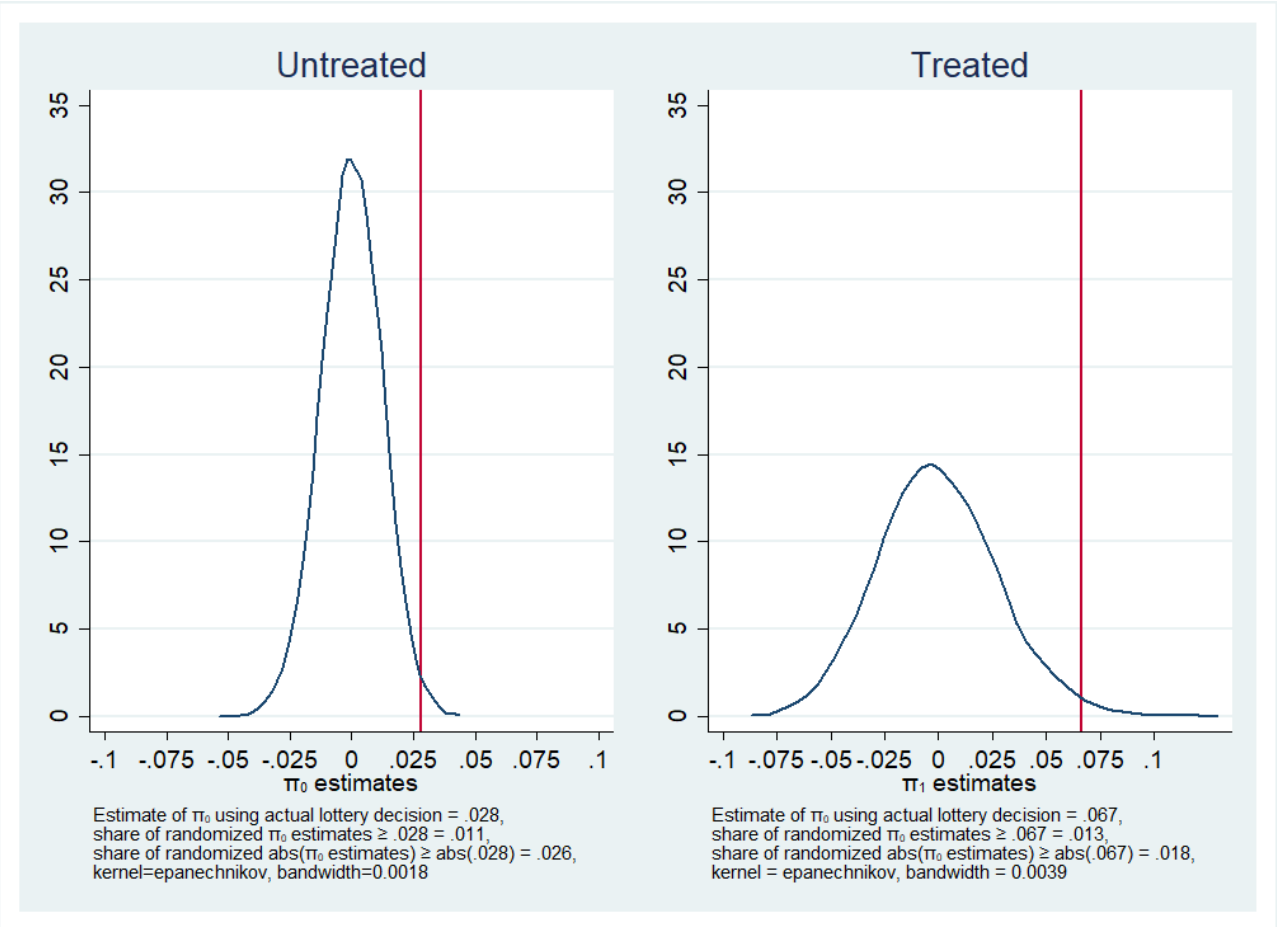
Selection into the experiment (treated)

Table 5: Testing selection into the experiment among the treated

	(1)	(2)	(3)	(4)
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 Crossovers	Treated + Crossovers + Late Crossovers	Crossovers + Late Crossovers	Crossovers + Late Crossovers
Treatment status	Treated	Treated	Treated	Treated

Robust standard errors in parentheses. OLS used in all specifications. *** p¡0.01, ** p¡0.05, * p¡0.1

Selection into experimental: randomization inference



Notes: Distributions constructed from 10,000 repetitions.

▸ t-test ▸ F-test ▸ C2-test

Within-study design analysis: Large observational positive and significant effects

Table 6: Observational analysis estimates of program effects.

	(1)	(2)	(3)	(4)	(5)
FYLP	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	0.88	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. For PSM we present the estimated average treatment on the treated as well as estimates of the ATE. *** p¡0.01, ** p¡0.05, * p¡0.1.

Even larger effects on subsample on which there is significant overlap

Table 7: Observational analysis on restricted sample

	(1)	(2)	(3)	(4)	(5)
FYLP	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	0.87	0.87	0.87	0.87	0.87
Controls	No	Yes	Yes	Yes	Yes
Estimation	OLS	OLS	Logit	PSM ATT	PSM ATE

The restricted sample uses only observation for which there is overlap with propensity scores greater than 0.1 and less than 0.9. *** p¡0.01, ** p¡0.05, * p¡0.1

▸ Decomposition

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

	(1)	(2)	(3)	(4)
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)	(0.03)	(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

▸ Conclusion

Empirical Findings

- Cannot conclude FYLC increased retention for students who complied with the randomization.
- The results seem to generalize throughout the experimental sample.
- Cannot say how it would have impacted those who did not select into the experiment.

Empirical Findings

- Cannot conclude FYLC increased retention for students who complied with the randomization.
- The results seem to generalize throughout the experimental sample.
- Cannot say how it would have impacted those who did not select into the experiment.

Empirical Findings

- Cannot conclude FYLC increased retention for students who complied with the randomization.
- The results seem to generalize throughout the experimental sample.
- Cannot say how it would have impacted those who did not select into the experiment.

Takeaways

- Define the population of interest.
- Balance tests on observed characteristics are insufficient for showing balance on unobserved characteristics.
- Platinum standard? Tie RCTs to random sample of or administrative data containing the population of interest and test selection into the RCT.
- Caution against using experimental results as the standard for comparison without first assessing the external validity of the experiment.

▸ Population of Interest

Takeaways

- Define the population of interest.
- Balance tests on observed characteristics are insufficient for showing balance on unobserved characteristics.
- Platinum standard? Tie RCTs to random sample of or administrative data containing the population of interest and test selection into the RCT.
- Caution against using experimental results as the standard for comparison without first assessing the external validity of the experiment.

▸ Population of Interest

Takeaways

- Define the population of interest.
- Balance tests on observed characteristics are insufficient for showing balance on unobserved characteristics.
- Platinum standard? Tie RCTs to random sample of or administrative data containing the population of interest and test selection into the RCT.
- Caution against using experimental results as the standard for comparison without first assessing the external validity of the experiment.

▸ Population of Interest

Takeaways

- Define the population of interest.
- Balance tests on observed characteristics are insufficient for showing balance on unobserved characteristics.
- Platinum standard? Tie RCTs to random sample of or administrative data containing the population of interest and test selection into the RCT.
- Caution against using experimental results as the standard for comparison without first assessing the external validity of the experiment.

▸ Population of Interest

Preview of empirical findings

- Both intent-to-treat and LATE reflect no discernible effect learning communities on first year retention.
- No discernible evidence that noncompliance is problematic.
  - RCT results seem to generalize within study sample.
- RCT results do not seem to generalize to the remaining freshman class.
  - The experimental sample is **negatively** selected on the basis of **observed** characteristics.
  - The experimental sample is **positively** selected on the basis of **unobserved** propensity to persist in college.
- The null RCT results are contradicted by large significantly positive observational results.
  - This divergence in results is due to the selection into the experiment on unobserved variables.

Preview of empirical findings

- Both intent-to-treat and LATE reflect no discernible effect learning communities on first year retention.
- No discernible evidence that noncompliance is problematic.
  - RCT results seem to generalize within study sample.
- RCT results do not seem to generalize to the remaining freshman class.
  - The experimental sample is **negatively** selected on the basis of **observed** characteristics.
  - The experimental sample is **positively** selected on the basis of **unobserved** propensity to persist in college.
- The null RCT results are contradicted by large significantly positive observational results.
  - This divergence in results is due to the selection into the experiment on unobserved variables.

Preview of empirical findings

- Both intent-to-treat and LATE reflect no discernible effect learning communities on first year retention.
- No discernible evidence that noncompliance is problematic.
  - RCT results seem to generalize within study sample.
- RCT results do not seem to generalize to the remaining freshman class.
  - The experimental sample is **negatively** selected on the basis of **observed** characteristics.
  - The experimental sample is **positively** selected on the basis of **unobserved** propensity to persist in college.
- The null RCT results are contradicted by large significantly positive observational results.
  - This divergence in results is due to the selection into the experiment on unobserved variables.

Preview of empirical findings

- Both intent-to-treat and LATE reflect no discernible effect learning communities on first year retention.
- No discernible evidence that noncompliance is problematic.
  - RCT results seem to generalize within study sample.
- RCT results do not seem to generalize to the remaining freshman class.
  - The experimental sample is **negatively** selected on the basis of **observed** characteristics.
  - The experimental sample is **positively** selected on the basis of **unobserved** propensity to persist in college.
- The null RCT results are contradicted by large significantly positive observational results.
  - This divergence in results is due to the selection into the experiment on unobserved variables.

Preview of empirical findings

- Both intent-to-treat and LATE reflect no discernible effect learning communities on first year retention.
- No discernible evidence that noncompliance is problematic.
  - RCT results seem to generalize within study sample.
- RCT results do not seem to generalize to the remaining freshman class.
  - The experimental sample is **negatively** selected on the basis of **observed** characteristics.
  - The experimental sample is **positively** selected on the basis of **unobserved** propensity to persist in college.
- The null RCT results are contradicted by large significantly positive observational results.
  - This divergence in results is due to the selection into the experiment on unobserved variables.

Preview of empirical findings

- Both intent-to-treat and LATE reflect no discernible effect learning communities on first year retention.
- No discernible evidence that noncompliance is problematic.
  - RCT results seem to generalize within study sample.
- RCT results do not seem to generalize to the remaining freshman class.
  - The experimental sample is **negatively** selected on the basis of **observed** characteristics.
  - The experimental sample is **positively** selected on the basis of **unobserved** propensity to persist in college.
- The null RCT results are contradicted by large significantly positive observational results.
  - This divergence in results is due to the selection into the experiment on unobserved variables.

Preview of empirical findings

- Both intent-to-treat and LATE reflect no discernible effect learning communities on first year retention.
- No discernible evidence that noncompliance is problematic.
  - RCT results seem to generalize within study sample.
- RCT results do not seem to generalize to the remaining freshman class.
  - The experimental sample is **negatively** selected on the basis of **observed** characteristics.
  - The experimental sample is **positively** selected on the basis of **unobserved** propensity to persist in college.
- The null RCT results are contradicted by large significantly positive observational results.
  - This divergence in results is due to the selection into the experiment on unobserved variables.

Preview of empirical findings

- Both intent-to-treat and LATE reflect no discernible effect learning communities on first year retention.
- No discernible evidence that noncompliance is problematic.
  - RCT results seem to generalize within study sample.
- RCT results do not seem to generalize to the remaining freshman class.
  - The experimental sample is **negatively** selected on the basis of **observed** characteristics.
  - The experimental sample is **positively** selected on the basis of **unobserved** propensity to persist in college.
- The null RCT results are contradicted by large significantly positive observational results.
  - This divergence in results is due to the selection into the experiment on unobserved variables.

▸ Setting

Defining the population of interest

Well...it depends on the question and audience.

Take the example of Moving to opportunity Jacobs (2004) and Chyn (2018) vs. Goering et al. (1999), Orr et al. (2003), Sanbonmatsu et al. (2011) and Chetty et al. (2016).

- Interpretation 1: What is the effect of vouchers on well-being?
- The self-selected may be the population of interest to policy makers, so long as the selection process and scale remain constant.

- Interpretation 2: Does neighborhood quality matter?
- This question is perhaps more relevant to social scientists and pertains to a broader population than those who self-select into a randomized study.

Defining the population of interest

Well...it depends on the question and audience.

Take the example of Moving to opportunity Jacobs (2004) and Chyn (2018) vs. Goering et al. (1999), Orr et al. (2003), Sanbonmatsu et al. (2011) and Chetty et al. (2016).

- Interpretation 1: What is the effect of vouchers on well-being?
- The self-selected may be the population of interest to policy makers, so long as the selection process and scale remain constant.
- Interpretation 2: Does neighborhood quality matter?
- This question is perhaps more relevant to social scientists and pertains to a broader population than those who self-select into a randomized study.

Defining the population of interest

Well...it depends on the question and audience.

Take the example of Moving to opportunity Jacobs (2004) and Chyn (2018) vs. Goering et al. (1999), Orr et al. (2003), Sanbonmatsu et al. (2011) and Chetty et al. (2016).

- Interpretation 1: What is the effect of vouchers on well-being?
- The self-selected may be the population of interest to policy makers, so long as the selection process and scale remain constant.

- Interpretation 2: Does neighborhood quality matter?
- This question is perhaps more relevant to social scientists and pertains to a broader population than those who self-select into a randomized study.

Defining the population of interest

Well...it depends on the question and audience.

Take the example of Moving to opportunity Jacobs (2004) and Chyn (2018) vs. Goering et al. (1999), Orr et al. (2003), Sanbonmatsu et al. (2011) and Chetty et al. (2016).

- Interpretation 1: What is the effect of vouchers on well-being?
- The self-selected may be the population of interest to policy makers, so long as the selection process and scale remain constant.

- Interpretation 2: Does neighborhood quality matter?
- This question is perhaps more relevant to social scientists and pertains to a broader population than those who self-select into a randomized study.

Defining the population of interest

Well...it depends on the question and audience.

Take the example of Moving to opportunity Jacobs (2004) and Chyn (2018) vs. Goering et al. (1999), Orr et al. (2003), Sanbonmatsu et al. (2011) and Chetty et al. (2016).

- Interpretation 1: What is the effect of vouchers on well-being?
- The self-selected may be the population of interest to policy makers, so long as the selection process and scale remain constant.

- Interpretation 2: Does neighborhood quality matter?
- This question is perhaps more relevant to social scientists and pertains to a broader population than those who self-select into a randomized study.

Population of interest for our setting

- Want to know the average effectiveness for those who enrolled in the FYLC.
- Some who received treatment did not enroll in the lottery.
- Further, in subsequent years the program was scaled up to 90 percent of the incoming freshman class.
- Accordingly, we take the entire freshman class as the population of interest.

▸ Populations

Recruitment, Randomization, and Compliance

- Over two years during consecutive summer enrollment sessions, freshman were encouraged to apply for the program.
- Of the 8,131 incoming freshman 1,565 applied to be part of the FYLC.
- 824 were selected for treatment.
- 170 of the lottery winners selected out of treatment.
- 108 of the 741 lottery losers took their spots.
- 117 of the 6,566 students who did not enter the lottery migrated into treatment.
- In the last few years the program has grown to serve 90 percent of the freshman class.

▸ Populations

Intuitive parametric approach examining selection

$$y_i = \mathbf{X_i}\gamma + D_i\beta + D_i e_i + \epsilon_i,$$

Let the following be the unobserved propensities to persist in college across lottery participation and treatment status:

- $E(\epsilon_i \mid \mathbf{x}, D = 0, L = 0) = \overline{\epsilon_{00}}$
- $E(\epsilon_i \mid \mathbf{x}, D = 0, L = 1) = \overline{\epsilon_{01}}$
- $E(\epsilon_i \mid \mathbf{x}, D = 1, L = 0) = \overline{\epsilon_{10}}$
- $E(\epsilon_i \mid \mathbf{x}, D = 1, L = 1) = \overline{\epsilon_{11}}$

Also, let the following be the unobserved responsiveness to treatment:

- $E(e_i \mid \mathbf{x}, D = 1, L = 0) = \overline{e_{10}}$
- $E(e_i \mid \mathbf{x}, D = 1, L = 1) = \overline{e_{11}}$

Differences by lottery participation within treatment status

The difference in outcomes conditional on $\mathbf{X} = \mathbf{x}$ within treatment status is given by the following:

Untreated
$E(Y_i \mid \mathbf{x}, D = 0, L = 1) - E(Y_i \mid \mathbf{x}, D = 0, L = 0) = \overline{\epsilon_{01}} - \overline{\epsilon_{00}}$

Treated
$E(Y_i \mid \mathbf{x}, D = 1, L = 1) - E(Y_i \mid \mathbf{x}, D = 1, L = 0) = \overline{e_{11}} + \overline{\epsilon_{11}} - \overline{e_{10}} - \overline{\epsilon_{10}}$

The later builds differences in the responsiveness to treatment into the tests.

Intuitive parametric tests for selection

Among the Untreated

$$E(y_i|\mathbf{x}, D_i = 0) = \mathbf{X_i}\boldsymbol{\gamma_0} + L_i\pi_0$$

A substantially or significantly non-zero $\hat{\pi_0}$ 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.

Among the Treated

$$E(y_i|\mathbf{x}, D_i = 1) = \mathbf{X_i}\boldsymbol{\gamma_1} + L_i\pi_1$$

A substantially or significantly non-zero $\hat{\pi_1}$ provides evidence of selection on unobserved characteristics and/or responsiveness to treatment, making the claim of external validity unreasonable.

Populations within the data

Name	Conditional outcome	Composition
Complacent Treatment	$E(Y_i \mid \mathbf{x}, D=1, L=1, Z=1) = \mathbf{x}\gamma + \beta + \overline{e_t} + \overline{\epsilon_t}$	Compliers and always-takers
Complacent Control	$E(Y_i \mid \mathbf{x}, D=0, L=1, Z=0) = \mathbf{x}\gamma + \overline{\epsilon_c}$	Compliers and never-takers
No-shows	$E(Y_i \mid \mathbf{x}, D=0, L=1, Z=1) = \mathbf{x}\gamma + \overline{\epsilon_n s}$	Never-takers
Crossovers	$E(Y_i \mid \mathbf{x}, D=1, L=1, Z=0) = \mathbf{x}\gamma + \beta + \overline{e_c o} + \overline{\epsilon_c o}$	Always-takers
Never-ever-takers	$E(Y_i \mid \mathbf{x}, D=0, L=0, Z=0) = \mathbf{x}\gamma + \overline{\epsilon_n}$	Never-ever-takers
Late-takers	$E(Y_i \mid \mathbf{x}, D=1, L=0, Z=0) = \mathbf{x}\gamma + \beta + \overline{e_l} + \overline{\epsilon_l}$	Late-takers

▸ RCT results

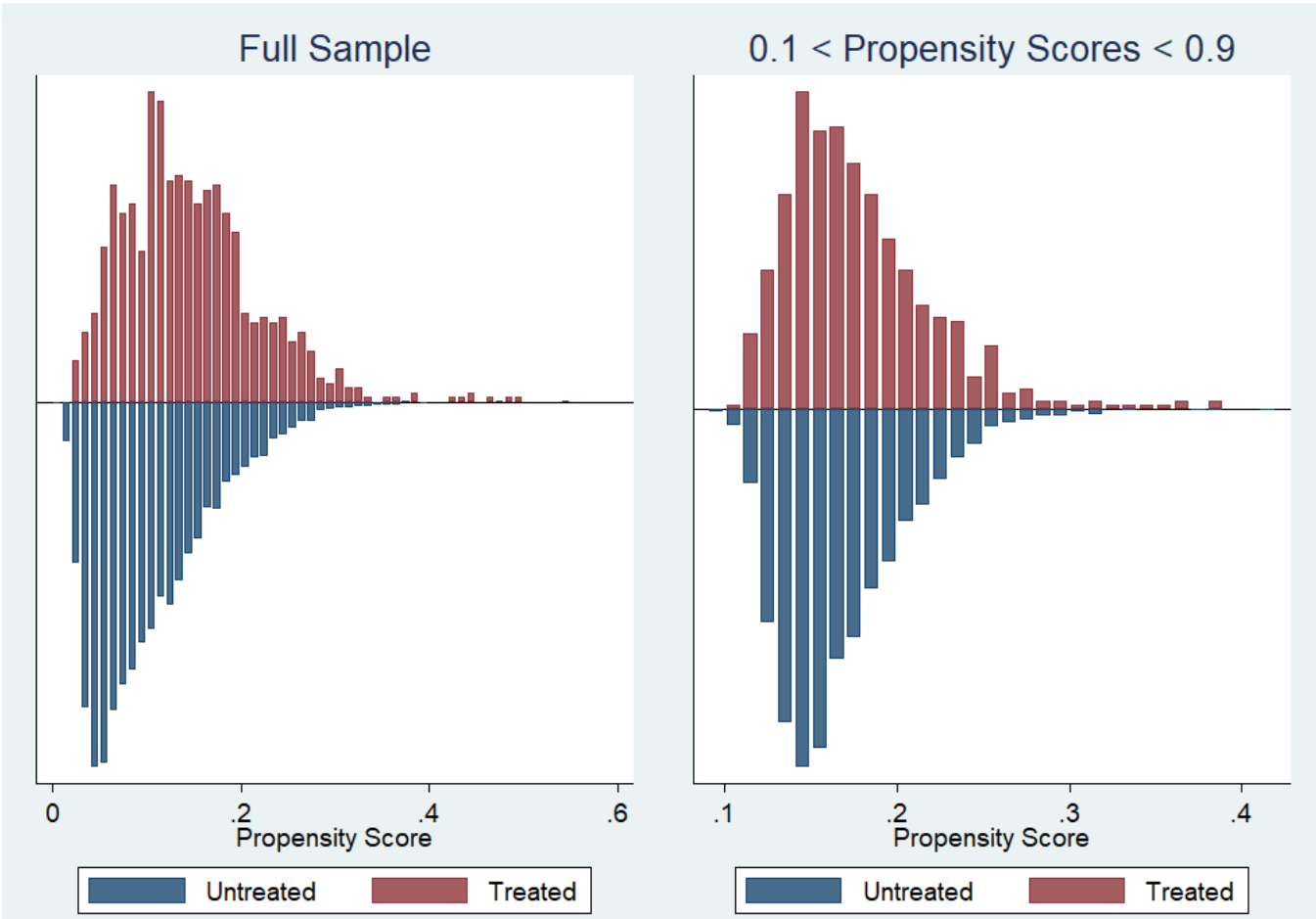
Within-study design

Observational analysis:

- We use OLS, Logit, Propensity Score Matching.
- We both examine unconditional differences and adjust for observed covariates.
- Conduct analysis on the whole sample and the sample for which propensity scores range 0.1-0.9.
- Following Huber (2013) we include interactions to decompose the OLS results to show what is driving them.

▸ RCT results

Overlap



Note: Propensity scores estimated using logit.

▸ RCT results

1st stage estimates of the effect of lottery on participation

Table: First Stage

	(1)	(2)
Won lottery	0.648***	0.648***
	(0.019)	(0.019)
Observations	1565	1565
Controls	No	Yes
Model	LPM	LPM

▸ RCT results

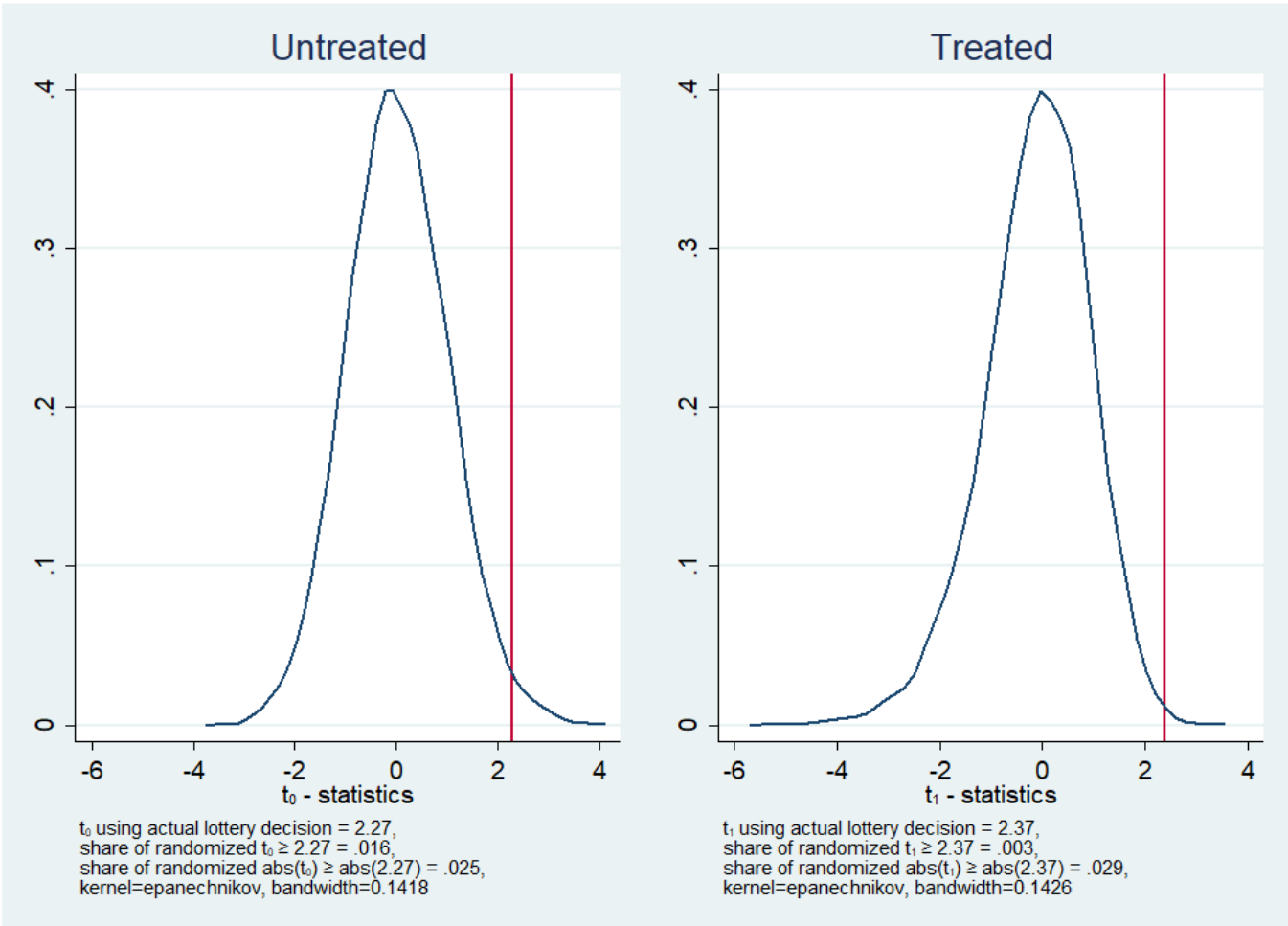
RCT estimates of the effects on GPA

Table: RCT estimates of the effects on 1st and 2nd year GPA

	(1) 1st Year GPA	(2) 1st Year GPA	(3) 2nd Year GPA	(4) 2nd Year GPA
Panel A: ITT				
Won lottery	0.016	0.018	0.015	-0.016
	(0.03)	(0.027)	(0.038)	(0.036)
Panel B: LATE				
FYLC	0.024	-0.027	0.022	-0.018
	(0.045)	(0.044)	(0.053)	(0.053)
Observations	1489	1489	662	662
Controls	No	Yes	No	Yes

▸ RCT results

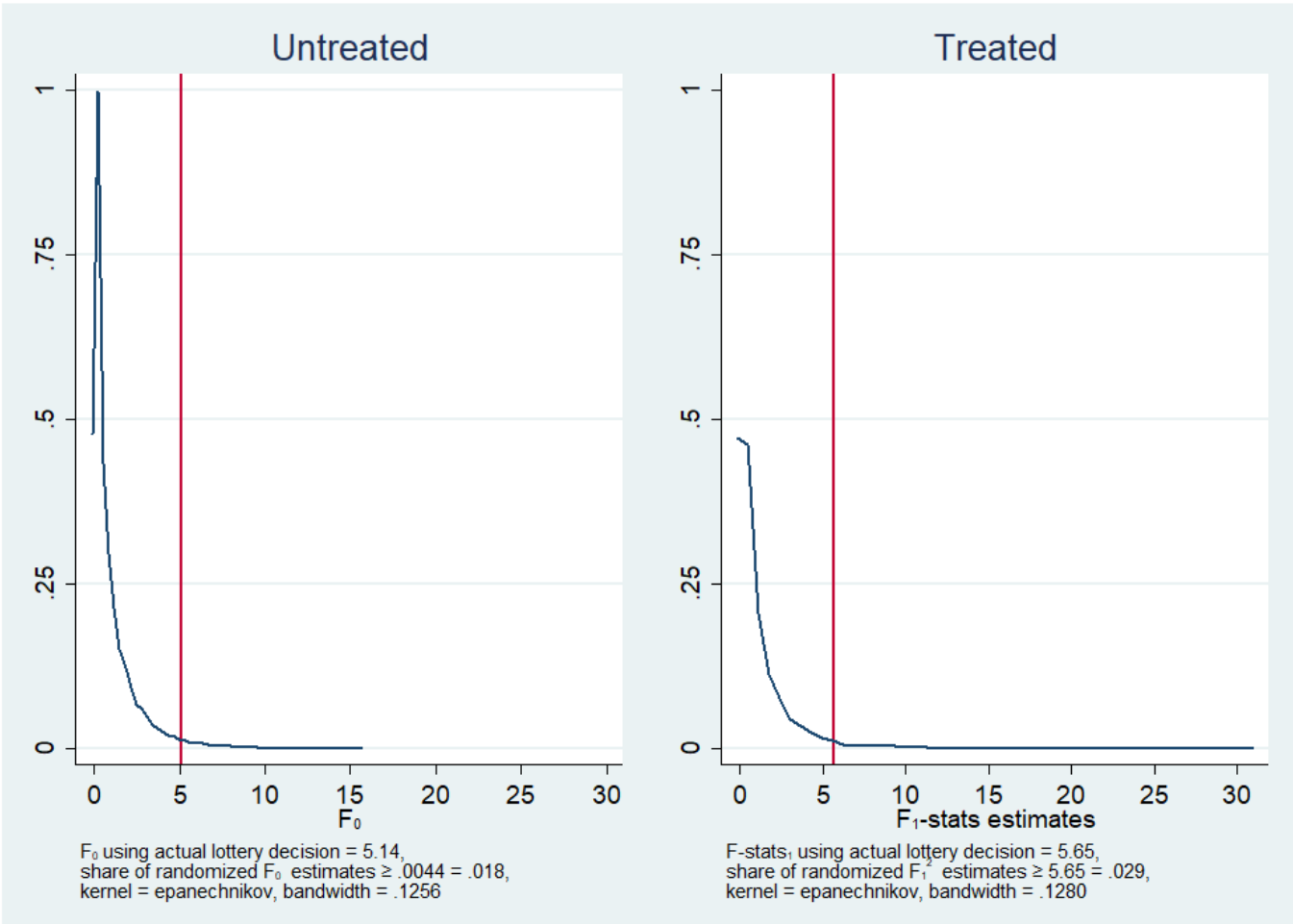
Selection into experimental: randomization inference



Notes: Distributions constructed from 10,000 repetitions.

▸ Selection

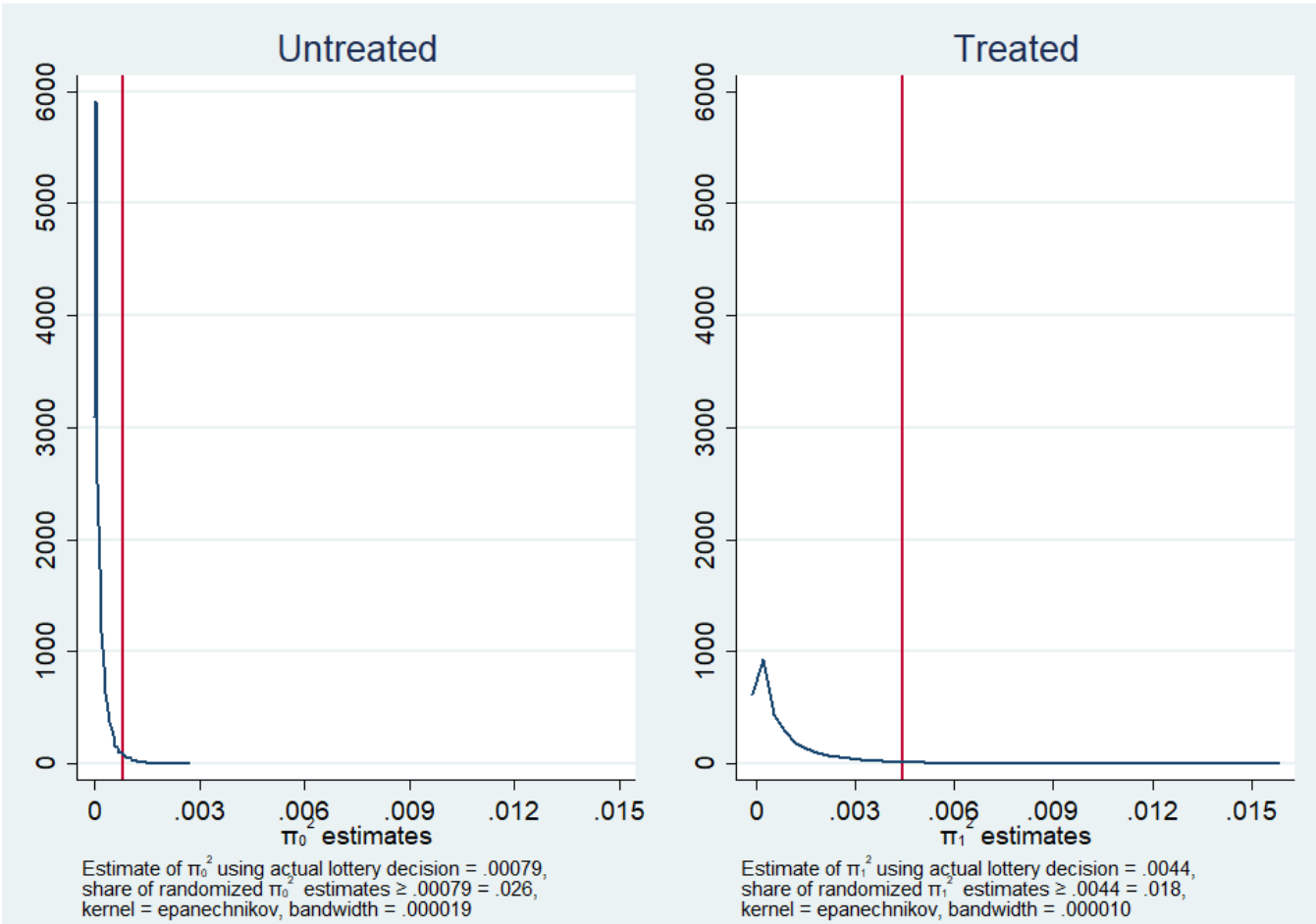
Selection into experimental: randomization inference



Notes: Distributions constructed from 10,000 repetitions.

Selection

Selection into experimental: randomization inference



Notes: Distributions constructed from 10,000 repetitions.

▸ Selection