The Consequences of Sorting for Understanding School Quality

Jesse Bruhn*

Job Market Paper

January 1, 2019

Click for most recent version

Abstract

I study the sorting of students to school districts using new lottery data from an inter-district school choice program in Massachusetts. I find that moving to a more preferred school district increases student math scores by 0.16 standard deviations. The program also generates positive effects on coursework quality, high-school graduation, and college attendance. Motivated by these findings, I develop a rich model of treatment effect heterogeneity and estimate it using a new empirical-Bayes-type procedure that leverages non-experimental data to increase precision in quasi-experimental designs. The estimator I propose is a weighted average of experimental and non-experimental variation, with the weights chosen according to the correlation of the heterogeneous effects across samples. I use the heterogeneous effects to examine Roy selection into the choice program. Students who would be negatively impacted by the program are both less likely to apply and, conditional on taking up an offer to enroll, are more likely to subsequently return to their home district. I find that this selection drives almost all of the program evaluation treatment effect identified with the lottery. The fact that families sort students to school districts according to potential benefit suggests that research relying on school choice lotteries to learn about differences in school quality may lack any broad claim to external validity.

Keywords: School Choice, Roy Selection, External Validity, Bayesian Modeling

JEL classification: C11, C36, H75, I21, I28, J24

* Department of Economics, Boston University. Contact: jessebr@bu.edu. Website: jessebruhn.com

I am indebted to Kevin Lang for his invaluable guidance and encouragement throughout this project. I would also like to thank Daniele Paserman, Ray Fisman, Marcus Winters, Scott Imberman, Ivan Fernandez-Val, Carrie Conaway, Pascual Restrepo, James Feigenbaum, Joshua Goodman, Kehinde Ajayi, Sam Bazzi, Phillip Ross, Bruno Martins, and Arthur Smith for helpful conversations and comments, as well as the participants at the Boston University Empirical Micro Workshop and the Boston University Labor Reading group for their excellent feedback. Finally, I would like to extend my gratitude to the Massachusetts Department of Elementary and Secondary Education for making this project possible.

Introduction

There is now a well documented causal link between educational inputs, test scores, and later life outcomes. Whether it is the size of a kindergarten classroom, the value added of a middle school teacher, or the type of high school a student attends, educational interventions have far reaching consequences for outcomes like teen pregnancy, incarceration, college attendance, and adult earnings (Cullen, Jacob & Levitt 2006, Chetty et al. 2011, Angrist et al. 2012, Chetty, Friedman & Rockoff 2014, Deming et al. 2014, Dobbie & Fryer 2015, Angrist et al. 2016). Thus understanding school quality is important for effectively targeting educational investments.

Recent work on school effectiveness leverages randomization in the school assignment process to generate estimates of quality differences across institutions. Since the results of the lottery are random, estimates of school quality based on comparisons between school choice lottery winners and school choice lottery losers are not confounded by higher ability or better-resourced students choosing to attend better schools. For this reason, researchers have used lottery estimates of school quality to construct novel measures of value added, to validate observational methods of ranking schools, and to estimate the relation between school effectiveness and educational inputs (Angrist, Pathak & Walters 2013, Dobbie & Fryer 2013, Deming et al. 2014, Abdulkadiroglu et al. 2017, Angrist et al. 2017).

While school choice lotteries may seem like an attractive tool for learning about effectiveness, individual students may nonetheless experiences test score gains by switching schools even in the absence of differences in average school quality. If students use choice programs to sort to schools on the basis of idiosyncratic benefit, then the gains identified by a comparison of lottery winners to losers have no straightforward connection to quality. More generally, school choice lottery estimates will not be externally valid in the presence of Roy selection (Walters 2017). Thus knowing whether and to what degree lottery identified test score gains are driven by sorting versus differences in quality is necessary for understanding the practical policy relevance of this body of work.

In this paper, I use random admission offers from an inter-district school choice program in Massachusetts to study the consequences of sorting for understanding school quality. I provide three main contributions. The first contribution is a causal evaluation of the impact of inter-district school choice on student outcomes using new, hand-collected lottery data. The second contribution is an examination of the role that Roy selection plays in generating the causal benefits of inter-district choice. The third contribution is a new method for estimating treatment effect heterogeneity that uses non-experimental data to increase precision in quasi-experimental designs.

I start with a program evaluation of inter-district school choice. I find that moving to a more preferred district increases student math scores by 0.16 standard deviations, with no effect on English Language Arts. The impact on math is large. The unadjusted 10th grade black-white math score gap in Massachusetts is 0.56 standard deviations. This result also stands in contrast to prior estimates of the effects of school choice in the traditional public school sector, which find little to no impact on test scores (for examples see Cullen, Jacob & Levitt 2006, Hastings, Neilson & Zimmerman 2012, Deming et al. 2014). I find that students who participate in the inter-district choice program are more likely to take advanced placement and other advanced classes. I also find positive effects on the probability that students who participate in the program graduate from high school and go on to attend a four-year college.

The findings from this evaluation are interesting because they represent the first lottery evaluation of a state-wide inter-district choice program. Such programs are common in the United States (Wixom 2016) and also controversial. Critics argue that because funding typically follows the student, inter-district choice drains educational resources from underprivileged communities (O'Connell 2017). Thus understanding the causal impact of inter-district choice is important for policy. Prior work has been limited to examining the consequences of within-district urban assignment mechanisms, choice to charter schools, the impact of private school vouchers, and race-based desegregation programs.[1]

Next I analyze the role that Roy selection plays in generating test score gains. I accomplish this by estimating a model of treatment effect heterogeneity that incorporates a rich set of student observables: lagged test scores, subsidized lunch recipiency, race/ethnicity, gender, and measures of student behavior. I find that the observed heterogeneity predicts student take-up behavior in a way that is consistent with Roy selection. Students who would be negatively impacted by the program are much less likely to apply; conditional on applying, negatively impacted students are less likely to take up a randomly assigned offer to enroll; and once enrolled, negatively impacted students are less likely to continue on in the program after their first year.

This finding is significant because selection on potential benefit drives a wedge between the local average effect identified by the lottery and the average treatment effect of interest:

[1]For recent examples of choice among traditional public schools, see Cullen, Jacob & Levitt (2006), Hastings, Neilson & Zimmerman (2012), Deming et al. (2014), and Abdulkadiroglu et al. (2017). For recent examples of choice to the charter sector, see Hoxby & Murarka (2009), Abdulkadiroglu et al. (2011), Dobbie & Fryer (2011), Angrist et al. (2012), and Angrist et al. (2016). For examples of the impact of private school vouchers, see Howell et al. (2002), Wolf et al. (2008), Mills & Wolf (2017), and Abdulkadiroglu, Pathak & Walters (2018). For race based programs, see Angrist & Lang (2004) or Bergman (2018).

school quality. To get a sense of the bias this selection induces, I use the observed heterogeneity to extrapolate the average treatment effect for the treated, the applicants and the non-applicants. I find that 38% of the treatment effect for the treated comes from post-lottery selection into enrollment and that 78% of the treatment effect for applicants is driven by pre-lottery selection into the applicant pool. Almost none of the local average effect identified with the lottery is the result of quality differences across districts. The fact that families sort students to school districts according to potential benefit suggests that research relying on school choice lotteries to learn about differences in school quality may lack any broad claim to external validity. These findings also add to a recent literature examining the relationship between selection and heterogeneity for understanding optimal policy (Walters 2017, Mogstad, Santos & Torgovitsky 2018, Hull 2018).

The final contribution of this paper is a new estimator that leverages non-experimental data to efficiently estimate heterogeneous treatment effect models in quasi-experimental designs. In order to study the sorting of students to districts on the basis of potential benefit, I must first fit a rich heterogeneous effects model using an instrumental variables (IV) strategy. Unfortunately, IV designs are notoriously noisy (Young 2017). This makes precise estimation of the heterogenous effects difficult with the lottery sample at my disposal. However, I show that corresponding estimates using observational data on the universe of public school students in Massachusetts are highly correlated with the estimates from the experimental sample. This suggests that the non-experimental data contains information that is useful for pinning down the local average effects identified by the IV design. I formalize this intuition by combining the experimental and non-experimental estimates within a hierarchical model. The estimator is consistent under the same conditions as IV and, under the assumption that the heterogeneous effects are normally distributed, it is more efficient.

The estimator I propose adds to an emerging literature in Economics that uses random-effects and other bayesian or quasi-bayesian methods to synthesize information from multiple sources (e.g. Hull 2018, Meager 2017, Meager 2018). In particular, the method outlined in Angrist et al. (2017) is closely related. The authors of that paper use a simulated method of moments approach that combines non-experimental and lottery identified value added in a hierarchical model to generate a complete quality ranking across oversubscribed and undersubscribed schools in Boston. The method I develop is similar in spirit to the just-identified version of their model; however, because I am only interested in efficiency gains, whereas Angrist et al. (2017) use the non-experimental data to solve an otherwise under-identified model, I do not need to model the first stage, reduced form, and least square's bias jointly within the parent distribution. This allows me to find a closed form solution with

simple, transparent intuition. And unlike Angrist et al. (2017), this approach allows for the possibility that the local average treatment effect identified by the lottery is different from the average treatment effect in the population.

This econometric method also offers a partial answer to a recent critique of instrumental variable designs. Young (2017) argues that due to their lack of precision, the typical IV design in economics generates no increase in knowledge beyond what is learned from the corresponding least squares regression. However, the decision about whether to use IV or least squares need not be binary. Provided the econometrician cares about a collection of parameters beyond the average treatment effect, the estimator I propose offers a principled way to average the IV and least squares estimates and thus fully leverage the available information.

1 Increasing Access with District Choice

The purpose of inter-district choice in Massachusetts is to weaken the link between geography and access to a high quality education. The program was originally established in 1993 as one portion of a broader set of education reforms known as the Massachusetts Educational Reform Act (MERA). Broadly speaking, the reforms centered around three areas: school funding, accountability, and access. To further the latter objective, MERA established provisions allowing for both charter schools and inter-district choice (Chester 2014). Between 2001 and 2016, over 70,000 students enrolled in a school outside their home district via the inter-district choice program. To put this number in context, over the same time span the charter sector in Massachusetts enrolled around 119,000 students.[2] Figure 1 shows enrollment in the inter-district choice and the charter sector over time.

At the district level, the program operates in several stages that may or may not culminate in a lottery for admission. By default, every public school district in Massachusetts participates in the program. However, each year the local school board may vote to opt out. If the school board votes to opt out, the district is not required to enroll students from other districts; however, voting to opt out does not preclude local students from using the program. The law then requires that participating districts project capacity and enrollment and make excess seats available to any student in the state. The projection methods are determined locally. Since 2001, nearly 200 districts out of approximately 295 traditional public school districts[3] in Massachusetts have enrolled at least one student via the program, with 156 districts

[2]Both calculations are my own and were made using administrative student micro-data provided by the Massachusetts Department of Elementary and Secondary Education.

[3]Over this period, some districts consolidated into regional districts.

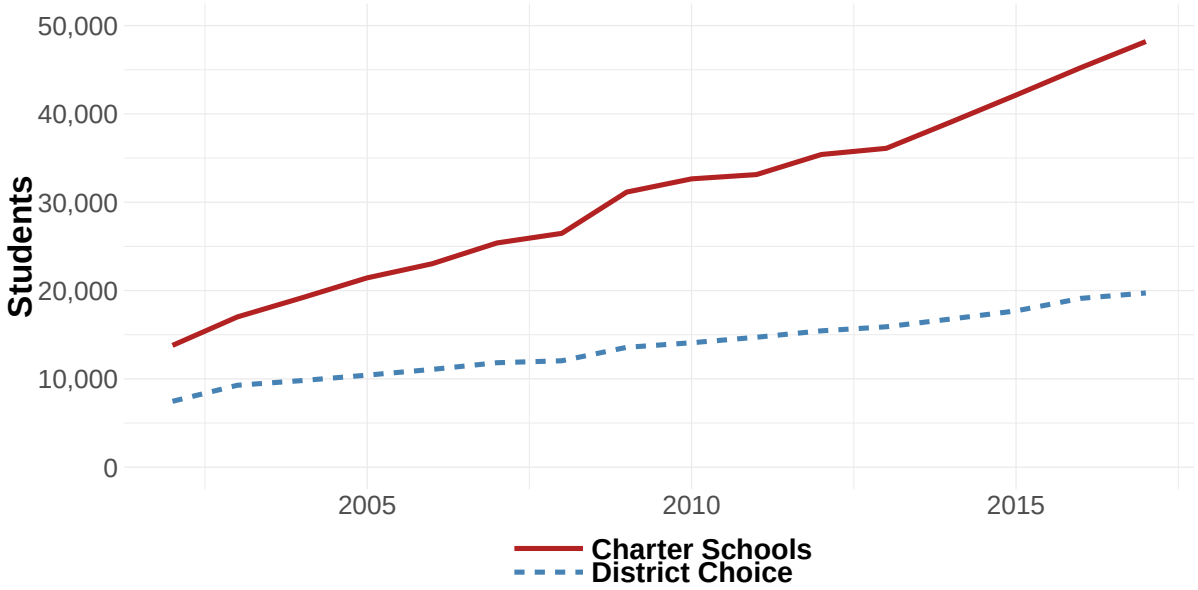

Figure 1: Enrollment in Inter-district Choice and Charter Schools Over Time

participating in an average year. Figure 2 shows the geospatial distribution of choice districts as of 2016. When the number of students who apply exceeds the number of seats available, the district is required to allocate the seats via lottery. Once a student is offered a spot in the district and accepts, she becomes a full public school student of the district until she graduates or leaves voluntarily. However, transportation is the responsibility of the family.[4] The sending district is then required to pay the receiving district the lesser of 75% of average per-pupil expenditures in the sending district or $5,000. However, the sending district must pay the full cost of any special education services as determined by the state funding formula. In practice, the $5,000 cap is binding for non-special education students.

The way the program is implemented in practice sometimes differs substantially from the text of the law. For example, an advisory memo from the Massachusetts Office of General Counsel concluded that the non-discrimination language in the law was so strong that even sibling preference should not be considered when administering lotteries for admissions purposes (Moody 1994). In practice, nearly every district offers some form of sibling preference.[5] In addition, there are a number of districts which are regularly oversubscribed yet conduct admissions on a first-come first-serve basis.[6] Finally, there are some portions of the law which

[4]There are some exceptions to this rule for students with disabilities.

[5]This assertion is based on conversations I had with state level program officials and district level administrators while collecting data.

[6]While collecting data, at least five districts indicated this to me, but not all districts offered this information when responding to my emails.

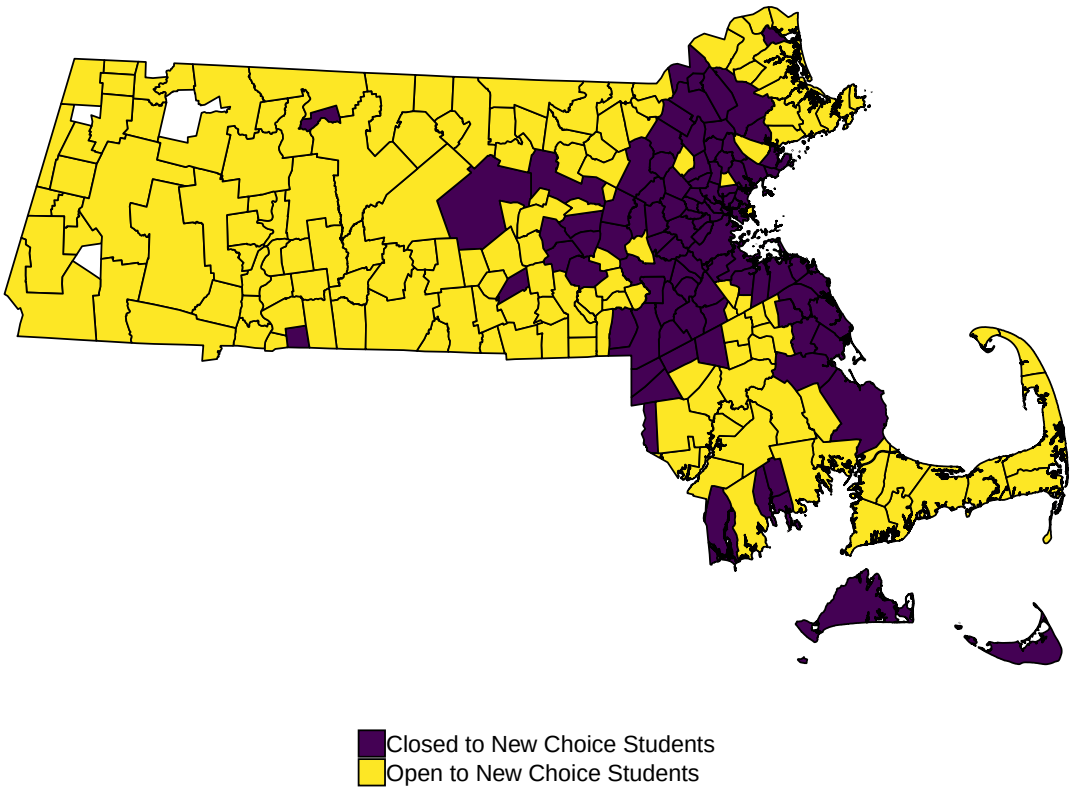

Figure 2: Inter-district Choice in 2016

simply never made it in to practice. For example, the original bill asked participating districts to submit their enrollment and capacity projections to the Massachusetts Department of Elementary and Secondary Education (DESE). I learned from my conversations with state level program administrators that, in practice, this information has never been collected.

2 Collecting District Choice Data in Massachusetts

The data I use for this project come from several sources. I start with hand collected lottery records from school districts in Massachusetts. I then match and merge these lottery records to administrative data on the universe of public school students in Massachusetts. These administrative data include information on standardized test scores, teachers and coursework, as well as data on college outcomes via an extract from the National Student Clearinghouse. I also make use of of several spreadsheets provided to me by DESE which describe information such as which districts were open to choice in a given year, how the structure and coverage of districts has changed over time, and the within district distribution of education spending.

Finally, I augment these sources with publicly available data on property values from the Massachusetts Department of Revenue and district level data on the parental characteristics of public school students from the Census' Education and Geographic Estimates project. I will now briefly discuss each of these data sources in turn. For a complete characterization of the data matching and cleaning process, see the online appendix.

2.1 New Lottery Data

In May of 2016, I contacted every public school district in the state of Massachusetts that had ever enrolled a student via inter-district choice and asked them to share their lottery records with me.[7] Of the districts I contacted, approximately 75% responded. Of the districts that responded, 36% confirmed that they had ever conducted a lottery. Typically, districts that did not conduct a lottery were not over-subscribed. A small number of districts accepted new students using a first-come first-serve procedure despite being over-subscribed. Of the districts that had ever conducted a lottery, 38% had maintained records that they were willing to share with me. By far the most common reason for not sharing data was poor record keeping. Some districts elected not to participate out of privacy concerns. Of the records I collected, a substantial portion were unusable due to insufficient documentation of the lottery process. Ultimately, I was left with approximately 3,000 student level lottery records from 203 lotteries across 14 districts.

Districts used a variety of randomization mechanisms to conduct the lotteries. The most common randomization method involved having a secretary or administrator randomly select some subset of the applicants to receive offers of admission. I code these random offers as a binary "initial offer" instrument. This randomization procedure was used in 91% of the lotteries in my sample. Typically, the remaining applicants were then randomly assigned a waitlist number. When available, I also code these numbers as a "waitlist number" instrument. There was one district which, for a single year in my data, randomly chose students from a waitlist pool instead of assigning them lottery numbers. I code these random offers as a binary "waitlist offer" instrument and include it for completeness. There was also one small district whose records consisted of randomly assigned lottery numbers, with no indication as to who actually received an offer of admission. For this district, I code the raw number as a "lottery number" instrument. In practice, all the lottery results I present in this paper are driven by initial offers.

[7] A number of these districts were vocational districts, internet based learning programs, or other non-traditional programs that I subsequently learned were not required to use a lottery based admissions process. For this reason, I don't count these districts when calculating response rates.

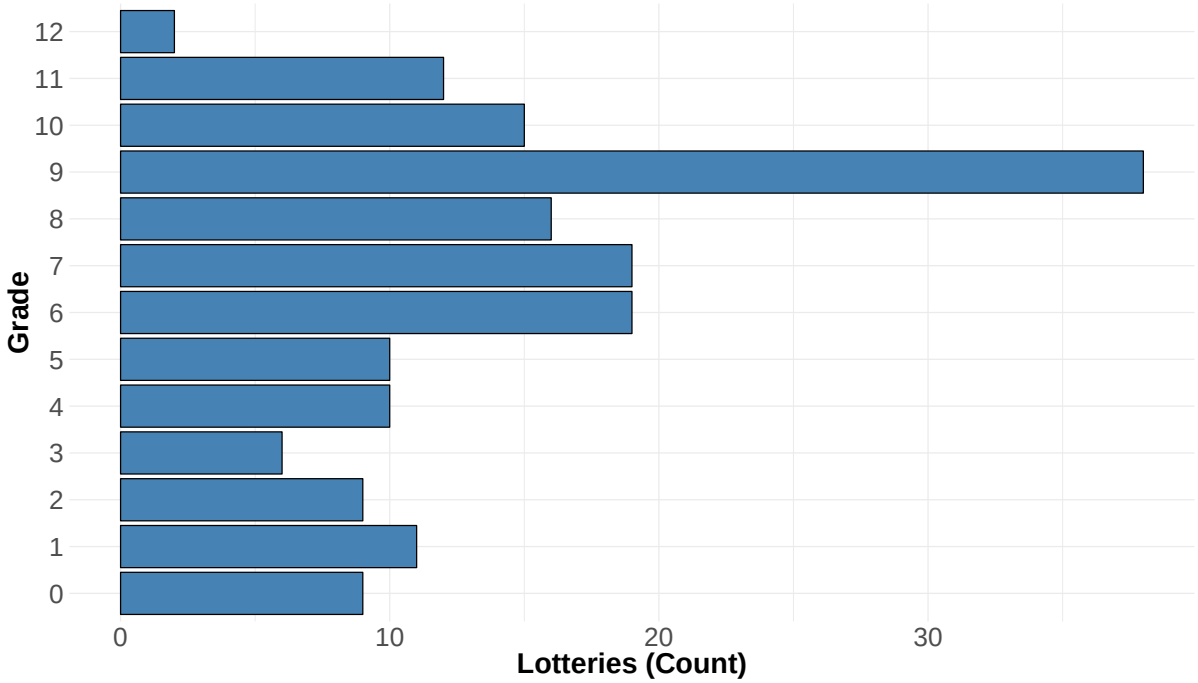

Figure 3: Distribution of Lotteries by Grade

The typical lottery in my sample is small. The average number of students I view in a single lottery is 9.6; the median is 7. The lotteries also span a considerable time-period. The earliest lottery in my data occurs in academic year 2002-2003; the latest occurs in academic year 2016-2017. None of the 2016-2017 lotteries are included in my estimation sample since, as of the time the analysis was conducted, the necessary outcome variables were unavailable post-lottery. Finally, I will note that the lotteries in my sample span all grade levels. However, as can be seen in figure 3, the lotteries are clustered at grades which are typically within-district, cross-school transition points for students.[8] For more detailed descriptive statistics regarding the raw lottery data, see the online appendix.

I merge these student lottery records to the data provided by DESE by looking for exact first and last name matches within the implied application grade / year. When available, I break ties using middle names / initials, home-town and date of birth. When town of residence is unavailable and I am otherwise unable to break a tie, I choose individuals that live within the empirical distribution of towns that lose students to the receiving district via choice. If I am unable to break a tie in this way, I consider the student un-matched and drop her from the sample. When this procedure fails to find any exact match, I repeat it using fuzzy first and last

[8]For example, students often move from grammar to middle school in the fifth, sixth, or seventh grade, and from middle to high-school in ninth grade.

name matching. For this reason, all of my specifications will include indicators for whether a student was matched via the exact or fuzzy version of the algorithm. Overall, I obtain an 89% match rate.

I will note here that my lottery sample exhibits some imbalance along predetermined characteristics. Figure 4 presents point estimates and two standard deviation intervals from a within-lottery regression[9] of all baseline observable and otherwise exogenous characteristics on the initial offer indicator for the sub-sample of students where I observe at least one test score prior to the lottery. The joint F-statistic across all pre-determined characteristics is 1.64. Of particular concern is the fact that the coefficient for black students is negative and the 2 standard deviation interval does not include zero. However, the administrators conducting the lottery could not directly observe race,[10] the magnitude of the coefficient is small, white students also have a negative point estimate, and the point estimate for black students is not significantly different than the point estimate for white students (or any other racial group). For these reasons, it seems unlikely that racial discrimination is the culprit. In the online appendix, I consider the possibility that this imbalance is driven by differential attrition and conclude that this is also unlikely to be the case.

While it is possible that the covariate imbalance is due to some form of cheating on the part of districts, I believe this is unlikely for two reasons. First, all of the districts that provided lottery data to me did so voluntarily and described to me in detail the process they used for randomization. Second, cheating would open the district up to potentially serious liability. As I discussed in section 1, the legal office in the department of education in Massachusetts concluded that the anti-discrimination language in the inter-district choice law was even stronger than that used in the charter sector. Further, there was no consequence for opting not to share data with me. Thus if a district was cheating, they had strong incentive to not provide me with data. One explanation for the imbalance is the possibility that some of the lottery records I obtained did not track things like sibling preference or late applications properly. Another potential explanation is that this imbalance is simply the product of sampling variation. In any event, I show in the online appendix that conditioning on earlier pre-lottery test scores increases my precision substantially and, more importantly, that such specifications pass all of the standard falsification tests used in lottery designs. For this reason, every specification in this paper using the lottery variation is restricted to the sample of

[9]Within lottery is the level of variation at which the instrument is randomly assigned. In practice, I do this by including lottery fixed effects. I also drop all students from this regression that received sibling preference or were indicated as applying late.

[10]Of course, it is possible that lottery administrators were able to infer race from student or parent names or that they were able to observe race if a student or her family dropped the application form off in person.

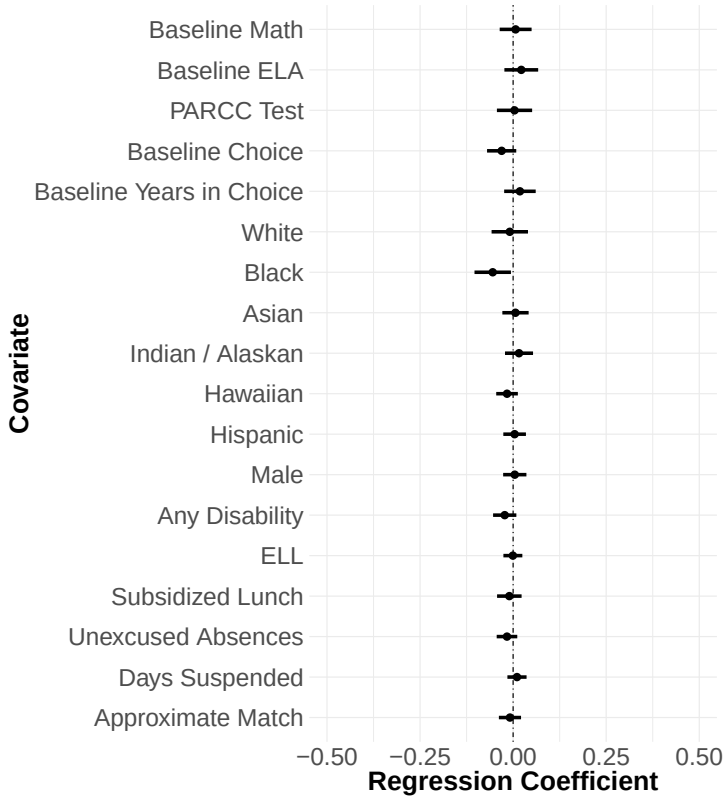

Figure 4: Covariate Balance by Initial Offer Status

students for whom I observe at least one test score prior to the lottery year and will include baseline test scores as controls.

2.2 Administrative Student Records and Other Data Sources

For this project, the state of Massachusetts provided me with data on the universe of public school students. I retrieved demographic and socioeconomic information from the Student Information Management System (SIMS) spanning academic years 2001-2002 through 2016-2017. This included variables related to race / ethnicity, gender, attendance, discipline, disability, and whether the student received a subsidized lunch, as well as the variables necessary for matching. It also includes administrative information on the district, school, and grade-level where students are assigned in a given year, including an indicator for whether a student was enrolled in a district via inter-district choice. I drop observations appearing in adult education programs, collaborative or special education schools, online schools, charter schools, and vocational schools.

I retrieve test scores from the Massachusetts Comprehensive Assessment System (MCAS)

spanning academic years 2001-2002 through 2016-2017. I standardize the test scores at the grade, year, and test-type[11] level to have mean zero and standard deviation one. I retrieve coursework taken by students from Student Course Schedule (SCS) data spanning academic years 2010-2011 through 2016-2017. I also use data on college attendance contained in an extract from the National Student Clearinghouse purchased by DESE.

For some auxiliary regressions, I make use of additional spreadsheets provided to me by the state level officials who administer the program. These spreadsheets describe district finances, as well as the outcome of the annual district level votes on choice status spanning academic years 2007-2008 to 2016-2017. I also make use of district level socio-economic and demographic data on parents of public school students from the Census' Education, Demographic and Geographic Estimates (EDGE) project, as well as data on property values which I downloaded from the Massachusetts Department of Revenue.

For further description of the various data sources along with a detailed break-down of the cleaning process, see the online appendix.

3 Program Take-up by Students and Districts

Students in my lottery sample are positively selected both relative to the state as a whole and relative to their home district peers. Table 1 illustrates this fact. The column labeled "All Districts," provides averages of observables across the entire state for students in test taking grades in academic years 2001-2002 through 2016-2017. The column labeled "Choice Students" restricts the state-wide sample to observations where a student is currently participating in inter-district choice. The column labeled "Sending Districts" restricts the state-wide sample to districts that lose a student to choice via a lottery I observe in my data. The column labeled "Lottery Sample" includes students found in my lottery data as observed at baseline.[12] The column labeled "Compliers" uses the method of Abadie & Kennedy (2003) to re-weight the lottery sample such that the averages reflect those of initial offer lottery compliers.

Compared to their home district peers, the lottery sample is disproportionately white, less likely to receive a subsidized lunch, less likely to be identified as limited English proficiency,

[11] The state transitioned testing regimes from the original MCAS exam to the Partnership for Assessment and Readiness for College Careers (PARCC) exam over the course of my sample frame. There are 3 years in my data where the old and new examinations appear simultaneously. For this reason, all regressions will also include test-type fixed effects.

[12] There are ≈ 80 students involved in lotteries that did not use an initial offer mechanism and that I do not include here. Including them does not meaningfully change the averages in this column, and excluding them facilitates calculating the complier averages in column four. To view the averages for the entire estimation sample, see the online appendix.

Table 1: Student Selection into Inter-District Choice

	All Students	Choice Students	Sending Districts	Lottery Sample	Compliers
Math	0.02σ	0σ	-0.21σ	0.11σ	0.05σ
ELA	0.02σ	0.04σ	-0.21σ	0.14σ	0.11σ
White	83%	93%	68%	90%	87%
Black	11%	6%	29%	10%	12%
Hispanic	15%	8%	27%	5%	3%
Male	51%	48%	51%	46%	47%
Subsidized Lunch	33%	28%	55%	21%	21%
Limited English	6%	1%	11%	0%	1%
Disability	12%	12%	13%	11%	12%
Days Attended	165.12	163.5	161.68	167.91	168.66
Observations	3,879,633	56,440	178,458	881	881

less likely to be diagnosed with a disability, and has higher average test scores. However, when compared to the state as a whole, the differences are smaller. One notable pattern is the enormous difference in subsidized lunch recipiency across sub-samples. This is likely due to the fact that transportation to the new district is the responsibility of the family. For this reason, we should expect families with the resources to transport their children long distances to be more likely to apply to the program and subsequently accept lottery offers.

At the district level, the decision not to opt out of inter-district choice is typically determined by a desire to supplement revenue. When a district observes that it has extra space in a classroom, in the sense that it is below the target student to teacher ratio in a given grade level, the district will use the program as a source of additional funds. However, in the greater Boston area, participation is quite low. This is likely due to the fact that many suburban districts in the Boston area participate in the METCO program. As discussed in Angrist & Lang (2004), METCO is the nation's oldest voluntary school desegregation program. It provides a separate mechanism for filling excess seats whereby predominantly white suburban districts enroll minority students from Boston. Thus METCO leads to a crowding out of inter-district choice.

These explanations are supported both by informal discussions I have had with district officials and by suggestive regressions in my data. Table 2 displays select coefficients from a joint regression of district characteristics on an indicator that takes a value of one in years when a district did not vote to opt out of inter-district choice. Column (1) displays select results from the joint regression estimated via OLS. Column (2) displays select results from the variables chosen when estimation is done using post-Lasso. Column (3) displays select results

Table 2: Select Predictors of District Participation

	Accepting New Choice Students		
	(1)	(2)	(3)
Student-Teacher Ratio	−0.07	−0.07	−0.002
	(0.02)	(0.02)	(0.01)
Per-Pupil-Spending: Pupil Services	0.15	0.23	0.03
	(0.08)	(0.07)	(0.05)
Metco Students (tens)	−0.01	−0.01	0.01
	(0.005)	(0.005)	(0.02)
Estimation Method	OLS	Post-Lasso	OLS
District/Year Fixed Effects	No	No	Yes
Dependent Variable Mean	0.55	0.55	0.55
Observations	2,280	2,280	2,280
Observations (Districts)	285	285	285
Adjusted R^2	0.34	0.31	0.88

from a joint regression that also includes district and year fixed effects; in effect, column (3) asks whether trends in the independent variables are predictive of changes in participation status. In levels, the student teacher ratio, various per-pupil expenditure categories, and the number of METCO students are predictive of the decision to participate. Other observables, such as the district demographic composition and urbanicity, are not. And almost none of the variables considered exhibit trends which predict changes in participation status. See the online appendix for complete results including the variables not displayed in table 2.

Finally, I note that as a result of this participation disparity, the net student gain / loss to choice is not evenly distributed across the state. Figure 5 shows the geographic distribution of the net gains and losses. The largest net winners and losers are concentrated in the middle and western regions. The winners tend to be suburbs and large regionalized school districts. The losers tend to be urban and rural.

4 Program Evaluation

In this section, I evaluate the benefits of inter-district choice for students that participate. For identification, I examine applicants to oversubscribed districts and compare the district

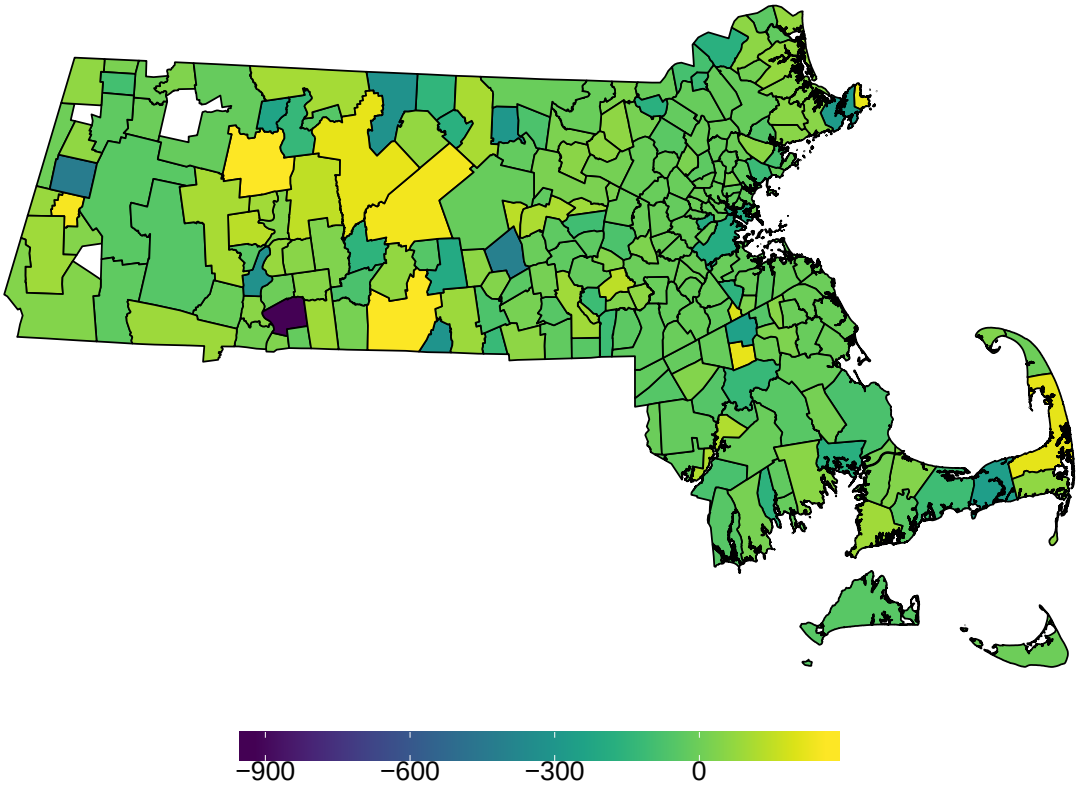

Figure 5: Net Student Gain/Loss to Inter-district Choice in 2016

choice lottery winners to the district choice lottery losers within a two-stage least squares framework. I find that participating in district choice causes large test score gains in math. I find no effect on English Language Arts scores. I also find that participating in district choice increases the quality of the coursework that students take. Finally, I provide evidence that participating in district choice increases the probability a student will graduate from high-school and attend a four-year college.

4.1 Identification and Estimation

Consider the following two-stage least squares framework:

$$y_{it} = \delta_0 + \beta d_i + \delta_\ell + \gamma W_i + \epsilon_{it} \tag{1}$$

$$d_{it} = \delta_0' + \Pi Z_i + \delta_\ell' + \gamma' W_i + \eta_{it} \tag{2}$$

Where y_{it} denotes the outcome of student i during a post-lottery period of time t (typically

an academic year), d_{it} is an indicator for whether student i was enrolled out of district via the choice program at time t, δ_ℓ is a lottery fixed effect,[13] δ_0 is a reference lottery, W_i are covariates observed at baseline,[14] and Z_i denotes the vector of four lottery instruments[15] discussed in section 2.

The parameter β identifies a local average treatment effect (LATE) specific to the instrument vector Z_i under a standard set of instrument-by-instrument conditions: exclusion, random assignment, first stage, and monotonicity (Imbens & Angrist 1994). Exclusion requires that the result of the lottery affect potential outcomes only via takeup of the treatment. Random assignment requires that within each lottery the results are, in fact, random. First stage requires that the results of the lottery change take-up behavior for some subset of the population (i.e. that $\Pi > 0$ for some element of Z_i). Monotonicity is a restriction on the heterogeneity of potential treatment status permitted in the first stage; it requires that all individuals whose behavior is changed by the results of the lottery behave consistently with respect to take-up. Provided these four conditions are satisfied, β is properly interpreted as the average treatment effect of moving to a more preferred school district for lottery compliers who applied to over-subscribed districts that maintained and were willing to share high quality lottery records. I save a discussion of heterogeneity and external validity for section 5.

I restrict the sample to the set of students appearing in my lottery data such that I observe at least one pre and one post lottery test score. I drop students who received sibling preference or applied late. When students apply to lotteries in multiple years, I randomly choose which observation to use. I also drop all students involved in a lottery if I am unable to match at least one student who receives a lottery offer and one student who does not; otherwise, the lottery would contribute no identifying variation to the estimate. Finally, I restrict the data to the set of student-year observations occurring after the lottery randomization.

For the standard errors, I follow the design based approach of Abadie et al. (2017) and cluster at the level at which treatment is assigned (i.e. the student). Other sensible approaches

[13] To be precise, a lottery is defined as the interaction of the grade, application district, and year where the student appears in my lottery data.

[14] All specifications will include an average of all test scores observed prior to the lottery year; academic year and grade fixed effects; indicators for PARCC testing; and indicators for whether or not a student was matched to the state data via an exact or fuzzy process. One district asked students who were not given a random initial offer whether or not they wanted to be included on the waitlist before assigning them a random waitlist number; I include an indicator where this happens in my data. However, the results are not sensitive to dropping these observations. I also had a district that, for one lottery, indicated "admission rounds" in their lottery spreadsheet without further explanation. For this reason, I also include indicators for these admissions rounds. The results are not sensitive to dropping this lottery. See the online appendix for more detail.

[15] These include random initial offers of attendance, random offers from the waitlist, lottery numbers, and waitlist numbers. However, 91% of the students in my estimation sample were involved in lotteries that used an initial offer mechanism. In practice, this instrument drives virtually all of the results I will present.

Table 3: Test Score Results

	Math				English Language Arts			
	OLS	RF	FS	2SLS	OLS	RF	FS	2SLS
Choice	−0.005			0.16	−0.05			−0.01
	(0.04)			(0.08)	(0.04)			(0.08)
Initial Offer		0.08	0.51			0.001	0.51	
		(0.04)	(0.03)			(0.04)	(0.03)	
Waitlist Offer		−0.10	0.96			−0.19	0.96	
		(0.13)	(0.05)			(0.21)	(0.05)	
Lottery Number		−0.004	0.01			0.004	0.01	
		(0.004)	(0.004)			(0.004)	(0.004)	
Waitlist Number		0.01	0.02			0.01	0.02	
		(0.01)	(0.01)			(0.01)	(0.01)	
F-Stat Excluded Instruments			120.9	120.9			120.1	120.1
Observations	1705	1705	1705	1705	1705	1705	1705	1705
Observations (students)	966	966	966	966	969	969	969	969
Adjusted R^2	0.67	0.67	0.33	0.66	0.56	0.56	0.33	0.56

would be to cluster at the school-by-grade level, as in Angrist, Pathak & Walters (2013) or at the level of the lottery. In practice, neither of these alternatives materially change the standard errors.

4.2 District Choice Benefits the Average Student who Participates

I begin with results on test scores. Table 3 shows ordinary least squares, reduced form, first stage and two-stage least squares results side-by-side for my baseline specification. The two-stage least squares estimates imply that the causal effect of moving to a more preferred district is to increase math test scores by 0.16 standard deviations. There is no detectable impact on English Language Arts.

The effects in table 3 are large in both absolute terms and relative to the existing literature on choice between traditional public schools. The unadjusted black-white test score gap in Massachusetts in 2016 was 0.56σ; hence the point estimate from inter-district choice represents approximately 30% of that gap. Prior lottery evaluations of choice between traditional public schools have examined the impact that attending a student's most preferred school has on test scores within the context of large, urban district assignment algorithms. In that environment, attending a most preferred school in does not typically impact test scores (Cullen, Jacob & Levitt 2006, Hastings, Neilson & Zimmerman 2012, Deming et al. 2014). For

Table 4: Coursework Results

	Class Type Indicator			
	AP	Remedial	General	Advanced
Choice	0.14	−0.07	0.01	0.20
	(0.06)	(0.03)	(0.01)	(0.05)
Mean Dependent Variable	0.19	0.09	0.99	0.28
Observations	809	2,418	2,418	2,418
Observations (students)	470	933	933	933
Adjusted R^2	0.26	0.12	0.04	0.36

additional specifications where I include pre-determined student level controls, as well as robustness checks using student fixed effects to achieve identification via trend changes across lottery winners and losers, see the online appendix.

Next I examine the impact that moving to a more preferred district has on coursework. For the coursework regressions, I am forced to drop a small number of students that only appear in the sample frame prior to the first year MA DESE kept student level records on courses taken. Table 4 presents results from the baseline two stage least squares using as an outcome indicators for whether the student was enrolled in coursework labeled as Advanced Placement (AP), Remedial, General, or Advanced. AP classes consist of a nationally recognized curriculum known for rigor and college preparedness. Remedial, General, and Advanced are designations from the state of Massachusetts. When examining AP coursework, I restrict the sample to years when students appear in grades 11 and 12, since access to AP coursework is uncommon at earlier grades.

Table 4 tells a consistent story: moving to a more preferred district increases the quality of coursework that a student takes. There is a substantial increase in the probability that students enroll in advanced and AP coursework, and a moderate decrease in the probability that a student enrolls in a remedial class. In the online appendix, I present additional results on coursework using intensive margin variation which suggests that the pattern of substitution moves children from remedial to general coursework, and from general to advanced.

Finally, I present results pertaining to the impact of inter-district choice on graduation and college attendance. For table 5, I restrict the data to the sample of students whose on-time graduation date relative to their lottery grade-year is 2016 or prior. Since the estimates are imprecise, I present both the reduced form and two-stage least squares estimates. The point

Table 5: Post-Secondary Outcomes Results

	Post-Secondary Outcome					
	Graduate		Attend-2yr		Attend-4yr	
Initial Offer	0.02		−0.05		0.04	
	(0.03)		(0.05)		(0.04)	
Choice		0.03		−0.08		0.06
		(0.05)		(0.08)		(0.07)
Observations (Students)	518	518	518	518	518	518
F-Stat Exluded Instruments		226.75		226.75		226.75
Dependent Variable Mean	0.88	0.88	0.39	0.39	0.61	0.61
Adjusted R^2	0.05	0.05	0.06	0.06	0.21	0.21

estimates from table 5 suggest that students who participate in inter-district choice are more likely to graduate from high-school and less likely to attend a two-year college. However, the decline in two-year attendance is approximately compensated for by an increase in four-year college attendance. This suggests that lottery winners are substituting four year college for two year college. Combined with the results on coursework, it is tempting to conclude that this is coming from the increase in college application competitiveness that access to advanced and AP coursework bestows upon lottery winners. However, this is purely speculative. It is not possible to rule out other potential mechanisms or even the absence of an effect.

5 School Quality and External Validity

A minimum definition of school quality is that it is equal to the expected test score gain a student randomly selected from the population would experience if sent to that institution.[16] It follows that to credibly relate estimates of test score gains from choice lotteries to school quality, we need to know if the local average treatment effect (LATE) identified with the lottery is equal to the average treatment effect (ATE) for the relevant student population. Thus whether, and to what degree, the program evaluation results presented in section 4 communicate information about school quality is at its core a question about externally validity.

[16] I call this a minimum criterion because, in the presence of treatment effect heterogeneity, it is not obvious how to properly define school quality. A stronger, but somewhat more natural, criterion would be that a school is higher quality if it benefits every student in the population relative to the reference school; however, the weaker criterion is still a reasonable measure for many practical applications despite the fact that optimal policy should, to the greatest degree possible, account for observed heterogeneity rather than rely on averages.

Of particular concern for the external validity of choice lottery estimates is the potential for test score gains to emerge from Roy selection. Simple models of economic behavior would predict that families should use inter-district choice to sort students to schools on the basis of potential benefit (Hoxby 2000). This selection on gains will drive a wedge between the LATE and the ATE by ensuring that students with higher average benefit are disproportionately likely to apply for inter-district choice, accept admissions offers, and subsequently remain in the program. Thus school choice can generate positive test score gains even when there are no quality differences across schools.

It is possible to test for this sorting under weak conditions. Consider the following simple version of the potential outcome framework:

$$y_i = d_i y_i^1 + (1 - d_i) y_i^0 = \beta_i d_i + y_i^0 \tag{3}$$

Were y_i is the observable test score of student i, d_i is a treatment indicator denoting whether the student accepted an offer to switch schools, (y_i^1, y_i^0) represents the student's test score in the treated and control state respectively, and $\beta_i = y_i^1 - y_i^0$ is the benefit of the program to student i. Let τ_i denote an indicator for whether or not a student applied to the program.

Then a necessary condition for the LATE to be externally valid is that application and take-up behavior are unrelated to potential benefit:

$$\beta_i \perp (d_i, \tau_i) \tag{4}$$

In general, a linear extrapolation is appropriate to any sub-sample of the population where this condition holds. Hence, I will refer to condition (4) as weak linearity.

With a rich model of heterogeneity, I can test weak linearity under weak conditions. Without loss of generality, suppose I am interested in testing for selection on post-lottery take-up behavior (d_i). Then weak linearity implies that $\mathbb{E}(\beta_i d_i) = 0$; however, β_i is unknowable and hence we cannot test this implication directly. Instead, let $k = k(X_i)$ be an injective mapping between covariates and student types as indexed by k. Suppose $\beta_i = \beta_k + \upsilon_i$ where β_k is the treatment effect for students of type k. Now I can test whether:

$$\mathbb{E}(\beta_k d_i) = 0 \tag{5}$$

A finding that $\mathbb{E}(\beta_k d_i) \neq 0$ would imply a violation of weak linearity except in the knife-edge case where the correlation between take-up behavior and the observable heterogeneity is exactly off-set by the correlation between take-up and the unobserved heterogeneity.[17] In practice, this is the test I will take to my data in section 7. In order to implement it, however, I will first need to estimate the observable heterogeneity (β_k).

6 Estimating Treatment Effect Heterogeneity

In order to understand the relation between potential benefit, application, and take-up behavior, I need to estimate a rich model of treatment effect heterogeneity. However, my estimation sample is only moderately sized ($\approx 1,000$ students), and I am using a noisy estimation procedure (two-stage least squares). This makes it difficult to precisely estimate the necessary number of interaction terms.

To overcome this technical challenge, I develop a new empirical-Bayes type estimator that uses non-experimental data to increase the precision of quasi-experimental estimates. The model assumes a hierarchical structure for the heterogeneity. This allows the posterior mean of the experimental estimates to incorporate information from the non-experimental data. The resulting estimator swaps noisy experimental variation for precise non-experimental variation according to the correlation of the heterogeneous effects across samples. The estimator is consistent under the same conditions as IV and, under the joint normality assumption required for the hierarchical model, it is more efficient. I also provide simulation evidence that the estimation procedure dominates standard methods on mean squared error.

6.1 A Hierarchical Model of Heterogeneous Effects

Suppose that we wish to estimate treatment effect heterogeneity in a population with I observations. Further, assume that a subset of size E from this population are exposed to some quasi-experiment, while the remaining $N = I - E$ are not. Let $k = k(X_i)$ be an injective mapping between covariates X_i and a student's type as indexed by k.

Suppose we are interested in estimating the following model:

[17] More precisely, $\mathbb{E}(\beta_k d_i) = -\mathbb{E}(v_i d_i)$ implies that it is possible to find $\mathbb{E}(\beta_k d_i) \neq 0$ even when the data generating process exhibits no selection on gains.

$$y_i = \beta_i d_i + u_i \tag{6}$$

$$\beta_i = \beta_k + v_i \tag{7}$$

Here, β_k is the local average treatment effect for individuals of type k identified via the quasi-experiment (e.g. a lottery design). Let $\hat{\beta}_k^e$ denote the estimate of β_k from the quasi-experiment, and let $\hat{\beta}_k^n$ denote an estimate using only observational data (e.g. a lagged test score model using the N observations not exposed to the experiment). Let the joint asymptotic distribution of the estimators be given by:

$$\begin{bmatrix} \hat{\beta}_k^e \\ \hat{\beta}_k^n \end{bmatrix} \overset{a}{\sim} \mathbb{N}\left(\begin{bmatrix} \beta_k \\ \beta_k + b_k \end{bmatrix}, \Omega_k \right) \tag{8}$$

Where b_k is the difference between the local average treatment effect β_k identified by the quasi-experiment and the estimand of the observational design. Note that up to this point, we have not assumed anything beyond what is ordinarily required for identification and inference.

In general, the econometrician may prefer the experimental estimates because with a compelling quasi-experiment these should be unbiased (or at least consistent) for the local average effect of interest. However, if the experimental sample E is small, or if the quasi-experiment requires a noisy technique such as IV (or both), the estimated heterogenous effects may still be far from the local average effect due to sampling variation. At the same time, the non-experimental estimates will be inconsistent for the local average effect in general. Despite this fact, the non-experimental estimates can still contain valuable information useful for pinning down the heterogeneous effects in the experimental sample. Intuitively, realizations of the estimators $(\hat{\beta}_k^e, \hat{\beta}_k^n)$ that are highly correlated are unlikely to emerge from chance alone. Hence such a realization should give the econometrician more confidence that the point estimates from the experiment are close to the local average effect of interest. The following model formalizes this intuition.

Assume a hierarchical model for the estimands of the experimental and non-experimental designs:

$$\begin{bmatrix} \beta_k \\ \beta_k + b_k \end{bmatrix} \sim \mathbb{N}\left(\begin{bmatrix} \beta_0 \\ \beta_0 + b_0 \end{bmatrix}, \Sigma \right) \tag{9}$$

Where β_0 is the center of the distribution of the heterogeneous effects identified by the experiment, and b_0 is the difference between the centers of the experimental and non-experimental distributions. The assumption that the estimands are jointly normal induces a Bayesian structure:

$$\mathbb{P}\left(\begin{bmatrix} \beta_k \\ \beta_k + b_k \end{bmatrix} \middle| \begin{bmatrix} \hat{\beta}_k^e \\ \hat{\beta}_k^n \end{bmatrix} \right) \propto \mathbb{P}\left(\begin{bmatrix} \hat{\beta}_k^e \\ \hat{\beta}_k^n \end{bmatrix} \middle| \begin{bmatrix} \beta_k \\ \beta_k + b_k \end{bmatrix} \right) \mathbb{P}\left(\begin{bmatrix} \beta_0 \\ \beta_0 + b_0 \end{bmatrix} \right) \tag{10}$$

With the parent distribution from the hierarchical model taking on the role of the prior. Specifying the joint distribution of the estimands in this way allows the posterior mode of the experimentally identified heterogeneous effects to be influenced by the realization from the non-experimental sample in a way that I will make precise later. First, I discuss identification.

Observe that in order to operationalize this model empirically, I will need values for Ω_k, Σ and $(\beta_0, \beta_0 + b_k)$. One option would be to specify a prior on these parameters and estimate the model in a fully Bayesian framework. Another option, and the one I pursue in this paper, is to estimate these quantities from the data and thus implement the model via an empirical-Bayes procedure. The main advantage of this approach is that I will be able to provide a simple analytical representation of the resulting estimator that makes transparent how the non-experimental variation is used to inform the posterior mode. The center of the joint distribution $(\beta_0, \beta_0 + b_k)$ is identified via the corresponding pooled regressions that assume no heterogeneity (i.e. $\beta_k = \beta$). The population covariance matrix Σ is identified by calculating the cross-design variance-covariance matrix: $cov(\hat{\beta}^e, \hat{\beta}^n)$. And the joint asymptotic covariance matrix is calculated from the residuals of the experimental and non-experimental heterogeneous effects regressions.[18] For more detail, see the online appendix.

From equation (10), we can use standard properties of the multi-variate normal distribution to calculate the posterior-mode of β_k as follows:

[18] When the quasi-experimental estimates are generated via ordinary least squares (as opposed to IV or 2SLS), this is analogous to estimating the covariance matrix of a seemingly unrelated regression model via Zellner (1962).

$$\beta_k^s = \beta_0 + \alpha_k(\hat{\beta}_k^e - \beta_0) + \delta_k(\hat{\beta}_k^n - \beta_0 - b_0) \tag{11}$$

Equation (11) consists of three terms. The first term (β_0) anchors the estimator to the center of the experimental distribution. The next two terms consist of a weighted average of the experimental variation in the heterogeneous effects ($\hat{\beta}_k^e - \beta_0$) and the non-experimental variation ($\hat{\beta}_k^n - \beta_0 - b_0$). For now, assume the off-diagonal elements of Ω_k are zero as would typically be the case when the observations in the experimental data are not also included in the non-experimental data.[19] The weights are given by:

$$\alpha_k = \frac{\phi_n^k - \rho^2}{\phi_n^k \phi_e^k - \rho^2} \tag{12}$$

$$\delta_k = \frac{\rho \frac{(\omega_e^k)^2}{\sigma_e \sigma_n}}{\phi_n^k \phi_e^k - \rho^2} \tag{13}$$

Where $\rho \equiv corr(\beta_k, \beta_k + b_k)$ is the correlation between the experimental and non-experimental estimands and $\phi_j^k \equiv \frac{\sigma_j^2 + (\omega_j^k)^2}{\sigma_j^2}$ is the inverse of a standard empirical Bayes weight,[20] commonly referred to as the signal to noise ratio. The parameters $(\omega_e^k, \omega_n^k, \sigma_e, \sigma_n)$ come from the diagonals of Ω_k and Σ. When $\rho = 0$, the system decouples and equation (11) reduces to a standard empirical-Bayes estimator applied to the experimental data alone. Otherwise, the resulting estimate is a mixture of the two sources of variation. I show in the online appendix that after plugging in the empirical counterparts for $(\alpha_k, \delta_k, \beta_0, \beta_0 + b_0)$, the resulting posterior modes are consistent under the same conditions as IV[21] and, under the normality assumption on the parent distribution, more precise than standard two-stage least squares. I also provide simulation evidence that the consensus estimates using all the data dominate the individual estimators (and their decoupled empirical-Bayes counterparts) on mean squared error. See the online appendix for more detail.

[19] For a more general expression, see the online appendix.

[20] Here $j = e$ and $j = n$ refer to the experimental and non-experimental weights respectively

[21] While this is true, a better model for the large sample behavior of this estimator might be to fix the ratio of the sample size between the experimental and non-experimental data. This should slow the rate of convergence for the experimental sample and thus preserve the experimental / non-experimental sample size disparity in the limit. However, this is left for future work.

6.2 Estimating Student Heterogeneity in Practice

In practice, I want to estimate a rich model of student level treatment effect heterogeneity using all of the available covariates at my disposal. However, some of these covariates are continuous or have many support points. Thus constructing indicators for student types based on their full interaction is infeasible. For this reason, I assume the heterogeneity takes the following form:

$$\beta_{it} = \beta_{k(X_{it})} + \upsilon_{it} = \alpha_0 + \alpha X_{it} + \upsilon_{it} \tag{14}$$

The vector X_{it} includes student age, indicators for race / ethnicity; lagged values for attendance, days suspended, and test scores; and lagged indicators for whether the student received a subsidized lunch or was diagnosed with a disability.

Moving to a two-stage least squares framework, this yields the following model for the experimental data:

$$y_{it} = \delta_0 + \delta_\ell + \beta_k d_{it} + \gamma_w W_i + \gamma_x X_{it} + \epsilon_{it} \tag{15}$$

$$\beta_k = \alpha_0^e + \alpha^e X_{it} + \upsilon_{it} \tag{16}$$

$$d_{it} = \delta_0' + \delta_\ell' + \pi_0 Z_{it} + \pi X_{it} Z_{it} + \gamma_w' W_i + \gamma_x' X_{it} + \eta_{it} \tag{17}$$

Note that in equation (16), I have added the superscript e to distinguish the important parameters estimated from the experimental data from those estimated using the non-experimental data (which I will superscript by n). In practice, I plug equation (16) into equation (15) and proceed with two-stage least squares to estimate (α_0^e, α^e) via the corresponding interaction terms.

For the non-experimental data, I consider the following model:

$$y_{it} = \delta_h + \delta_g + \delta_t + \beta_k d_{it} + \theta_x X_{it} + u_{it} \tag{18}$$

$$\beta_k = \alpha_0^n + \alpha^n X_{it} + \upsilon_{it}' \tag{19}$$

Where δ_h, δ_g, and δ_t are home district, grade, and academic year fixed effects respectively. In practice, I plug equation (19) into equation (18) and proceed with ordinary least squares to estimate (α_0^n, α^n) via the corresponding interaction terms. Thus the comparison I have

Table 6: Comparison of Pooled Models

	Standardized Math Test Score	
	2SLS	OLS
Choice	0.19	0.08
	(0.08)	(0.003)
F-Stat Excluded Instruments	149.79	
Observations	1,705	6,549,952
Observations (Students)	966	1,784,770
Adjusted R^2	0.68	0.44

in mind with equation (18) is between two children who would by default be assigned to the same grade and district during academic year t and who have similar values for the covariates X_{it}; however, the first child has left the home district via inter-district choice ($d_{it} = 1$), while the second has not ($d_{it} = 0$). Note that I drop all students used in the quasi-experiment to estimate (α_0^e, α^e) from the observational sample used to estimate (α_0^n, α^n).

Before proceeding to the fully heterogeneous models, I first present a comparison of estimates from the fully pooled versions that assume no heterogeneity (i.e. $\alpha^e = \alpha^n = 0$). The coefficients on the treatment indicator from these pooled regressions are the estimates of β_0 and $\beta_0 + b_0$ that I use in the parent distribution when estimating the cross-design posterior-modes. Table 6 contains the results. Note that the estimate of β_0 here using two-stage least squares is different from the estimate of β_0 found in the program evaluation due to the inclusion of the vector X_{it} in equation (15). The non-experimental estimate appears to indicate a moderate benefit to participating in inter-district choice. However, the point estimates across designs are quite far apart.

Next I estimate the fully heterogeneous models. Figure 6 plots the predicted treatment effects from the non-experimental model against the predicted treatment effects from the experimental model over the support points of X_{it} contained in the experimental data.[22] While the two sets of estimated treatment effects are not one to one, there is still a moderately strong relationship between them. The correlation between the two sets of estimates is 0.35. This suggests that knowledge of the heterogeneous effects from the non-experimental model is informative about the value we would expect in the experimental model. Hence, it seems

[22]To be precise, the experimental treatment effect is given by $\hat{\beta}_k^e = \hat{\alpha}_0^e + \hat{\alpha}^e X_{it}$ and the non-experimental treatment effect is given by $\hat{\beta}_k^n = \hat{\alpha}_0^n + \hat{\alpha}^n X_{it}$ where X_{it} comes from an observation in the lottery sample.

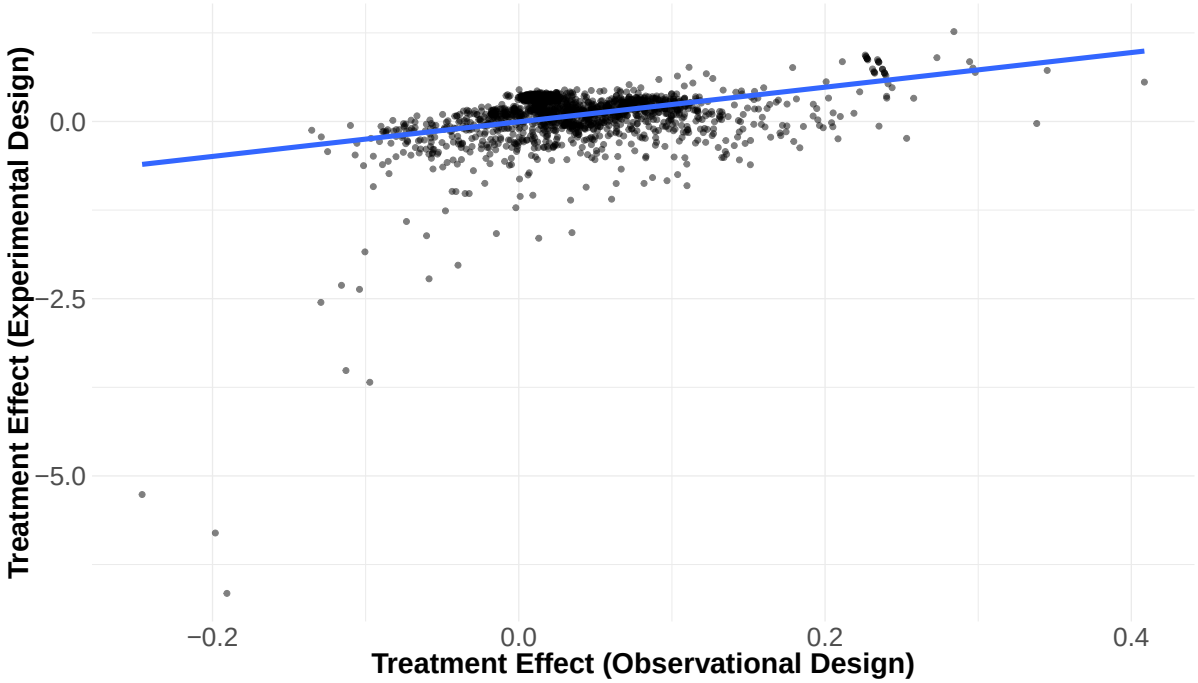

Figure 6: Correlation Across Experimental and Non-Experimental Models

reasonable to use a hierarchical model to incorporate information from the non-experimental data into the estimates.

Next I estimate the consensus posterior-modes. Figure 7 provides a visualization of how the estimator mixes the two sources of information in practice. For each support point X_{it} in the experimental data, figure 7 plots its rank in the distribution of experimental treatment effects against the predicted treatment effect from the experimental model (denoted by purple circles), the non-experimental model (denoted by green triangles), and the consensus posterior mode (denoted by yellow squares).[23] Thus we can observe directly, for each observation in the data, how much mixing occurs between the experimental and non-experimental predicted values.

Finally, I will note here that in principle is also possible to extract implied consensus regression coefficients from the estimated posterior modes.[24] Thus it is also possible to see directly how much of the consensus estimate is driven by the underlying sources of heterogeneity. This is useful when the sources of heterogeneity are themselves relevant for policy. In

[23] I trim a small number of observations whose predicted value in the experimental sample would be less than negative one. I do this to keep the scale of the y-axis small, which makes it easier to see in the figure how the consensus posterior modes mix the corresponding experimental and non-experimental estimates.

[24] These are just a linear combination of the posterior modes $\hat{\alpha}^s = (X'X)^{-1}X'\hat{\beta}^s$ and hence have a known distribution.

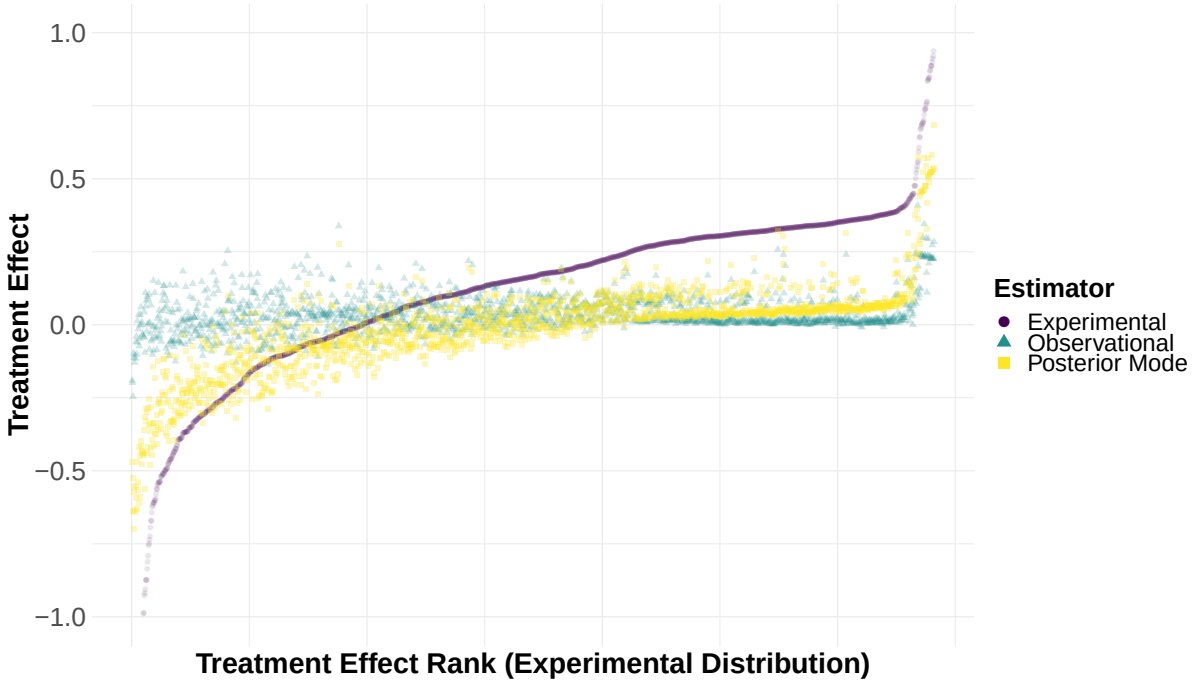

Figure 7: Visualizing Cross-Design Mixing

the online appendix, I provide simulation evidence that both the predicted heterogeneous effect and the implied consensus regression coefficients do better than their IV, OLS, or standard empirical-Bayes counterparts on mean squared error relative to the corresponding population values. For a more detailed examination of the factors driving the observed heterogeneity in my data, see the online appendix.

7 Inter-district Choice and Roy Selection

In this section, I examine the consequences that Roy selection has for the interpretation of the program evaluation LATE identified with the lottery. To test for selection on gains, I examine three phases of the admissions process and, in each case, I find that treatment effect heterogeneity is predictive of the take-up decision. First I examine the sub-population of students who have already taken up offers to switch districts. I find that students who are negatively impacted by the program are more likely to subsequently return to their home district. Second, I re-examine the first stage of the two-stage least squares estimates from the program evaluation. I find that holding constant the outcome of the lottery, students who are negatively impacted by the program are less likely to take-up treatment. Third, I extrapolate treatment effects to the pool of students who were eligible to apply for the school choice slots

in my lottery data. I find that students who would be negatively impacted by the program are less likely to apply.

I conclude by using observed heterogeneity to extrapolate the average benefit of inter-district choice to students that took up offers of treatment, to students that applied, and to students that did not apply. I find that 38% of the treatment effect for the treated comes from post-lottery selection into enrollment and that 78% of the treatment effect for applicants is driven by pre-lottery selection into the applicant pool. Almost none of the lottery LATE is attributable to differences in average quality across districts.

7.1 Testing for Selection on Gains

Recall from section 5 that for a lottery estimate to identify educational quality differences, it must be the case that weak linearity holds: individual benefit (β_i) is unrelated to both pre-lottery application behavior (τ_i) and post-lottery take-up behavior (d_i). Hence we should expect to find patterns of treatment effect heterogeneity that are consistent with no selection on gains: $\mathbb{E}(d_i\beta_i) = \mathbb{E}(\tau_i\beta_i) = 0$. Since individual potential benefit (β_i) is unobserved, I cannot test for selection on gains directly. Instead, I will test $\mathbb{E}(\beta_k d_i) = 0$ and $\mathbb{E}(\beta_k \tau_i) = 0$, where β_k is the observable heterogeneity. This is a valid test for selection on gains provided we rule out the knife edge case where the correlation between take-up behavior and unobserved heterogeneity exactly off-sets the correlation between take-up behavior and the observed heterogeneity.

This discussion motivates tests of weak linearity via models of the following form:

$$d_i = \alpha + \rho\beta_k + \epsilon_i \tag{20}$$

Where $\rho \neq 0$ indicates a failure of weak linearity, and $\rho > 0$ implies Roy selection. However, the parameter ρ is difficult to interpret directly since β_k is measured in units of standardized test score gains.

Another natural test of selection on gains is to ask whether students who would be negatively impacted by the treatment are less likely to apply or to take it up. This motivates models of the following form:

$$d_i = \alpha + \rho\mathbb{1}(\beta_k < 0) + \epsilon_i \tag{21}$$

Here $\rho \neq 0$ implies a violation of weak linearity, with $\rho < 0$ indicating Roy selection. Speci-

fications like (21) have the advantage of a straightforward interpretation.

7.2 Assessing the Impact of Roy Selection

First, I restrict the sample to students that I use for lottery estimation and who also accept an offer to enroll in a district outside of their home district. I then restrict the data to student-years after the first post-lottery year and estimate the following model:

$$d_{it} = \delta_g + \delta_d + \delta_t + \rho\hat{\beta}_k + \epsilon_{it} \tag{22}$$

Where δ_g, δ_d and δ_t are grade, district, and academic year fixed effects; d_{it} is an indicator for whether student i participated in choice in year t; and $\hat{\beta}_k$ is the estimated heterogeneous treatment effect. With model (22), the comparison I have in mind is between two students who accepted lottery offers and are now attending school under the choice program in the same grade, district and year. The parameter ρ tells me whether students with high potential test score gains are more likely to remain in the program relative to those with low potential test score gains.

To look at the participation decision, I use the entire lottery estimation sample and revisit the first stage of the two-stage least squares, but now including $\hat{\beta}_k$ as a predictor:

$$d_{it} = \delta'_0 + \delta'_\ell + \rho\hat{\beta}_k + \pi Z_{it} + \gamma'_w W_i + \eta_{it} \tag{23}$$

The comparison I have in mind with model (23) is between two students who entered the same lottery and had a similar lottery outcome. The parameter ρ tells me whether students with high potential benefit are more likely to take up treatment than those with low potential benefit.

Finally, I wish to compare the potential benefit of students who applied to the inter-district choice program to those who were eligible to apply but did not. In theory, every student in the state is eligible to enter every lottery. In practice, commuting costs make it unreasonable for students to apply to choice spots far away from their home. To find a reasonable group of comparison students, I use the empirical distribution of home districts for each lottery[25] and only consider students in the relevant grades / districts. Since the pool of eligible students

[25] In other words, if only students from districts A and B appear in lottery 1, I only consider students from districts A and B as lottery eligible for the purposes of finding a comparison group.

is large, I use a randomly chosen 1% sub-sample within grade, year, and district. I consider all students that appear in my lottery estimation sample as having applied.[26] I then estimate models of the following form:

$$\tau_i = \delta_g + \delta_d + \delta_t + \rho\hat{\beta}_k + \epsilon_i \tag{24}$$

Where δ_g, δ_d and δ_t are grade, district, and year fixed effects and τ_i is an indicator for whether student i did, in fact, enter the lottery for which they were eligible. With model (24), the comparison I have in mind is between two students currently in the same grade, district, and year who are eligible to enter one of the lotteries in my sample. The parameter ρ tells me whether the students with high potential benefit are more likely to apply.

For all three models, I also estimate specifications where I replace $\hat{\beta}_k$ with an indicator for negative potential benefit $\mathbb{1}(\hat{\beta}_k < 0)$. As I argued in the previous section, the magnitudes in these models are easier to interpret. For a general discussion of the procedure I used to estimate $\hat{\beta}_k$, see section 6. To ensure there is no mechanical correlation between the participation indicators (d_{it}, τ_i) and the estimated heterogeneity $(\hat{\beta}_k)$, I calculate the heterogeneous effects for these models using a leave-lottery out jack-knife procedure (in the case of the observations in the lottery data) or a split sample procedure (in the case of the non-experimental observations). See the online appendix for more detail. I calculate asymptotic standard errors clustered at the student level and, to account for the increased variability introduced by the generated regressor, I also calculate standard errors using the parametric bootstrap by resampling from the distribution of $\hat{\beta}_k$. In all cases, I choose the most conservative value.

Table 7 reveals important selection at each stage of the admissions and enrollment process. Students with a negative predicted treatment effect are 17% less likely to apply. Conditional on applying and receiving a randomly assigned offer, they are 5% less likely to enroll. Conditional on enrolling, they are 8% less likely to continue in the program. These results continue to hold in the reduced form linear specification.

Taken together, the results in table 7 suggest it is unlikely that potential benefit is unrelated to application and take-up. This implies that the program evaluation LATE is not externally valid and hence unrelated to school quality. However, if the component of selection on gains that is driven by the sorting of students to schools is small, it is possible that the LATE identified by the lottery is still "close" to the quantity of interest in the sense that

[26] I continue to exclude students that received preferences in the lottery or applied late, and I also continue to exclude students that were missing a baseline test score since I am unable to calculate the necessary heterogeneous effect.

Table 7: Testing for Selection on Gains

	Take-up Indicator					
	Continue		Participate		Apply	
Heterogeneous Effect	0.10		0.15		0.32	
	(0.07)		(0.10)		(0.05)	
Heterogeneous Effect < 0		−0.08		−0.05		−0.17
		(0.03)		(0.04)		(0.02)
Subsample	Ever-Enrolled	Ever-Enrolled	Applicants	Applicants	Eligible	Eligible
Observations	860	860	1,621	1,621	2,730	2,730
Observations (students)	395	395	894	894	2,730	2,730
Dependent Variable Mean	0.85	0.85	0.46	0.46	0.38	0.38
Adjusted R^2	0.38	0.38	0.32	0.32	0.21	0.22

the majority of the estimated effect could still be driven by average quality differences across schools.

To quantify the magnitude of the wedge induced by the Roy selection, I average the predicted heterogeneous effects for three sub-populations: the treated, the applicants, and the non-applicants. For this exercise to be valid, the extrapolation from the complier population to the applicants and non-applicants must be accurate conditional on the observed heterogeneity. This will be the case when there is no selection on the unobserved heterogeneity: $v_i \perp (d_i, \tau_i) = 0$. This assumption is unlikely to be true. However, I note that this assumption is strictly weaker than the stronger weak linearity assumption that implicitly drives much of the interpretation of lottery estimates in the literature. Thus the exercise generates value by demonstrating in practice how far from the truth estimates that do not account for heterogeneity can be.

I find that virtually all of the test score gains generated by inter-district choice are driven by selection. The average treatment effect on the treated[27] is $.11\sigma$, the average treatment effect for applicants is $.08\sigma$, and the average treatment effect on non-applicants $.02\sigma$. This suggest that 38% of the treatment on the treated comes from post-lottery selection into the program, and that 78% of the treatment effect for applicants is driven by selection into the applicant pool.[28] The point estimates suggest that at most 18% of the LATE can be attributed

[27] There are three possible explanations for why the estimate here is lower than the program evaluation LATE: 1) it is constructed with the consensus posterior modes and hence shrunk towards the non-experimental estimate, 2) it is a student weighted average as opposed to a conditional variance weighted average, and 3) it includes the extrapolated effect to always takers, instead of being solely based on compliers.

[28] There were over 170,000 eligible applicants which is large relative to the number that applied ($\approx 1,000$). Hence the treatment for the non-applicants is effectively the population average treatment effect in this case.

to differences in average quality across sending and receiving districts.[29] Finally, I will note that if there is also selection on unobserved heterogeneity, we would expect the extrapolated estimates presented here to be an upper bound. Thus I cannot rule out the possibility that the entirety of the program evaluation LATE is the result of sorting.

8 What Can Lotteries Say About School Quality?

In this paper, I have shown how the sorting of students to school districts on the basis of potential benefit leads to lottery estimates of test score gains that have no straightforward connection to school quality. I accomplish this in three steps. First, I document that the inter-district choice program is substantially beneficial to students who participate. Inter-district choice increases math test scores and the quality of coursework students take as well as increasing the probability a student graduates from high-school and goes on to attend a four year college. Next, I provide a new method for estimating treatment effect heterogeneity. This method leverages information contained in non-experimental data by positioning the heterogeneity within a hierarchical model. The resulting estimator is a weighted average of experimental and non-experimental variation, with the weights chosen according to the correlation of the heterogeneous effects across samples. Finally, I show that the heterogeneous treatment effects associated with inter-district choice predict student take-up behavior in a manner that is consistent with Roy selection. I find that this Roy selection is responsible for almost the entirety of the program evaluation treatment effect identified with the lottery. Taken together, these results suggest that research using lotteries to identify school quality should exercise caution with regard to the external validity of their estimates.

The fact that families sort students to districts on the basis of potential benefit fits within a broader pattern of facts in the literature which suggest that some of the gains to charter attendance are conditional on initial selection into a large urban district. Within Boston, charter takeovers and expansion generate lottery gains commensurate with already established charters (Abdulkadiroglu et al. 2016, Cohodes, Setren & Walters 2018). This suggests that the charter model generates a real quality difference for students within Boston. However, the effect of charters in Massachusetts outside of urban areas is negative (Angrist, Pathak & Walters 2013). Indeed, a recent meta-analysis of charter effectiveness found that controlling for the quality of a student's fall-back option attenuates much of the effect of factors associated with the highly-touted set of charter teaching practices known as the "No Excuses" philoso-

[29] This is based on the ratio of the treatment for the treated and the ATE. If I use the program evaluation LATE instead of the treatment on the treated, this number would change to 12.5%.

phy (Chabrier, Cohodes & Oreopoulos 2016). This is consistent with the idea that selection across districts is an important mediator of effective educational practices. Why selection at this more aggregate level leads to an equilibrium where some students in urban areas appear to be so poorly served by the teaching methods of the traditional public education system relative to charters is an open question.

Last, I will note that the patterns of heterogeneity and selection I find across districts call into question the use of test scores for the purpose of evaluating and ranking schools. As pointed out in Hoxby (2000), simple Tiebout models imply that in equilibrium students should be sorted among districts based on school types and individual ability to benefit. In a world where test-score gains are driven by more aggregate levels of sorting, ranking schools on the basis of test score gains is unlikely to be a useful exercise. Put simply, there are no straightforward policy implications from the fact that Jane experiences smaller test score gains at the school where she is best suited than Jill experiences at the school where she is best suited. On the other hand, leveraging heterogeneity to design an education system that provides students opportunities to better match with the education type that best suits them seems like a promising area for future work.

References

Abadie, Alberto, and John F. Kennedy. 2003. "Semiparametric instrumental variable estimation of treatment response models." *Journal of Econometrics*, 113: 231–263.

Abadie, Alberto, Susan Athey, Guido Imbens, and Jeffrey Wooldridge. 2017. "When Should You Adjust Standard Errors for Clustering?"

Abdulkadiroglu, Atila, Joshua D. Angrist, Peter D. Hull, and Parag A. Pathak. 2016. "Charters without Lotteries: Testing Takeovers in New Orleans and Boston." *American Economic Review*, 106(7): 1878–1920.

Abdulkadiroglu, Atila, Joshua D. Angrist, Susan M. Dynarski, Thomas J. Kane, and Parag A. Pathak. 2011. "Accountability and Flexibility in Public Schools: Evidence from Boston's Charters And Pilots." *The Quarterly Journal of Economics*, 126(2): 699–748.

Abdulkadiroglu, Atila, Joshua D. Angrist, Yusuke Narita, and Parag A. Pathak. 2017. "Research Design Meets Market Design: Using Centralized Assignment for Impact Evaluation." *Econometrica*, 85(5): 1373–1432.

Abdulkadiroglu, Atila, Parag A. Pathak, and Christopher R. Walters. 2018. "Free to Choose: Can School Choice Reduce Student Achievement?" *American Economic Journal: Applied Economics*, 10(1): 175–206.

Angrist, Joshua D., and Kevin Lang. 2004. "Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program." *American Economic Review*, 94(5): 1613–1634.

Angrist, Joshua D., Parag A. Pathak, and Christopher R. Walters. 2013. "Explaining Charter School Effectiveness." *American Economic Journal: Applied Economics*, 5(54): 1–27.

Angrist, Joshua D., Peter D. Hull, Parag A. Pathak, and Christopher R. Walters. 2017. "Leveraging Lotteries for School Value-Added: Testing and Estimation." *The Quarterly Journal of Economics*, 132(2): 871–919.

Angrist, Joshua D., Sarah R. Cohodes, Susan M. Dynarski, Parag A. Pathak, and Christopher R. Walters. 2016. "Stand and Deliver: Effects of Boston's Charter High Schools on College Preparation, Entry, and Choice." *Journal of Labor Economics*, 34(2): 275–318.

Angrist, Joshua D., Susan M. Dynarski, Thomas J. Kane, Parag A. Pathak, and Christopher R. Walters. 2012. "Who Benefits from KIPP?" *Journal of Policy Analysis and Management*, 31(4): 837–860.

Bergman, Peter. 2018. "The Risks and Benefits of School Integration for Participating Students: Evidence from a Randomized Desegregation Program."

Chabrier, Julia, Sarah Cohodes, and Philip Oreopoulos. 2016. "What Can We Learn from Charter School Lotteries?" *Journal of Economic Perspectives*, 30(3): 57–84.

Chester, Mitchell D. 2014. "Building on 20 Years of Massachusetts Education Reform." Massachusetts Department of Elementary and Secondary Education.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review*, 104(9): 2633–2679.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane W. Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom

Affect Your Earnings? Evidence from Project Star." *The Quarterly Journal of Economics*, 126(4): 1593–1660.

Cohodes, Sarah, Elizabeth Setren, and Christopher Walters. 2018. "Can Successful Schools Replicate? Scaling Up Boston's Charter School Sector."

Cullen, Julie B., Brian A. Jacob, and Steven Levitt. 2006. "The Effect of School Choice on Participants: Evidence from Randomized Lotteries." *Econometrica*, 74(5): 1191–1230.

Deming, David J., Justine S. Hastings, Thomas J. Kane, and Douglas O. Staiger. 2014. "School Choice, School Quality, and Postsecondary Attainment." *American Economic Review*, 104(3): 991–1013.

Dobbie, Will, and Roland G. Fryer. 2011. "Are High-Quality Schools Enough to Increase Achievement Among the Poor? Evidence from the Harlem Children's Zone." *American Economic Journal: Applied Economics*, 3(3): 158–187.

Dobbie, Will, and Roland G. Fryer. 2013. "Getting Beneath the Veil of Effective Schools: Evidence From New York City." *American Economic Journal: Applied Economics*, 5(4): 28–60.

Dobbie, Will, and Roland G. Fryer. 2015. "The Medium-Term Impacts of High-Achieving Charter Schools." *Journal of Political Economy*, 123(5): 985–1037.

Hastings, Justine, Christopher Neilson, and Seth Zimmerman. 2012. "The Effect of School Choice on Intrinsic Motivation and Academic Outcomes." *NBER Working Paper*, 18324.

Howell, William G., Patrick J. Wolf, David E. Campbell, and Paul E. Peterson. 2002. "School vouchers and academic performance: results from three randomized field trials." *Journal of Policy Analysis and Management*, 21(2): 191–217.

Hoxby, Caroline, and Sonali Murarka. 2009. "Charter Schools in New York City: Who Enrolls and How They Affect Their Students' Achievement." *NBER Working Paper*, 14852.

Hoxby, Caroline M. 2000. "Does Competition Among Public Schools Benefit Students and Taxpayers?" *American Economic Review*, 90(5): 1209–1238.

Hull, Peter. 2018. "Estimating Hospital Quality with Quasi-experimental Data."

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467.

Meager, Rachael. 2017. "Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature."

Meager, Rachael. 2018. "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments." *American Economic Journal: Applied Economics*, forthcoming.

Mills, Jonathan N., and Patrick J. Wolf. 2017. "Vouchers in the Bayou: The Effects of the Louisiana Scholarship Program on Student Achievement After 2 Years." *Educational Evaluation and Policy Analysis*, 39(3): 464–484.

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky. 2018. "Using Instrumental Variables for Inference about Policy Relevant Treatment Effects." *Econometrica*, forthcoming.

Moody, Sandra. 1994. "Advisory Opinion on School Choice." Massachusetts Department of Education.

O'Connell, Scott. 2017. "School choice initiative drains $ from hard-hit districts, critics say." *The Telegram and Gazette*.

Walters, Christopher R. 2017. "The Demand for Effective Charter Schools." *Journal of Political Economy*.

Wixom, Micah. 2016. "Open Enrollment: Overview and 2016 Legislative Update." Education Commission of the States, Denver.

Wolf, Patrick, Babette Gutmann, Michael Puma, Brian Kisida, Lou Rizzo, and Nada Eissa. 2008. "Evaluation of the DC Opportunity Scholarship Program: Impacts after Two Years." National Center for Education Evaluation and Regional Assistance.

Young, Alwyn. 2017. "Consistency without Inference: Instrumental Variables in Practical Application."

Zellner, Arnold. 1962. "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias." *Journal of the American Statistical Association*, 57(298): 348–368.