The Consequences of Sorting for Understanding School Quality

Jesse Bruhn*

August 11, 2020

Click here for a version with in-text figures/tables.

Abstract

I study the sorting of students into 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 0.19 standard deviations. The program also improves coursework quality and increases the probability of high school graduation and college attendance. Motivated by these findings, I develop a feature-rich model of treatment effect heterogeneity and estimate it using a new application of empirical Bayes. The estimator I propose improves accuracy by synthesizing information contained in the choices of students who are exposed to the lottery with information contained in the choices of students who are not. I use the heterogeneous effects to examine selection into the choice program. Students who benefit from the program are more likely to apply, and conditional on taking up an offer to enroll, they are more likely to continue on in the program beyond their first year. I find that this Roy-type selection drives nearly 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 a 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, Brown University. jesse_bruhn@brown.edu. / 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; the Boston University Labor Reading group; the Northeast Economics of Education Workshop; and the applied micro seminars at the University of Rochester, Queen's University; Syracuse University, the University of Chicago Booth School of Business, Notre Dame, the University of Toronto, Colby College, Brandeis University, the University of California San Diego, Rutgers University, Brown University, Kansas State University, and Princeton University 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.

1 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 et al., 2006; Chetty et al., 2011; Angrist et al., 2012; Chetty et al., 2014; Deming et al., 2014; Dobbie and Fryer, 2015; Angrist et al., 2016). Understanding institutional quality is therefore important for effectively targeting educational investments.

Recent work on educational effectiveness leverages randomization embedded within the school assignment process to estimate quality differences across institutions. Since the outcomes of lotteries are random, estimates of school quality based on a comparison between applicants who win a school choice lottery and applicants who lose a school choice lottery 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 institutional value added, validate observational methods of ranking schools, and estimate the relation between effectiveness and educational inputs (Angrist et al., 2013; Dobbie and 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. It could be that some schools are better for each student or that each school is better for some students. In either case, a school choice lottery may make the receiving school look effective. But to design the best system, the policymaker must know which mechanism is at work. In general, school choice lottery estimates are not externally valid if the decision to apply to an institution is correlated with the student's potential benefit (Walters, 2018). Thus knowing whether and to what degree lottery-identified test score gains are driven by Roy selection versus differences in quality is necessary for understanding the practical policy implications of this body of academic 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 participation in inter-district school choice on student outcomes using new lottery data. The second contribution is an examination of Roy selection and its role in generating the causal effects of inter-district choice. The third contribution is a new method for estimating treatment effect

heterogeneity using empirical Bayes.

I start with a program evaluation of inter-district school choice in Massachusetts. The program allows primary school students to apply for open seats in neighboring school districts. Importantly, the law requires seats to be rationed via a lottery whenever there is excess demand. I use new data on the outcomes of these lotteries to identify the causal effect of participating in inter-district choice. I find that moving to a more preferred district increases student math scores by 0.19 standard deviations, with no effect on English Language Arts (ELA). The impact on math is large. Angrist et al. (2013) find that, on average, choosing to attend a charter middle school or high school in Massachusetts generates math test score gains of 0.21 and 0.27 standard deviations, respectively. My results also stand in contrast to prior estimates of the effect of school choice in the traditional public school sector, which find little to no impact on test scores (see, e.g., Cullen et al., 2006; Hastings et al., 2012; Deming et al., 2014). I find that participation in the inter-district choice program increases the probability that students take Advanced Placement (AP) and other advanced classes. I also find positive but imprecise effects on the probability that students graduate from high school and go on to attend a four-year college.

The findings from this evaluation are important because they represent the first lottery evaluation of a statewide inter-district choice program. Such programs are common in the United States (Wixom, 2016) and are also controversial. Critics argue that because funding for this type of program typically follows the student, inter-district choice drains educational resources from underprivileged communities (O'Connell, 2017). Thus, quantifying the benefits of inter-district choice is important for understanding the value of these programs for the students who choose to use them. Prior work has examined the impact of within-district urban assignment mechanisms, choice to charter schools, private school vouchers, and race-based desegregation programs.[1]

Next I examine Roy selection in the inter-district choice context. 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 student behavior measures. Students of lower socioeconomic status, students with disabilities, and students with discipline problems appear to be hurt by the program, while older, higher

[1]For recent examples of choice among traditional public schools, see Cullen et al. (2006), Hastings et al. (2012), Deming et al. (2014), and Abdulkadiroglu et al. (2017). For recent examples of choice to the charter sector, see Hoxby and Murarka (2009), Abdulkadiroglu et al. (2011), Dobbie and 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 and Wolf (2017), and Abdulkadiroglu et al. (2018). For race-based programs, see Angrist and Lang (2004) and Bergman (2018).

ability students tend to benefit. I find that the observed heterogeneity predicts student take-up behavior that is consistent with Roy selection. Students who would be positively impacted by the program are more likely to apply. Conditional on applying, those who benefit from the program are more likely to take up a randomly assigned offer to enroll, and once enrolled, those who benefit are more likely to continue on after their first year.

Roy selection is significant because it drives a wedge between the local average treatment effect (LATE) identified by the lottery and the average treatment effect (ATE) of interest: district quality. To quantify the wedge's magnitude, I use the observed heterogeneity to extrapolate the ATE for the treated, the applicants and the non-applicants. I find that 45% of the treatment effect for the treated comes from postlottery selection into enrollment, and 67% of the treatment effect for applicants is driven by prelottery selection into the applicant pool. Almost none of the LATE identified with the lottery is the result of quality differences across districts.

This finding is important because it provides insight into the potential domain of external validity for a body of work that uses school choice lotteries to study educational effectiveness (Angrist et al., 2013, 2017; Dobbie and Fryer, 2013; Deming et al., 2014; Abdulkadiroglu et al., 2017). For example, Angrist et al. (2017) uses school choice lotteries within the Boston school district to estimate the benefit of closing a low value-added school. However, real world accountability systems such as the one actually employed in Massachusetts often compare schools across district boundaries. In fact, the Massachusetts Board of Elementary and Secondary Education (DESE) cited low levels of student test score growth relative to the rest of the state among the reasons they voted to place the Holyoke Public School District in receivership in 2015 (Massachusetts Department of Elementary and Secondary Education, 2015). State-level accountability systems that use value added for decision-making will not necessarily generate welfare gains in the presence of sorting across districts. Prior work has also used lotteries to argue that certain educational practices are generally effective mediators of educational quality (Angrist et al., 2013; Dobbie and Fryer, 2015). Scaling up these practices will unlikely be able to generate the anticipated benefits if the original gains identified from the lotteries emerged in part from a sorting mechanism.[2]

The final contribution of this paper is a new application of empirical Bayes to the esti-

[2]There is compelling evidence that the sorting mechanism is less important within large urban districts. For example, Deming et al. (2014) demonstrates that students in the Charlotte-Mecklenburg school district in North Carolina only benefit from school choice when they gain access to a higher quality school. Further, Angrist et al. (2017) provides evidence based on reweighting methods that school effects within the Boston Public School district are approximately linear. Finally, Abdulkadiroğlu et al. (2020) find that, conditional on peer quality, parental preferences are uncorrelated with school effectiveness or match quality in New York City. In principle, there is nothing inconsistent between these findings and the present work.

mation of treatment effect heterogeneity. This application leverages non-experimental data to more accurately estimate heterogeneous treatment effects in a quasi-experimental design. To study the sorting of students to districts on the basis of potential benefit, I must first fit a feature-rich heterogeneous treatment effects model using an instrumental variables (IV) strategy. Using IV is necessary to correct for postlottery selection on the margin of treatment take-up. Unfortunately, IV designs are notoriously noisy (Young, 2017), making accurately estimating the heterogeneous effects difficult with the lottery sample at my disposal. However, I show that corresponding estimates of the heterogeneous effects using observational data on the universe of public school students in Massachusetts are highly correlated with the estimates from the experimental sample. This finding suggests that the non-experimental data contains information useful for pinning down the LATEs identified by the lottery. I formalize this intuition by combining the experimental and non-experimental treatment effect estimates within a hierarchical model. I show that the estimator is consistent under the same conditions as IV, and I provide simulations that suggest the procedure dominates among other common estimators with respect to the mean squared error.

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 (see, e.g., Angrist et al., 2017; Hull, 2018; Meager, 2017, 2018; Chetty and Hendren, 2018; Rothstein, 2018). In particular, the method outlined in Angrist et al. (2017) is closely related. The authors 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 Angrist et al. (2017)'s model. However, I am only interested in gains with respect to accuracy, whereas they use the non-experimental data to extrapolate treatment effects to schools that are not oversubscribed. As a result, I do not need to model the first-stage, reduced-form, and bias jointly within the parent distribution of the hierarchical model, allowing me to find a closed-form representation of the estimator with a simple, transparent intuition. And unlike Angrist et al. (2017), my approach allows for the possibility that the LATE identified via the lottery is different from the ATE in the population. This is important because assuming equality between the LATE and ATE would effectively assume away the Roy selection I am looking for.

2 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 of 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.[3] Figure 1 shows enrollment in the inter-district choice and the charter sector over time.

[Figure 1 about here.]

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 they do, the district is not required to enroll students from other districts. Nonetheless, 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[4] in Massachusetts have enrolled at least one student via the program, with 156 districts 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, the student's family is responsible for transportation.[5] 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.

[3] Both calculations are my own and were made using administrative student micro-data provided by the DESE.

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

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

[Figure 2 about here.]

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 nondiscrimination 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,[6] and there are a number of districts that are regularly oversubscribed yet conduct admissions on a first-come, first-serve basis.[7] Finally, there are some portions of the law that simply never made it into practice. For example, the original bill asked participating districts to submit their enrollment and capacity projections to the DESE. I learned from my conversations with state-level program administrators that this information has never been collected.

3 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 college outcomes via an extract from the National Student Clearinghouse (NSC). I also make use of several spreadsheets provided to me by the 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. I will now briefly discuss each of my primary data sources in turn. For a more detailed discussion of the primary data sources, as well as more detailed descriptions of the less frequently used data sources, see Appendix A.1.

3.1 New Lottery Data

In May 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

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

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

with me.[8] Of the districts I contacted, approximately 75% responded, and of the districts that responded, 36% confirmed they had ever conducted a lottery. Typically, districts that did not conduct a lottery were not oversubscribed. A small number of districts accepted new students using a first-come, first-serve procedure despite being oversubscribed. 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. Ninety-one percent of the lotteries in my sample used this randomization procedure. 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 that, 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, the results in this paper are driven by initial offers; see Appendix B.2 for results that only use the initial offer instrument.

The typical lottery in my sample is small. The average number of students I view in a single lottery is 9.6, and the median is 7. The lotteries also span a considerable time period. The earliest lottery in my data occurs in academic year 2002–2003, while 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 postlottery. The lotteries in my sample also span all grade levels. However, as Figure 3 shows, the lotteries are clustered at grades that are typically within-district, cross-school transition points for students.[9] For more detailed descriptive statistics regarding the

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

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

lottery data, including histograms of lotteries by year and size, see Appendix A.3.

[Figure 3 about here.]

I merge these student lottery records to the data provided by the 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, hometown 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 unmatched and drop her from the sample. When this procedure fails to find any exact match, I repeat it using fuzzy first and last name matching. For this reason, all of my specifications include indicators for whether a student was matched via the exact or fuzzy version of the algorithm. Overall, I obtain an 89% match rate. For further discussion of the procedure used to match the lottery data to the state data, see Appendix A.2.

My lottery sample exhibits some imbalance along predetermined characteristics. Figure 4 presents point estimates and two standard error intervals from a within-lottery regression[10] of all baseline observable and otherwise exogenous characteristics on the initial offer indicator for the subsample of students where I observe at least one test score prior to the lottery. The joint F-statistic across all predetermined characteristics is 2.[11] Of particular concern is the fact that the coefficient for black students is negative and the two standard error interval does not include zero. However, I note that the administrators conducting the lottery could not directly observe race,[12] the coefficient's magnitude 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 Appendix A.4, I consider the possibility that this imbalance is driven by differential attrition and conclude that this is also unlikely to be the case.

[Figure 4 about here.]

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

[10] Within lottery is the level of variation at which the instrument is randomly assigned. I leverage this variation by including lottery fixed effects in the regression. I also drop from this regression all students who received sibling preference or were indicated as applying late.

[11] Rounded to the hundredth decimal place.

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

lottery data did so voluntarily and described to me in detail their randomization process, and there was no consequence for opting to not share data with me. Second, cheating would open the district up to potentially serious liability. As I discussed in Section 2, 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. Thus if a district was cheating, it had a strong incentive to not provide me with data.

One potential 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 Appendix A.5 that conditioning on earlier prelottery test scores increases my precision substantially, and more importantly, 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 students for whom I observe at least one test score prior to the lottery year and will include baseline test scores as controls.

3.2 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 information 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 included 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. Unless otherwise noted, I drop observations from the state data that appear in nontraditional public school environments. These include collaborative schools, charter schools, vocational schools, agricultural schools, adult education programs, virtual schools, institutional schools, and residential/deaf programs.

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[13] level to have a mean of zero and a standard deviation of one. I retrieve coursework taken by students from Student Course Schedule (SCS) data spanning

[13]The state transitioned testing regimes from the original MCAS exam to the PARCC exam over the course of my sample frame. There are three years in my data where the old and new examinations appear simultaneously. For this reason, all regressions will also include test-type fixed effects.

academic years 2010–2011 through 2016–2017. I also use data on college attendance contained in an extract from the NSC purchased by the 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. For further description of the various data sources along with a detailed breakdown of the cleaning process, see Appendix A.1.

4 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 Students" provides averages of observable characteristics 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 statewide sample to observations where a student is currently participating in inter-district choice. The column labeled "Sending Districts" restricts the statewide sample to districts that lose a student to choice via a lottery I observe in my data. The column labeled "Lottery Sample" restricts the statewide sample to students found in my lottery data as observed in the year when they applied.

[Table 1 about here.]

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 and 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 subsamples, likely due to the fact that the family is responsible for transportation to the new district. 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 to subsequently accept lottery offers.

At the district level, the decision to not 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, likely due to the fact that many suburban districts in the Boston area participate in the Metropolitan Council for Educational Opportunity

(METCO) program. As discussed in Angrist and Lang (2004), METCO is the nation's oldest voluntary school desegregation program, and 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 using district characteristics to predict 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 ordinary least squares (OLS), and column (2) displays select results from the variables chosen when estimation is performed using post-Lasso. Column (3) displays select results from a joint regression that also includes district and year fixed effects; in effect, the column 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 that predict changes in participation status. See Appendix B.4 for complete results including the variables not displayed in Table 2.

[Table 2 about here.]

Finally, 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, and the losers tend to be urban and rural.

[Figure 5 about here.]

5 Program Evaluation

In this section, I evaluate the benefits of inter-district choice for students who participate. For identification, I examine applicants to oversubscribed districts and compare the district choice lottery winners to the district choice lottery losers within a two-stage least squares (2SLS) framework. I find that participating in district choice causes large test score gains in math, and I find no effect on ELA scores. I also find that participating in district choice increases

the quality of the coursework that students take. Finally, I provide suggestive evidence that participating in district choice increases the probability a student will graduate from high school and attend a four-year college.

5.1 Identification and Estimation

Consider the following 2SLS framework:

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

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

where y_{it} denotes the outcome of student i during a postlottery period of time t (typically an academic year), δ_0 is an intercept, 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,[14] W_i are covariates observed at baseline,[15] and Z_i denotes the vector of four lottery instruments[16] discussed in Section 3.

The parameter β identifies a 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 and Angrist, 1994). Exclusion requires that the lottery's result affects potential outcomes only via take-up of the treatment, and random assignment requires that within each lottery the results are in fact random. The first stage requires that the lottery results 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 lottery results behave consistently with respect to take-up. Provided these four conditions are

[14] 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.

[15] 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 (see Appendix B.3). 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 Appendix B.3).

[16] 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. See Appendix B.2 for results when the sample is restricted to students exposed to initial offer lotteries.

satisfied, β is properly interpreted as the ATE of moving to a more preferred school district for lottery compliers who applied to oversubscribed districts that maintained and were willing to share high quality lottery records. I save a discussion of heterogeneity and external validity for Section 6.

I restrict the sample to the set of students appearing in my lottery data such that I observe at least one pre- and one postlottery 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 from that lottery 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 would be to cluster at the school-by-grade level, as in Angrist et al. (2013) or at the lottery level. In practice, neither of these alternatives materially change the standard errors.

5.2 District Choice Benefits the Average Student Who Participates

I begin this section with the results on test scores. Table 3 shows OLS, reduced-form, first-stage, and 2SLS results side-by-side for my baseline specification. The 2SLS estimates imply that the causal effect of moving to a more preferred district is to increase math test scores by 0.19 standard deviations. There is no detectable impact on ELA.

[Table 3 about here.]

The effects in Table 3 are large in both absolute terms and relative to the existing literature on choice between traditional public schools. For example, Angrist et al. (2013) find that, on average, choosing to attend a charter middle and high school in Massachusetts generates math test score gains of 0.21 and 0.27 standard deviations, respectively. For ELA, Angrist et al. (2013) find that choosing to attend a charter middle school in Massachusetts increases test scores by 0.075 standard deviations, and for high schools they find increases of 0.206 standard deviations. 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 does not typically impact test scores (Cullen et al., 2006;

Hastings et al., 2012; Deming et al., 2014). For additional specifications where I include predetermined student-level controls, as well as specifications that use student fixed effects and year-by-lottery fixed effects to achieve identification via a comparison of trend changes across winners and losers within lottery, see Appendix B.1. All results in Table 3 are robust to these more demanding specifications.

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 who only appear in the sample frame prior to the first year the DESE kept student-level records on courses taken. Table 4 presents results from the baseline 2SLS using, as an outcome, indicators for whether the student was enrolled in coursework labeled as AP, remedial, general, or advanced. AP classes consist of a nationally recognized curriculum known for rigor and college preparedness, and remedial, general, and advanced are designations from the state of Massachusetts. When examining AP coursework, I restrict the sample to years when students progressing normally would appear in grades 11 and 12, since access to AP coursework is uncommon at earlier grades.

[Table 4 about here.]

Table 4 tells a consistent story: moving to a more preferred district increases the quality of the student's coursework. 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 Appendix B.5, I present additional results on coursework using intensive margin variation that suggests the pattern of substitution moves students 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 2SLS estimates. The point estimates from Table 5 suggest that students who participate in inter-district choice are more likely to graduate from high school and are 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, suggesting 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.

[Table 5 about here.]

6 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.[17] It follows that to credibly relate estimates of test score gains from choice lotteries to institutional quality, we need to know if the LATE identified with the lottery is equal to the ATE for the relevant student population. Thus whether, and to what degree, the program evaluation results presented in Section 5 communicate information about school quality is, at its core, a question about externally validity.

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 choice programs 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 to the program, accept admission offers, and subsequently remain in the program after the first year. Thus school choice can generate positive test score gains even when there are no quality differences across insitutions.

It is possible to test for this sorting under weak conditions. Consider the following simple model of potential outcomes:

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

Here 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 program's benefit 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:

[17] 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.

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

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

Existing literature has looked for violations of weak linearity by testing for "selection on gains" via generalized Roy models (Walters, 2018; Hull, 2018; Mogstad et al., 2018). This approach relies on the existence of a continuous instrument that shifts different student types in and out of compliance as a means of identifying the joint distribution of (β_i, d_i) (Heckman and Vytlacil, 2001; Cornelissen et al., 2016). In the school choice context, prior work has argued that distance to receiving institution is conditionally randomly assigned to student families and is also excluded from the outcome equation, and hence it can serve as an instrument (e.g., Walters, 2018). Unfortunately, I do not observe student addresses in the state data, and it is inconsistently available in the lottery data, so I am unable to calculate precise measures of student distance to receiving institution. In addition, I note that while it may be plausible that distance to a school is randomly assigned to families within small geographic areas such as Boston, it seems intuitively less plausible that distance is randomly assigned across the entire state of Massachusetts conditional on covariates, which is what would be required to use these methods here. For these reasons, I am unable to apply the generalized Roy framework in this context.

However, observe that with a sufficiently rich model of observable treatment effect heterogeneity, I can still look for evidence of selection on gains under weak conditions. Without loss of generality, suppose I am interested in testing for selection on postlottery 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 $\beta_i = \beta_k + v_i$, where β_k is the treatment effect for type k students. Note that β_k is, in principle, identified from the data. 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 offset by the correlation between take-up and the unobserved heterogeneity.[18] In practice, this is the test I will take to my data in Section 8. To implement it, however, I will first need to estimate the observable heterogeneity (β_k).

7 Estimating Treatment Effect Heterogeneity

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 (2SLS). This makes it difficult to precisely estimate the necessary number of interaction terms.

To overcome this technical challenge, I develop a new application of empirical Bayes that uses non-experimental data to fully leverage the available information when forming quasi-experimental estimates of treatment effect heterogeneity. The model assumes a hierarchical structure for the heterogeneity, allowing the posterior mode of the experimental estimates to incorporate information from 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 I provide simulation evidence that the estimation procedure dominates standard methods, as measured via the mean squared error over the collection of heterogeneous effects.

In this section I also explore the drivers of the observed heterogeneity. I find that students of lower socioeconomic status, students with disabilities, and students with discipline problems appear to be hurt by the program, while older, high-ability students tend to benefit.

7.1 A Hierarchical Model of Treatment Effect Heterogeneity

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:

[18] More precisely, $\mathbb{E}(\beta_k d_i) = -\mathbb{E}(v_i d_i)$ implies it is possible to find $\mathbb{E}(\beta_k d_i) \neq 0$ even when the data-generating process (DGP) 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 LATE for type k individuals 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 LATE 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 LATE 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 heterogeneous effects may still be far from the LATE due to sampling variation. At the same time, the non-experimental estimates may be inconsistent for the LATE in general as a result of some form of selection. Nevertheless, the non-experimental estimates can still contain valuable information useful for pinning down the heterogeneous effects in the experimental sample. Intuitively, highly correlated realizations of the estimators $(\hat{\beta}_k^e, \hat{\beta}_k^n)$ 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 LATE 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 exper-

iment, and b_0 is the difference between the centers of the experimental and non-experimental distributions. The assumption that the estimands are random induces a Bayesian structure:

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

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 to operationalize this model empirically, I will need values for Ω_k, Σ, and $(\beta_0, \beta_0 + b_0)$. One option would be to specify a prior on these parameters and to 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 to 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 $(\beta_0, \beta_0 + b_0)$ is identified via the corresponding pooled regressions that assume no heterogeneity (i.e., $\beta_k = \beta$), and the population covariance matrix Σ is identified via the residuals of the pooled and unpooled models.[19] The joint asymptotic covariance matrix is calculated from the residuals of the experimental and non-experimental heterogeneous effects regressions.[20] For more detail, see Appendix C.

From equation (10), we can use standard properties of the multivariate normal distribution to calculate the posterior mode of β_k in the experimental data as follows:

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

Equation (11) consists of three terms. The first term (β_0) anchors the estimator to the cen-

[19] Intuitively, the residuals of the treated units from the pooled model contain the observable heterogeneity plus sampling variation, while the residuals of the treated units for the unpooled model contain only the sampling variation; hence the difference in residuals for treated units across models contains only the heterogeneity, and thus the variance of the heterogeneous effects may be calculated in a straightforward manner. This generalizes naturally to the case of multiple quasi-experimental/observational designs. See Appendix C for mathematical detail.

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

ter 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.[21] Then 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,[22] 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 Appendix C 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.[23]

I also provide simulation evidence that the consensus estimates using all of the data dominate the individual estimators (and their decoupled empirical Bayes counterparts) in terms of the mean squared error over the collection of heterogeneous effects. This is true even under certain violations of the normality assumption on the parent distribution, and it is true when the procedure is applied to a DGP calibrated to match the actual data/model I use for estimation in the next section. See Appendix C for more detail.

7.2 Estimating Student Heterogeneity 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

[21] For a more general expression, see Appendix C.

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

[23] 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. Doing 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.

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:

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

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. In Appendix C.3, I provide simulation evidence that models assuming a linear approximation for the heterogeneity perform well on simulated data calibrated to match my actual data despite the fact that the heterogeneous effects' linearity implies a potential for violations of the normality assumption on the parent distribution.

Moving back to the 2SLS framework, the linearity assumption yields the following model for the experimental data:

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

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

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

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). To estimate the model, I plug equation (16) into equation (15) and proceed with 2SLS to recover (α_0^e, α^e) via the corresponding interaction terms.

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

$$y_{it} = \delta_{hgt} + \beta_k d_{it} + \theta_x X_{it} + u_{it} \tag{18}$$

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

where δ_{hgt} is a home district (h) by grade (g) by academic year (t) fixed effect. To estimate the model, I plug equation (19) into equation (18) and proceed with OLS to recover (α_0^n, α^n) via the corresponding interaction terms. Thus the comparison I have 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 models 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 shows the results. Note that the estimate of β_0 here using 2SLS is not mechanically identical to the estimate of β_0 found in the program evaluation due to the inclusion of vector X_{it} in equation (15).

[Table 6 about here.]

In general, 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. This could be due to uncontrolled selection bias in the non-experimental sample, a LATE that diverges from the ATE in the lottery sample, sampling variation, or a mixture of all three.

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.[24] While the two sets of estimated treatment effects are not one-to-one, there is still a moderately strong relation between them (a correlation of 0.35), especially when considering that the measurement error in the estimates will tend to drive the slope toward zero. This strong relation 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 reasonable to use a hierarchical model to incorporate information from the non-experimental data into the estimates.

[Figure 6 about here.]

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

[24] 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.

purple circles), the non-experimental model (denoted by green triangles), and the consensus posterior mode (denoted by yellow squares).[25] Thus we can observe directly, for each observation in the data, how much mixing occurs between the experimental and non-experimental predicted values.

[Figure 7 about here.]

7.3 What Factors Predict the Observed Heterogeneity?

In this section, I show that students of lower socioeconomic status, students with disabilities, and students with discipline problems appear to be hurt by the program, while older, high-ability students tend to benefit. To explore which covariates drive the observed heterogeneity, I use the posterior modes from the preceding section to estimate the treatment effect for various subsamples of the data relative to the average student in the sample.[26] Figure 8 displays the results. Each point in the figure shows the difference between the treatment effect for the indicated subgroup and the remaining students in the sample. The lines represent two standard error intervals, which I calculate under the distributional assumptions of the hierarchical model.[27] Figure 8 appears to tell a consistent story: students who would traditionally be considered "high-needs" are hurt upon switching schools, while higher ability, older students tend to reap the benefits. Demographic variables are largely unrelated to the treatment effect's size.

[Figure 8 about here.]

Why this particular pattern emerges is impossible to know definitively with the data at my disposal. However, given that the flows of students in the sample are largely from districts with a high concentration of high-needs students to districts with high concentrations of "high test score" students (see Table 1), it is tempting to speculate that the mix of educational inputs in sending/receiving districts are more heavily tailored to match the needs of

[25] I trim a small number of observations from the figure whose predicted value in the experimental sample would be less than −1. I do this to keep the scale of the y-axis small, which makes it easier to visualize how the consensus posterior modes mix the corresponding experimental and non-experimental estimates in practice.

[26] To be precise, I project the consensus posterior modes onto an intercept and a subgroup indicator (e.g., $\hat{\beta}_k = \alpha + \omega d_k + u_k$, where $\hat{\beta}_k$ is the posterior mode of the heterogeneous effect for a type k student, d_k is an indicator for whether the student belongs to the subgroup, and ω is the average difference between the students who are and are not in the subgroup).

[27] The subgroup differences are a linear combination of the posterior modes and hence have a known distribution under the assumptions of the hierarchical model.

the particular student populations that have sorted into them. For example, the sending districts may have more guidance counselors on staff to address disciplinary/behavioral issues and have more teachers who specialize in helping students with disabilities. In turn, the receiving districts may make it easier to take advanced coursework because they have a higher concentration of students with high test scores. In fact, this latter possibility is reinforced via the finding from Section 5 that winning an inter-district lottery causes students to take more advanced classes.

More broadly, this interpretation of the results is consistent with additional evidence in the literature. For example, there is compelling evidence from a developing country context that shows school-student matches are important drivers of test score gains (Bau, 2019). When these patterns of heterogeneity are taken together with the patterns of sorting I document in the next section, this interpretation is also consistent with findings from the experimental literature that parents use information about their child's ability to select appropriately matched educational inputs (Dizon-Ross, 2019). However, I would like to emphasize that it is impossible to know with certainty that this is the mechanism given the data and variation at my disposal. The evidence in favor of this interpretation is entirely circumstantial.

8 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 subpopulation of students who have already taken up offers to switch districts. I find that students who are positively impacted by the program are more likely to continue on in the program; students who are hurt by the program are more likely to return to their home district. Second, I reexamine the first stage of the 2SLS estimates from the program evaluation. I find that holding constant the outcome of the lottery, students who benefit from the program are more 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 are likely to benefit from the program based on observables are also more likely to apply.

I conclude by using the observed heterogeneity to extrapolate the average benefit of inter-district choice to students who took up offers of treatment, to students who applied, and to students who did not apply. I find that 45% of the treatment effect for the treated comes from postlottery selection into enrollment and 67% of the treatment effect for applicants is driven

by prelottery selection into the applicant pool. Almost none of the lottery LATE is attributable to differences in average quality across districts.

8.1 Testing for Selection on Gains

Recall from Section 6 that for a lottery estimate to identify educational quality differences, weak linearity must hold: individual benefit (β_i) is unrelated to both prelottery application behavior (τ_i) and postlottery take-up behavior (d_i). Hence we should expect to find patterns of treatment effect heterogeneity 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.[28] 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 offsets 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 positively impacted by the treatment are more 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. Specifications like (21) have the advantage of a straightforward interpretation.

[28] In principle, if I had access to a continuous instrument, it would also be possible to test for this selection directly using methods from the marginal treatment effects literature (see Cornelissen et al., 2016, for an overview); however, my lottery instrument is binary, and I was unable to isolate another source of compelling variation for this purpose. For this reason, I have chosen to utilize an approach based on observables instead.

8.2 Assessing the Impact of Roy Selection

First, I restrict the sample to students who 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 postlottery year and estimate the following model:

$$d_{it} = \delta_{gdt} + \rho\hat{\beta}_k + \epsilon_{it}, \tag{22}$$

where δ_{gdt} is a grade-by-district-by-academic year fixed effect, 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 recovered in the preceding section. With model (22), the comparison I have in mind is between two students who accepted lottery offers and are now attending school outside of their home district via the choice program in the same receiving district, grade, 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 2SLS,[29] but now include $\hat{\beta}_k$ 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[30] and only consider students in the relevant grades/districts. Since the pool of eligible students is large and the estimation procedure for the heterogeneous effects plus the bootstrap procedure I use for the standard errors is computationally intensive, I form the comparison group using

[29] See the discussion around equation (2) for a complete set of variable definitions for the first-stage equation.

[30] 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.

a randomly chosen 1% subsample within grade, year, and home district. I consider all students who appear in my lottery estimation sample as having applied.[31] I then estimate models of the following form:

$$\tau_i = \delta_{gdt} + \rho\hat{\beta}_k + \epsilon_i, \tag{24}$$

where δ_{gdt} is a grade-by-home-district-by-academic-year fixed effect 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 want to exploit 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 positive 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 7. 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 Appendix C 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 a 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 positive potential benefit are 10% more likely to apply. Conditional on applying and receiving a randomly assigned offer, they are 5% more likely to enroll, and conditional on enrolling, they are 8% more likely to continue on in the program after their first year.

[Table 7 about here.]

Taken together, the results in Table 7 suggest that 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 does not identify average quality differences between sending and

[31] I continue to exclude students who received preferences in the lottery or applied late as well as students who were missing a baseline test score since I am unable to calculate the necessary heterogeneous effect.

receiving districts. 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 lottery-identified LATE is still "close" to the quantity of interest in the sense that the majority of the estimated effect could still be driven by average quality differences across institutions.

To quantify the magnitude of the wedge induced by the Roy selection, I average the predicted heterogeneous effects for three subpopulations: 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: $v_i \perp (d_i, \tau_i)$. This assumption is unlikely to be true. However, I note that this assumption is strictly weaker than the stronger version of no selection on gains that implicitly drives much of the interpretation of lottery estimates in the literature. Thus the present 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 ATE on the treated[32] is 0.11σ, the ATE for applicants is 0.06σ, and the ATE on non-applicants is 0.02σ. This suggests that 45% of the treatment on the treated comes from postlottery selection into the program and 67% of the treatment effect for applicants is driven by selection into the applicant pool.[33] The point estimates suggest that at most, 18% of the lottery LATE can be attributed to differences in average quality across sending and receiving districts.[34] Finally, note that if there is also Roy selection on the 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.

9 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

[32] 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 toward 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.

[33] 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 ATE in this case.

[34] 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 11%.

connection to institutional 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 by 0.19 standard deviations and the quality of coursework students take as well as increases the probability a student graduates from high school and goes on to attend a four-year college.

Next, I provide a new application of empirical Bayes 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 consistent with Roy selection. I find that Roy selection on the basis of observable characteristics can explain 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 beyond school district boundaries.

The fact that families sort students to districts on the basis of potential benefit fits within a broader pattern of facts in the literature that suggest 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 et al., 2018), suggesting the charter model generates a real quality improvement for students within Boston. However, the effect of charters in Massachusetts outside of urban areas is negative (Angrist et al., 2013). Indeed, a recent meta-analysis of charter effectiveness found that controlling for the quality of a student's fallback option attenuates much of the effect of factors associated with the highly touted set of charter teaching practices known as the "No Excuses" philosophy (Chabrier et al., 2016). This finding 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, 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. 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, my results suggest that leveraging heterogeneity to design an education system that encourages better student-school matches across district lines may be a promising area for future work.

References

Abadie, A., Athey, S., Imbens, G., and Wooldridge, J. (2017). When Should You Adjust Standard Errors for Clustering? Working Paper. Standford University, MIT, and Michigan State University.

Abdulkadiroglu, A., Angrist, J. D., Dynarski, S. M., Kane, T. J., and Pathak, P. A. (2011). Accountability and Flexibility in Public Schools: Evidence from Boston's Charters And Pilots. *The Quarterly Journal of Economics*, 126(2):699–748.

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

Abdulkadiroglu, A., Angrist, J. D., Narita, Y., and Pathak, P. A. (2017). Research Design Meets Market Design: Using Centralized Assignment for Impact Evaluation. *Econometrica*, 85(5):1373–1432.

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

Abdulkadiroğlu, A., Pathak, P. A., Schellenberg, J., and Walters, C. R. (2020). Do parents value school effectiveness? *American Economic Review*, 110(5):1502–1539.

Angrist, J. D., Cohodes, S. R., Dynarski, S. M., Pathak, P. A., and Walters, C. R. (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, J. D., Dynarski, S. M., Kane, T. J., Pathak, P. A., and Walters, C. R. (2012). Who Benefits from KIPP? *Journal of Policy Analysis and Management*, 31(4):837–860.

Angrist, J. D., Hull, P. D., Pathak, P. A., and Walters, C. R. (2017). Leveraging lotteries for school value-added: Testing and estimation. *The Quarterly Journal of Economics*, 132(2):871–919.

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

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

Bau, N. (2019). Estimating an Equilibrium Model of Horizontal Competition in Education. CEPR Discussion Paper No. DP13924. University of Toronto.

Bergman, P. (2018). The Risks and Benefits of School Integration for Participating Students: Evidence from a Randomized Desegregation Program. IZA Discussion Paper No. 11602. Columbia University.

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

Chester, M. D. (2014). Building on 20 Years of Massachusetts Education Reform. Technical report, Massachusetts Department of Elementary and Secondary Education.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star. *The Quarterly Journal of Economics*, 126(4):1593–1660.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the impacts of teachers II: teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9):2633–2679.

Chetty, R. and Hendren, N. (2018). The impacts of neighborhoods on intergenerational mobility II: County-level estimates. *The Quarterly Journal of Economics*, 133(3):1163–1228.

Cohodes, S., Setren, E., and Walters, C. (2018). Can Successful Schools Replicate? Scaling Up Boston's Charter School Sector. NBER Working Paper No. 25796.

Cornelissen, T., Dustmann, C., Raute, A., and Schönberg, U. (2016). From LATE to MTE: Alternative methods for the evaluation of policy interventions.

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

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

Dizon-Ross, R. (2019). Parents' Beliefs about Their Children's Academic Ability: Implications for Educational Investments. *American Economic Review*, 109(8):2728–2765.

Dobbie, W. and Fryer, R. G. (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, W. and Fryer, R. G. (2013). Getting beneath the veil of effective schools: Evidence from New York City. *American Economic Journal: Applied Economics*, 5(4):28–60.

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

Gelman, A. and Loken, E. (2013). The Garden of Forking Paths: Why Multiple Comparisons Can be a Problem, Even When there is No 'Fishing Expedition' or 'P-Hacking' and the Research Hypothesis was Posited Ahead of Time. Columbia University and Penn State University.

Hastings, J., Neilson, C., and Zimmerman, S. (2012). The Effect of School Choice on Intrinsic Motivation and Academic Outcomes. *NBER Working Paper*, 18324.

Heckman, J. J. and Vytlacil, E. (2001). Policy-Relevant Treatment Effects. *American Economic Review*, 91(2):107–111.

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

Hoxby, C. and Murarka, S. (2009). Charter Schools in New York City: Who Enrolls and How They Affect Their Students' Achievement. *NBER Working Paper*, 14852.

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

Hull, P. (2018). Estimating Hospital Quality with Quasi-experimental Data. MIT.

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

Massachusetts Department of Elementary and Secondary Education (2015). Holyoke Public Schools Level 5 District Turnaround Plan. Technical report, Massachusetts Department of Elementary and Secondary Education.

Meager, R. (2017). Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature. MIT.

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

Mills, J. N. and Wolf, P. J. (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, M., Santos, A., and Torgovitsky, A. (2018). Using Instrumental Variables for Inference about Policy Relevant Treatment Effects. *Econometrica*, 86(5):1589–1619.

Moody, S. (1994). Advisory Opinion on School Choice. Technical report, Massachusetts Department of Education.

National Student Clearinghouse (2020). StudentTracker - National Student Clearinghouse.

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

Rothstein, J. (2018). Inequality of Educational Opportunity? Schools as Mediators of the Intergenerational Transmission of Income. NBER Working Paper No. 24537.

Walters, C. R. (2018). The Demand for Effective Charter Schools. *Journal of Political Economy*, 126(6):2179–2223.

Wixom, M. (2016). Open Enrollment: Overview and 2016 Legislative Update. Technical report, Education Commission of the States, Denver.

Wolf, P., Gutmann, B., Puma, M., Kisida, B., Rizzo, L., and Eissa, N. (2008). Evaluation of the DC Opportunity Scholarship Program: Impacts after Two Years. Technical report.

Young, A. (2017). Consistency without Inference: Instrumental Variables in Practical Application. London School of Economics.

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

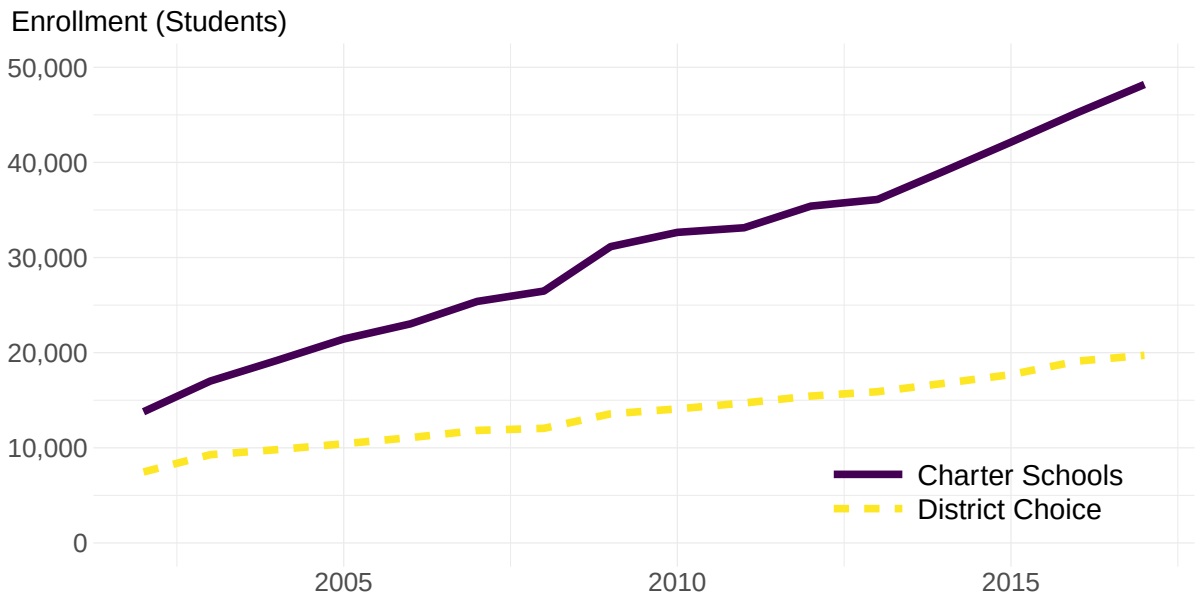

Figure 1: Enrollment in Inter-District Choice and Charter Schools over Time

Note: This figure plots statewide enrollment in the Massachusetts Inter-District Choice Program and charter schools from academic year 2001–2002 to academic year 2016–2017 using data from the Massachusetts Student Information Management System.

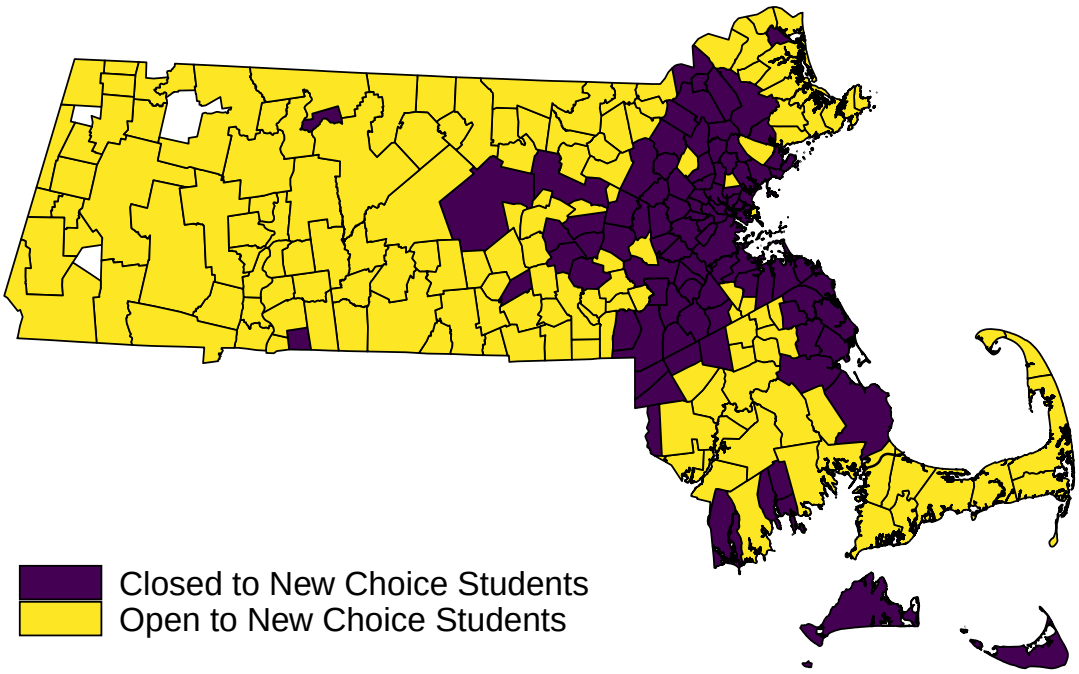

Figure 2: Inter-District Choice in 2016–2017

Note: This figure displays the inter-district choice participation status of Massachusetts' school districts in academic year 2016–2017. Each polygon in the figure represents the boundaries of a 12th grade school district. Note that some districts regionalize at the high school level but not at the primary or middle school level. Since students admitted in primary grades via inter-district choice are eligible to follow the feeder patterns of the receiving district, I aggregate district participation for earlier grades to the appropriate 12th grade boundary. Thus a district is considered "closed" to inter-district choice if all school districts aggregated to the 12th grade boundary voted to opt out of the program in 2016–2017; a district is considered "open" to inter-district choice if at least one district aggregated to the 12th grade boundary did not vote to opt out.

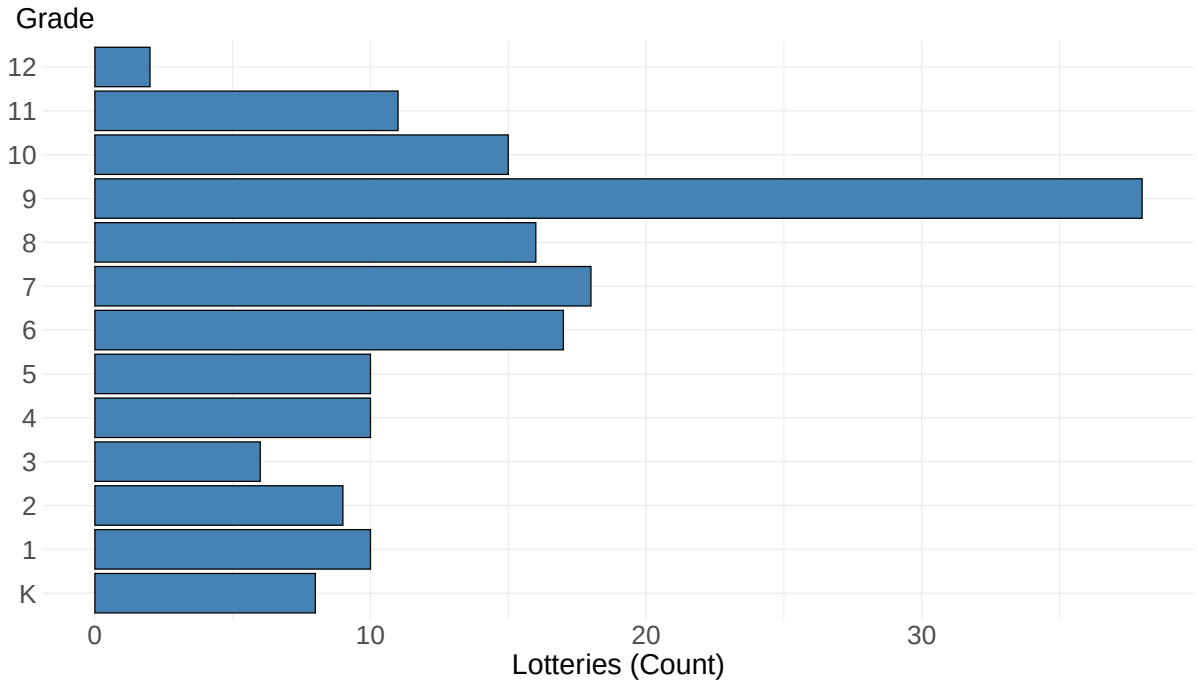

Figure 3: Lottery Distribution by Grade

Note: This figure displays the distribution of lotteries by grade for the districts where I collected data. I define a lottery as occurring at the entry grade by academic year in the receiving district, provided some form of randomization was used to ration an open seat.

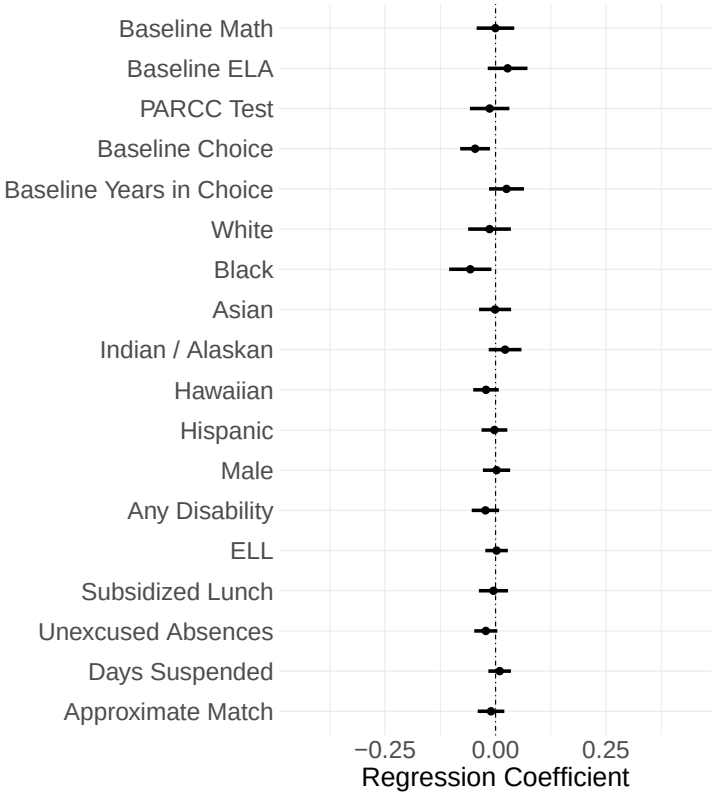

Figure 4: Covariate Balance by Initial Offer Status

Note: This figure visualizes covariate balance by plotting coefficients (β) and two standard error intervals from a regression of the form $d_i = \beta X_i + \delta_\ell + \epsilon_i$, where d_i is an indicator for whether a student received an initial offer from a lottery, X_i is the vector of covariates listed on the y-axis of this figure, and δ_ℓ is a lottery fixed effect. To facilitate comparability across regression coefficients, I standardize the vector of covariates X_i to have a mean of zero and a standard deviation of one prior to estimation. The sample of students includes all students exposed to an initial offer lottery instrument. The x-axis is scaled such that its length reflects two standard deviations of residual variation in d_i after projecting out the lottery fixed effects. Standard errors are clustered at the student level.

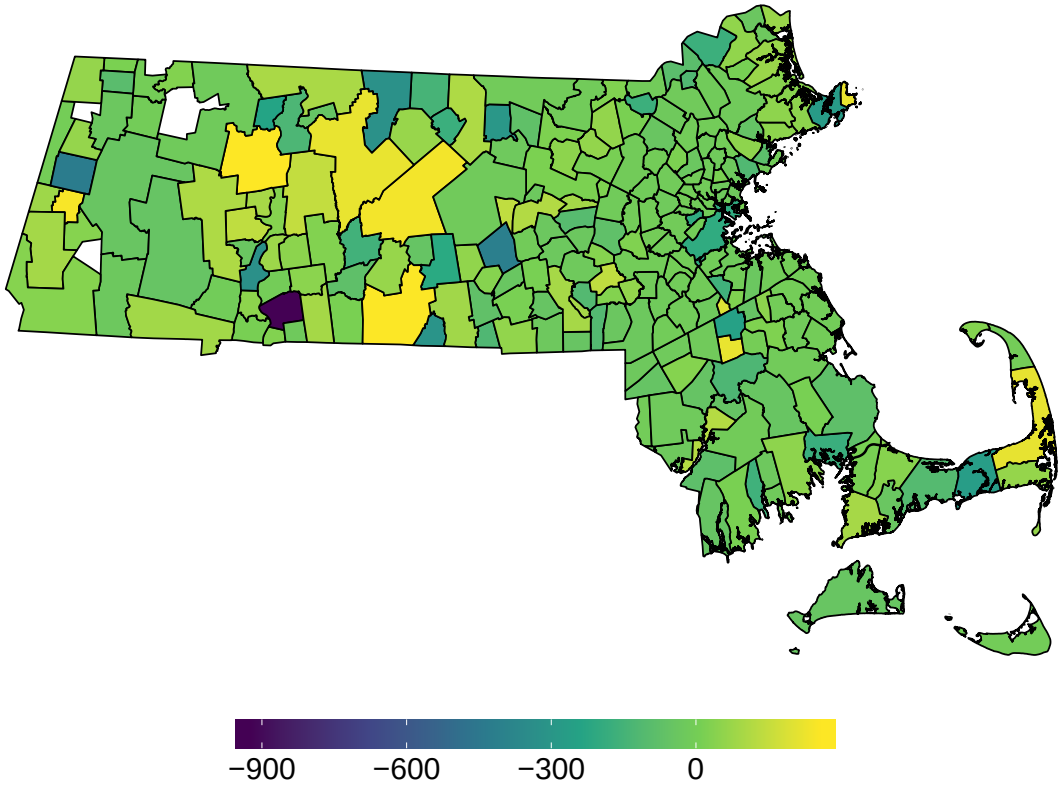

Figure 5: Net Student Gain/Loss to Inter-District Choice in 2016

Note: This figure displays the total change in student enrollment from inter-district choice for each Massachusetts school district for academic year 2016–2017. Each polygon in the figure represents the boundaries of a 12th grade school district. Note that some districts regionalize at the high school level but not at the primary or middle school level. Since students admitted in primary grades via inter-district choice are eligible to follow the feeder patterns of the receiving district, I aggregate enrollment for earlier grades to the appropriate 12th grade boundary. Students are considered a choice "gain" for a district if they are listed in the Student Information Management System (SIMS) data as enrolled via the inter-district choice program. A district "loses" a student if that student is listed in SIMS with a town of residence that feeds into the district, but the student is enrolled in a different district as a result of inter-district choice.

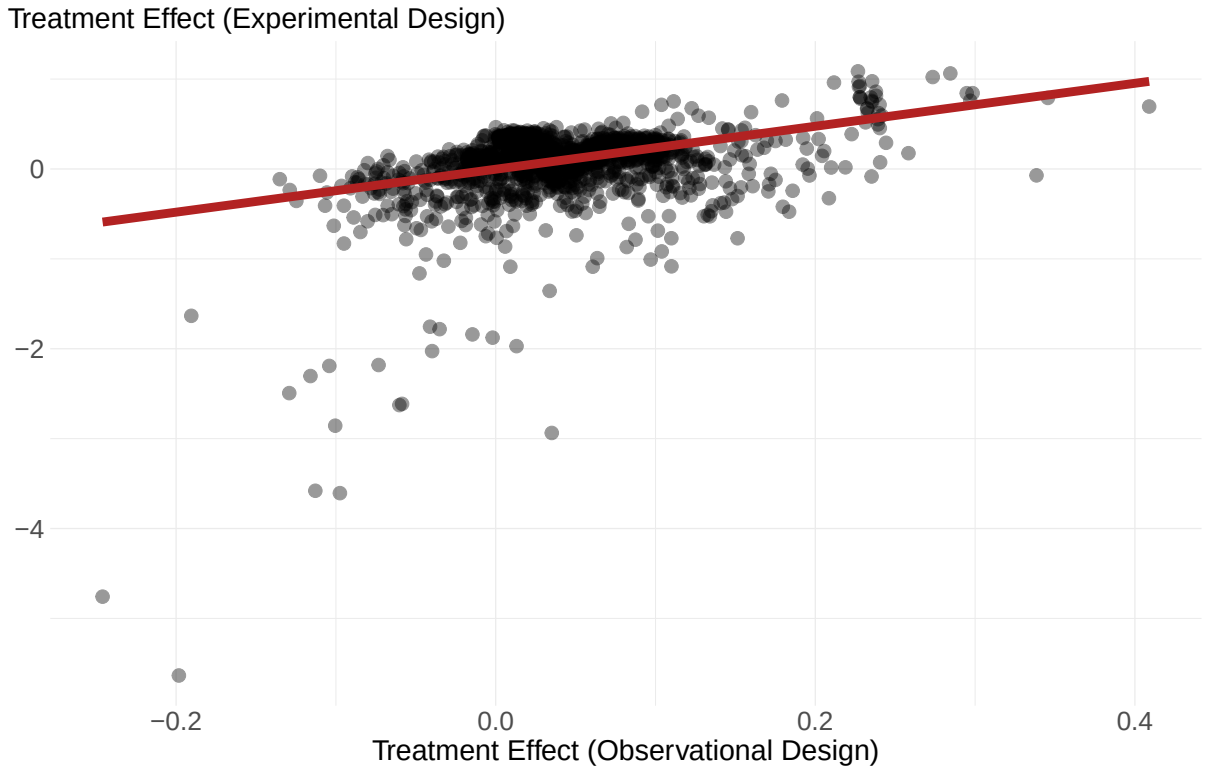

Figure 6: Correlation across Experimental and Non-Experimental Models

Note: This figure 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. 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 $(\alpha_0^e, \alpha^e, \alpha_0^n, \alpha^n)$ come from estimating the models described by equations (15), (16), (17), (18), and (19), and where X_{it} comes from an observation in the lottery sample. The correlation between the experimental and non-experimental estimates is 0.35.

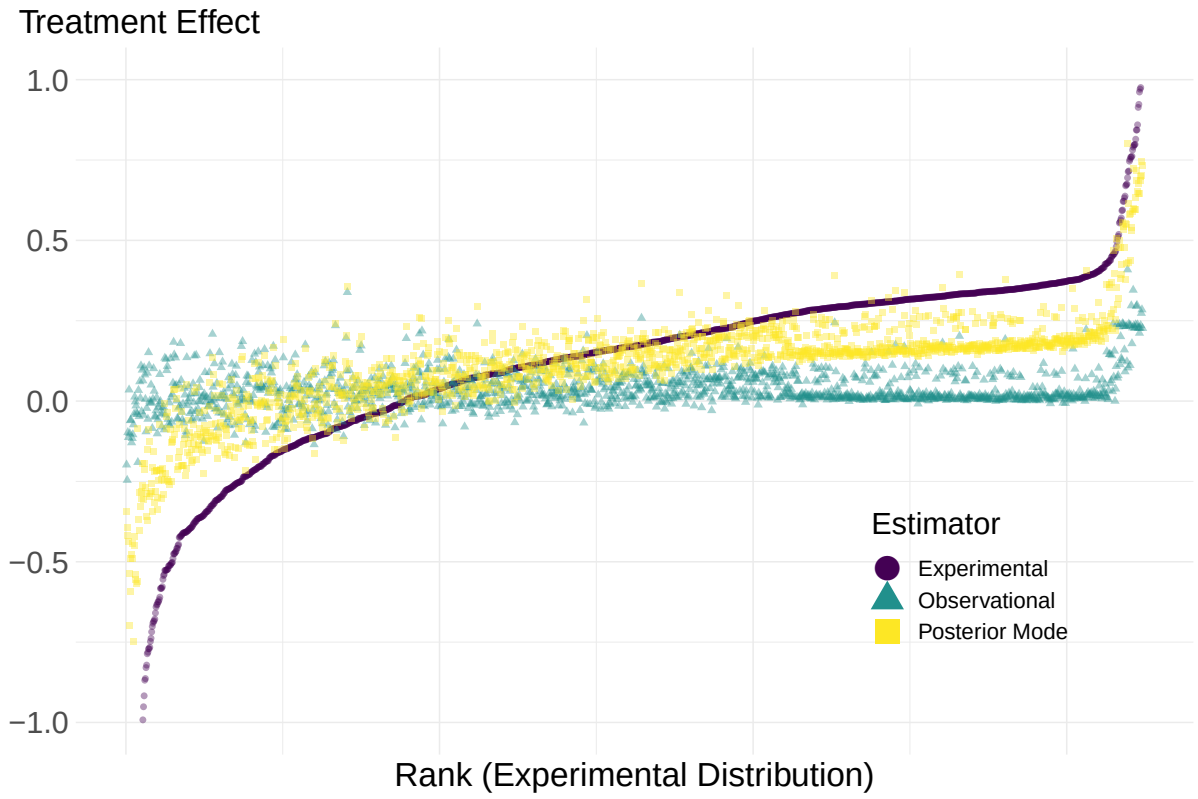

Figure 7: Visualizing Cross-Design Mixing

Note: This figure visualizes how the empirical Bayes estimator mixes the experimental and non-experimental information in practice. The x-axis plots the rank of a type k student 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).

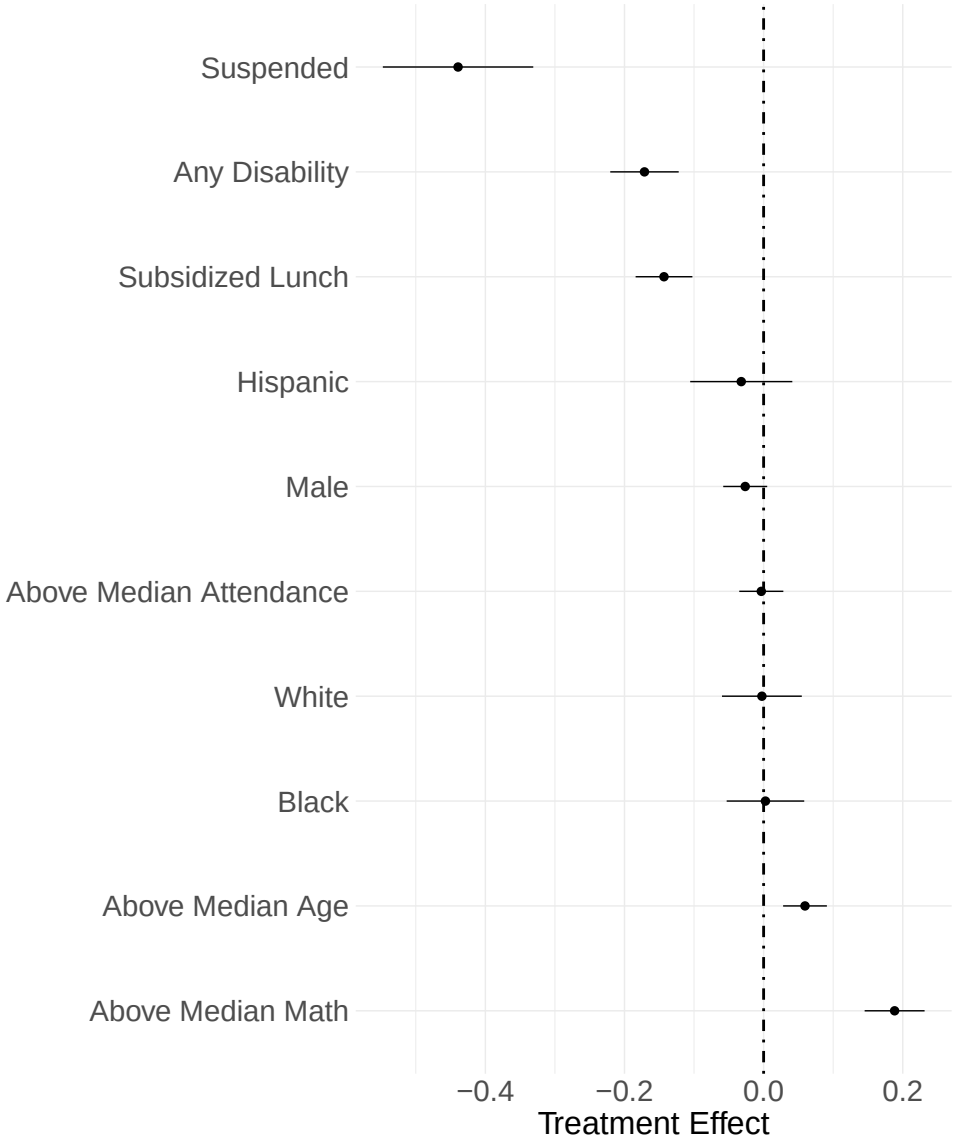

Figure 8: Predictors of Heterogeneous Treatment Effects

Note: This figure plots the predictors of treatment effect heterogeneity for the inter-district school choice program. Points come from regressions of the form $\hat{\beta}_k = \alpha + \omega d_k + u_k$, where $\hat{\beta}_k$ is the posterior mode of the heterogeneous effect for a type k student, d_k is an indicator for whether the student belongs to the subgroup, and ω is the average difference between the students who are and are not in the subgroup. Lines represent two standard error intervals around ω.

Table 1: Student Selection into Inter-District Choice

	All Students	Choice Students	Sending Districts	Lottery Sample
Math	0.01σ	-0.02σ	-0.25σ	0.11σ
ELA	0.01σ	0.03σ	-0.25σ	0.14σ
White	83%	93%	69%	90%
Black	11%	6%	29%	9%
Hispanic	15%	8%	27%	6%
Male	51%	48%	51%	47%
Subsidized Lunch	33%	28%	56%	21%
Limited English	6%	1%	11%	0%
Disability	14%	13%	15%	11%
Days Attended	163.41	162.13	159.08	168.05
Observations	4,890,552	66,817	305,673	968
Observations (Students)	1,528,038	33,251	155,473	968

Note: The sample used for constructing this table includes all students in Massachusetts attending traditional public schools. The column labeled "All Students" provides averages of observable characteristics 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 statewide sample to observations where a student is currently participating in inter-district choice. The column labeled "Sending Districts" restricts the statewide sample to districts that lose a student to inter-district choice via a lottery I observe in my data. The column labeled "Lottery Sample" restricts the statewide sample to students found in my lottery data as observed in the year when they applied.

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)	(0.04)
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

Note: The coefficients in this table come from models of the form $d_{jt} = \beta x_{jt} + u_{jt}$, where d_{jt} is an indicator for whether district j participated in inter-district choice in year t. x_{jt} is a vector of covariates that includes measures of student test scores, demographic characteristics, disciplinary records, and attendance, all averaged at the district-year level; average teacher experience and demographics; per-pupil measures of expenditure across ten categories; an indicator for urbanicity; and the total number of students, METCO students, schools, and teachers present in the district. For the complete list of coefficients from all predictors, see Appendix B.4.

Table 3: The Impact of Inter-District Choice on Test Scores

	Math				English Language Arts			
	OLS	RF	FS	2SLS	OLS	RF	FS	2SLS
Choice	0.01			0.19	−0.05			0.01
	(0.04)			(0.08)	(0.03)			(0.08)
Initial Offer		0.10	0.52			0.01	0.51	
		(0.04)	(0.03)			(0.04)	(0.03)	
Waitlist Offer		−0.12	0.96			−0.15	0.96	
		(0.15)	(0.06)			(0.20)	(0.06)	
Lottery Number		−0.003	0.01			0.005	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			117.2	117.2			114.9	114.9
Observations	1702	1702	1702	1702	1702	1702	1702	1702
Observations (students)	959	959	959	959	961	961	961	961
Adjusted R^2	0.66	0.66	0.32	0.65	0.56	0.56	0.32	0.56

Note: The table shows results from the two-stage least squares model outlined in equations (1) and (2). All of the table's estimates are from specifications that use my preferred set of controls. These include a lottery fixed effect, a baseline average of test scores observed prior to randomization, academic year and grade fixed effects, indicators for PARCC testing, indicators for whether or not a student was matched to the state data via an exact or fuzzy process, and indicators for waitlist requests or admissions rounds. The sample used for estimation includes all students from the lottery data who I was able to match to the state data who were involved in competitive lotteries, did not receive a sibling preference, did not apply late, and for whom I observe at least one test score prior to randomization. Note that the results are identical if I omit lotteries where waitlist requests or admissions rounds were used, and they are robust if I restrict to the set of students exposed only to the initial offer instrument. The results are also robust to dropping all controls except the lottery fixed effect and baseline test score average, and they are robust to models that include student fixed effects and year-by-lottery fixed effects so that the identification comes entirely from comparing test score trends within a lottery across winners and losers. For these robustness checks and others, see Appendix B.

Table 4: Impact of Inter-District Choice on Coursework

	Class Type Indicator			
	AP	Remedial	General	Advanced
Choice	0.14	−0.07	0.003	0.19
	(0.06)	(0.03)	(0.01)	(0.05)
Mean Dependent Variable	0.19	0.09	0.99	0.27
Observations	805	2,413	2,413	2,413
Observations (students)	467	911	911	911
Adjusted R^2	0.27	0.09	0.05	0.36

Note: The results in this table come from the two-stage least squares model outlined in equations (1) and (2). All of the table's estimates are from specifications that use my preferred set of controls. These include a lottery fixed effect, a baseline average of test scores observed prior to randomization, academic year and grade fixed effects, indicators for PARCC testing, indicators for whether or not a student was matched to the state data via an exact or fuzzy process, and indicators for waitlist requests or admissions rounds. The sample used for estimation includes all students from the lottery data that I was able to match to the state data who were involved in competitive lotteries, did not receive a sibling preference, did not apply late, and for whom I observe at least one test score prior to randomization. I also drop a small number of students who appear in my data only prior to the period when the department of education in Massachusetts kept records on student coursework. The column labeled "AP" further restricts the sample to observations that occur in grades 11 and 12, since Advanced Placement courses are typically unavailable to students in earlier grades.

Table 5: Suggestive Evidence on Medium-Run Outcomes

	Post-Secondary Outcome					
	Graduate Highschool		Attend 2 Year		Attend 4 Year	
Initial Offer	0.01		−0.05		0.05	
	(0.03)		(0.05)		(0.04)	
Choice		0.03		−0.09		0.07
		(0.07)		(0.10)		(0.09)
Mean Dependent Variable	0.88	0.88	0.38	0.38	0.61	0.61
Estimation Method	OLS	2SLS	OLS	2SLS	OLS	2SLS
F-Stat Excluded Instruments		66.5		66.5		66.5
Observations (Students)	537	537	537	537	537	537
Adjusted R^2	0.04	0.04	0.05	0.05	0.21	0.21

Note: This table's results come from the two-stage least squares model outlined in equations (1) and (2). All of the estimates in this table include as controls lottery fixed effects, a baseline average of test scores as observed prior to randomization, and indicators for whether or not a student was matched to the state data via an exact or fuzzy process. The sample used for estimation includes all students from the lottery data that I was able to match to the state data who were involved in competitive lotteries, did not receive a sibling preference, did not apply late, had a test score on file prior to the randomization, and who had an on-time graduation date prior to spring of 2016.

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	75.32	
Observations	1,705	6,549,949
Observations (Students)	966	1,784,773
Adjusted R^2	0.68	0.44

Note: This table compares results from the experimental and non-experimental pooled models. The column labeled "2SLS" provides results from equations (15), (16), and (17), with the further restrictions that $\alpha^e = \pi = 0$ so that there is no heterogeneity (i.e., $\beta_k = \beta$ is a constant). The estimation sample for the 2SLS column is identical to that used for the program evaluation in Section 5. The column labeled OLS provides results from equations (18) and (19) with a similar restriction on the heterogeneous effects ($\alpha^n = 0$). It is estimated using the sample of students in Massachusetts who do not appear in the lottery sample.

Table 7: Testing for Selection on Gains

	Take-up Indicator					
	Continue		Participate		Apply	
Heterogeneous Effect	0.10		0.15		0.15	
	(0.07)		(0.10)		(0.06)	
Heterogeneous Effect > 0		0.08		0.05		0.10
		(0.03)		(0.04)		(0.02)
Subsample	Ever-Enrolled	Ever-Enrolled	Applicants	Applicants	Elligible	Elligible
Observations	860	860	1,621	1,621	2,520	2,520
Observations	395	395	894	894	2,198	2,198
Dependent Variable Mean	0.85	0.85	0.46	0.46	0.32	0.32
Adjusted R^2	0.38	0.38	0.32	0.32	0.23	0.24

Note: This table presents results from models (22), (23), and (24). The "Ever-Enrolled" sample includes all students in my lottery data who ever accepted a lottery offer. The "Applicants" sample includes all students who appear in initial offer lotteries. The "Eligible" sample includes the "Applicant" sample plus a random 1% sample of students who appear in the same district-grades as the applicants during their application year.

Appendix to the Consequences of Sorting for Understanding School Quality

Intended for Online Publication Only

A Data Appendix

A.1 Primary Data Sources

Lottery Data

The lottery data I collected came in many forms. Most districts were able to provide me with a spreadsheet containing the relevant information. One district only maintained paper records and recorded the lottery results on the physical record. I scanned these paper records and compiled the results by hand. Another district only maintained records in the form of digital copies of letters, which were mailed to student families and contained the outcome of the lottery. I also compiled these by hand.

As discussed in the body of the paper, a number of districts that shared data with me did not document their lottery process in sufficient detail to warrant inclusion in my sample. Typically, this was due to the fact that every student in the lottery eventually received an offer of admission (after earlier admitted students declined), and the secretary wrote over the initial lottery results when indicating who received these later offers. There was also one district that used a complex scheme of highlighting, strikeouts, bold, italics, and shading cells to encode information related to their lottery-based admissions process. Reverse engineering the outcomes of this lottery ultimately proved impossible.

From the lottery records, I extracted student names, application grade and year, lottery preferences (e.g., sibling or late application), and lottery results. As discussed in the main text, this resulted in four different lottery instrument types. Nearly every district used what I term an "initial offer" instrument whereby some students were randomly selected and initial offers of admission were made to the family either via email or phone. I coded this as a binary indicator. Some districts then randomized students to positions on a waitlist, and I coded the numerical value of these positions as a "waitlist number" instrument. There is one district that, for one lottery in my data, did not use waitlist numbers and instead randomly chose students from the waitlist pool as spaces became available. I coded this as a binary "waitlist offer" instrument. Some of the first-stage and reduced-form results using this instrument

look odd. However, omitting the instrument due to odd-looking results could, in theory, create bias as a result of specification search (Gelman and Loken, 2013), and thus I included the instrument in my main specifications. That said, due to the small sample size involved, and the fact that all important results in the paper go through when I drop this instrument, I am not concerned by the odd-looking results. Finally, there was one district that assigned students random lottery numbers but did not record which students received initial offers. I coded these random number as a "lottery number" instrument.

When available, I also extracted date of birth and town of residence from the lottery data, since these were useful for matching. Town of residence was frequently misspelled, and I corrected these by hand as necessary. Observations in the lottery data frequently contained a census-designated place rather than a town of residence that would be recognizable in the state data. Where this occurred, I used zip code and publicly available information online to determine the town in which the census-designated place was located, and I replaced the census-designated place with the appropriate town.

As discussed in the main paper, a number of the lotteries exhibited idiosyncrasies. For example, there was one district that provided me a spreadsheet of lottery results that had a column labeled "admission rounds" and the numbers 1–3 entered into the corresponding cells below. The secretary I spoke with was unable to recall what this information was in reference to. For this reason, I coded indicators for each admission round and included them in the relevant specifications. There was also one district that, after the initial lottery results, asked students whether they wanted to be included on the waitlist before randomly assigning them waitlist numbers. I created indicators for this and included them in my main specifications. Omission of these idiosyncratic lotteries does not alter the results (see Appendix B for details).

Massachusetts Comprehensive Assessment System Data

The state of Massachusetts provided MCAS data to me for the spring test administration spanning the years 2002 through 2017. In all cases, I dropped students taking alternative assessments. For years 2002–2014, I used raw MCAS scores and standardized them within year and grade to have a mean of zero and a standard deviation of one. In 2015 and 2016, Massachusetts piloted a next-generation assessment based on the Partnership for Assessment of Readiness for College Careers (PARCC). For those years, some students in the state took the PARCC, some took the MCAS, and others took both. In 2017, the state transitioned entirely to the PARCC. Thus for the years 2015–2017, I took the raw MCAS scores wherever available. When unavailable, I used raw PARCC scores. For the 2015 test administration, the state was

unable to locate raw PARCC scores, so I used PARCC theta scores in their place.[35] Thus for each of these years, I standardized the test scores at the year-grade-test type level (raw MCAS, raw PARCC, PARCC theta) to have a mean of zero and a standard deviation of one. In addition, I included indicators for test types (raw MCAS, raw PARCC, PARCC theta) in the relevant specifications. For specifications that included lagged or baseline test scores, I also included lagged or baseline test type indicators.[36]

Student Information Management System Data

The state of Massachusetts provided information on the universe of public school students in Massachusetts spanning academic years 2001–2002 through 2016–2017. These data contained information on student names, gender, birth dates, assigned schools, grade, attendance, race, ethnicity, disabilities, and free and reduced-price lunch status. They also contained a variable describing which students were enrolled in a district via the inter-district school choice program. For years 2006–2007 through 2016–2017, students could identify as multiple races (e.g., black and white). Prior to 2006, students were restricted to choosing only one. For this reason, I coded the race/ethnicity variables as a series of indicators that take a value of one if a student in the given year identified as belonging to the given category. Since all specifications in the main text include year fixed effects, doing this will account for the fact that the meaning of the variables changes over time.

Student Course Schedule Data

The state of Massachusetts provided me with student course scheduling data spanning academic years 2010–2011 through 2016–2017. From these data, I determined which courses were AP classes by searching for "AP" within the course name. The remaining course designations I use in the paper (advanced, general, and remedial) come from a variable already contained in the data. I then counted the number of each course type each student in the state is enrolled in during the given academic year.

National Student Clearinghouse Data

The state of Massachusetts contracts with the NSC to produce data on postsecondary outcomes for students. The NSC data itself contain information on college enrollment and com-

[35] These are transformed versions of the raw scores meant to adjust for question difficulty using techniques from item response theory.

[36] The one exception here is the regressions for postsecondary outcomes. I did not include these indicators here since all test scores observed at baseline for this sample are raw MCAS and are hence unnecessary.

pletion from over 3,600 universities enrolling 98% of all college students in the United States (National Student Clearinghouse, 2020). The data I received were split into two folders based on whether the student had or had not graduated from high school at the time NSC conducted the search. I used this division to determine whether or not a student graduated from high school. I coded variables for two- and four-year college attendance based on enrollment dates contained in the NSC data. Because only a small minority of my sample would have an on-time four-year college graduation date of 2016 or earlier, I did not examine outcomes related to college completion.

EPIMS Data

The EPIMS data come from information transmitted from school districts to the state of Massachusetts. The data were provided to me in ten separate SPSS .sav files, each one corresponding to an academic year. The unit of observation in these data is a teacher-school-course-section-term. After standardizing variable names, I merged the files into a single data set at the teacher-year-school-course-term level. From there, I resolved inconsistencies in the data. For example, the gender variable sometimes coded males as "M" and others as "m"; I ensured such coding was common across all years.

Auxiliary Sources

Finally, I also made use of a small number of auxiliary data sources:

- Choice Votes over Time: The DESE provided spreadsheets to me containing the results of votes for district choice participation over time. These files contained a number of naming inconsistencies across years, which I fixed.
- Mapping between Town of Residence and Assigned District: The DESE also provided me with a spreadsheet describing the mapping between town of residence and assigned school district at each grade level for the year 2017. This spreadsheet also contained a tab describing how this mapping had changed over time. For example, a number of smaller districts regionalized over my sample period. I used this second tab to construct a panel data frame that describes the mapping between town of residence and assigned district over my entire sample period for the purposes of determining the “home district” of students who participate in inter-district choice.
- District Spending: The DESE also provided a spreadsheet containing detailed data on

expenditures across various categories by school district from academic years 2008–2017. These data required very little cleaning.

A.2 Matching Details

To match the state data to the lottery data, I first looked for students with exact first and last name matches in the appropriate grade and year. When available, I would break ties with date of birth, followed by town of residence as necessary. If town of residence was unavailable and I was unable to produce a unique match using birth date, I would look for unique first name/last name matches within the empirical distribution of towns such that I either observed a student apply from that town in the lottery data or I observed a student enroll from that town in the state data. When I was unable to break a tie in this manner, I would consider the student unmatched and drop them from the lottery sample.

When I was unable to find an exact first name/last name match anywhere in the state, I repeated the algorithm from the previous paragraph using fuzzy first name/last name matching. I would calculate the Levenshtein distance[37] between the first and last name of the observation in my lottery data and the rest of the students in the state enrolled in the appropriate grades/years/towns, and I would then restrict the state data to observations falling within a Levenshtein distance of two. When this procedure did not produce a unique match, I would further break ties using birth date (if available). At this point, if the student remained without a unique match, I would consider the student unmatched and drop them from the lottery sample.

A.3 Additional Lottery Descriptive Statistics

In this section, I present some additional descriptive information related to the lotteries. Figure A.1 is a histogram of lotteries by the number of students involved, and Figure A.2 shows the distribution of lotteries over time.

[Figure A.1 about here.]

[Figure A.2 about here.]

[37] Given two strings "a" and "b," the Levenshtein distance calculates the minimum number of insertions, deletions, and substitutions necessary to turn b into a.

A.4 Differential Attrition

Differential attrition is problematic if winning the lottery affects the probability of a student subsequently appearing in the state data. For example, if high-ability lottery losers attrit by leaving for private schools, the postlottery winners would be higher ability, on average, even in the absence of a causal effect of inter-district school choice, and would hence lead to a biased result.

To check for this possibility, I start with the raw lottery data and restrict it to the sample of students I am ever able to match in the state data. I drop students who received sibling preference or applied late. I then regress an indicator for whether I observe a student in the postlottery period on the lottery instruments and a vector of lottery fixed effects.[38] Table A.1 presents the results. Column (1) presents results for the entire lottery sample, and column (2) controls for demographic characteristics. Column (3) further restricts the sample to students I observe in the data at baseline and includes additional baseline controls.[39] Column (4) further restricts the sample to the set of students for whom I observe a baseline test score and includes these baseline scores as controls. In all cases, it would appear that winning the lottery is unrelated to the probability a student subsequently appears in the state data. Thus it is unlikely that the results in the main body of the paper are affected by differential attrition.

[Table A.1 about here.]

A.5 Falsification Tests

In this section, I present the results of a standard IV falsification test for math and ELA test scores. Intuitively, if the IV exclusion restriction holds, then for subsamples or time periods where the first stage is known to be zero, we should not find a reduced-form relation between the instrument and the outcome. In the school choice context, this means there should be no relation between winning a lottery offer and prelottery test scores.

Consistent with the covariate imbalance discussed in the main body of the paper, I find that the lotteries do not pass this test. As I argue in the main body, it is likely that this is due either to sampling variation or poor record keeping on the part of some districts with respect to things like sibling preference or late application. Given the strong anti-discrimination language of the state legislation, the timing of how lotteries were conducted, discussions I

[38] As mentioned in the main body of the paper, there was one lottery that had "admission rounds" indicated in their spreadsheets without further explanation. I also include indicators where this happens. There was also one district that asked lottery losers whether they wanted to be included on the waitlist before randomly assigning waitlist numbers. I include indicators where this happens as well.

[39] Indicators for any disability, subsidized lunch status, and English language learners.

had with district administrators related to their lottery process, and the incentives around sharing data with me, I believe it is unlikely that the failure here is due to cheating. In any event, whatever the cause of the imbalance, I need to correct for it to achieve consistent estimates of the parameter of interest.

[Table A.2 about here.]

[Table A.3 about here.]

Importantly, I find that conditional on an earlier test score, there is no relation between the baseline scores and the lottery results. To show this, I start by restricting the sample to the set of students I can match to the state data who were involved in competitive lotteries, who were not indicated as receiving a sibling preference in the lottery nor as having applied late, and who also have at least one prelottery test score. I then regress the most recent prelottery test score on the vector of lottery instruments and a set of lottery fixed effects.[40] Tables A.2 and A.3 present the results. Column (1) presents the otherwise uncontrolled comparison, and column (2) shows results including other baseline demographic controls. Column (3) restricts the sample to the set of students for whom I observe at least two test scores prior to the lottery and presents results for the otherwise uncontrolled comparison. Column (4) adds baseline controls to the two-test sample, and column (5) uses the two-test sample but exchanges the baseline demographic controls for a second prelottery test score. Column (6) presents results for the two-test sample controlling both for demographic characteristics and the second prelottery test score.

As we can see from Tables A.2 and A.3, conditioning on an earlier test score substantially reduces the unexplained variation and eliminates the relation between the lottery vector and the baseline test scores. For this reason, all lottery specifications in the main text are restricted to the sample where I observe at least one test score prior to randomization and includes an average of prelottery test scores as a control.

[40] As mentioned in the main body of the paper, there was one lottery that had "admission rounds" indicated in their spreadsheets without further explanation; I also include indicators where this happens. There was also one district that asked lottery losers whether they wanted to be included on the waitlist before randomly assigning waitlist numbers, and I include indicators where this happens as well.

B Supplemental Results and Robustness Checks

B.1 Specifications with Additional Controls

In this section, I present evidence that the results are robust to including additional sets of controls. Table A.4 presents the results. Columns (1) and (4) replicate the results from the main specification in the text, and columns (2) and (5) add controls for demographic characteristics and other baseline observables. These controls include indicators for race, ethnicity, and gender; age (measured in days). They also include indicators for whether the student received a subsidized lunch at baseline, had any disability at baseline, number of days suspended at baseline, number of unexcused absences at baseline, and whether the student was labeled as an English language learner at baseline.

Columns (3) and (6) expand the sample to include all student years observed prior to the lottery and replaces all controls with a set of student fixed effects and year-by-lottery fixed effects.[41] Hence columns (3) and (6) represent an IV difference-and-difference design that generates reduced-form and first-stage estimates by comparing the trends in test scores and choice status across lottery winners and losers relative to the lottery's date. In all cases, the results continue to hold.

[Table A.4 about here.]

B.2 Results Using Only Initial Offer Instrument

In this section, I present evidence that the results are robust to omitting the less frequently observed instruments. I do this by restricting the sample to the set of lotteries that used initial offer instruments and otherwise replicating the baseline specification. Table A.5 shows the results. Columns (1) and (3) replicate the main specification from Table 3 using all instruments. Columns (2) and (4) drop students who were not involved in initial offer lotteries and only uses the initial offer instrument in the first stage.

[Table A.5 about here.]

[41] To be precise, the 2SLS specification used for columns (3) and (6) is $y_{it} = \alpha_i + \beta d_{it} + \delta_{\ell t} + \epsilon_{it}$ with the first stage given by $d_{it} = \alpha_i' + \Pi Z_i + \delta_{\ell t}' + \eta_{it}$. In these equations, α_i and α_i' are fixed effects for student i, while $\delta_{\ell t}$ and $\delta_{\ell t}'$ are fixed effects defined by the interaction of the lottery fixed effects and the calendar year fixed effects. The other variables/parameters in these two equations are as defined in the main body of the text.

B.3 Omitting Idiosyncratic Lotteries

In this section, I present evidence that the results are not driven by lotteries with idiosyncratic randomization procedures. Recall that there was one lottery in my data where the district had labeled "admission rounds" in the lottery records without explanation as to how these were used. For the specifications in the main text, I control for this using admission round indicators. In addition, there was one district that, after initially randomly selecting students to receive initial offers, would ask students whether or not they wanted to be assigned a random waitlist number. For the specifications in the main text, I control for this with an indicator for the students who are affected. Table A.6 presents robustness checks where I instead omit these lotteries. Columns (1) and (4) replicate the specifications from Table 3. Columns (2) and (5) drop the lottery with admission rounds, while columns (3) and (6) drop the waitlist request lotteries.

[Table A.6 about here.]

B.4 District Take-Up

Table A.7 presents a series of regressions where I predict a year-by-district indicator for participation in inter-district choice since 2009 with district-level observable characteristics. Column (1) estimates the model with OLS and includes nearly all observables at my disposable: demographic composition of students and teachers, average test scores, rates of suspensions and unexcused absences, an urban indicator, a METCO participation indicator, and per-pupil expenditures across 11 categories. Column (2) presents results from a post-Lasso regression where model selection was performed over the set of variables included in column (1). For visual clarity, the only variables displayed in the table are those that were Lasso selected. Columns (3) and (4) provide the results from regressions using the time-varying observables from columns (1) and (2) but including district and time fixed effects. In effect, columns (3) and (4) ask if trends in the predictors are related to changes in the status of choice.

While trends in the predictors appear to be unrelated to the decision to participate in choice,[42] in levels there are a number of economically meaningful covariates, suggesting that over the short term, participation in choice is driven largely by the geographic distribution of covariates. While the most important predictors appear to be average test scores, I will not speculate on what that implies for the decision to participate. Test scores are highly correlated with many other observables, making the relation difficult to interpret. On the

[42] This is at least true over a seven-year time span.

other hand, other economically meaningful covariates such as the student-teacher ratio and per-pupil expenditures appear to agree with informal conversations I have had with district administrators. The decision to participate in choice is often driven by a desire to supplement revenue subject to class-size constraints.

[Table A.7 about here.]

B.5 Coursework Results: Intensive Margin

Table A.8 provides results on coursework using intensive margin variation. I replicate the specifications used to analyze coursework in Table 4, but I use the number of classes of each type the student was enrolled in as an outcome variable. While noisy, the point estimates show that the number of general classes has also increased, suggesting that the coursework substitution works by pushing students from remedial into general classes and from general classes into advanced or AP classes. Note that there is no adding up constraint on these coefficients, since there are many classes in the data (e.g., gym and music) that do not receive any of the three labels, suggesting that some of the increases are also being driven by substitution away from non-academic classes.

[Table A.8 about here.]

C Statistical Appendix

C.1 Additional Details of Heterogeneous Effect Estimation Procedure

To generate the consensus posterior modes of the heterogeneous effects that incorporate both the experimental and non-experimental variation, I proceed in five steps:

1. Estimate the center of the parent distribution.
2. Estimate the heterogeneous effects.
3. Estimate the variance-covariance matrix of the parent distribution.
4. Estimate the asymptotic covariance matrix of the heterogeneous effects.
5. Calculate the posterior modes.

Estimating the Center of the Parent Distribution

To recover the center of the experimental distribution $\hat{\beta}_0$, I estimate the following model on the lottery sample using 2SLS:

$$y_{it} = \delta_0 + \delta_\ell + \beta_0 d_{it} + \gamma_w W_i + \gamma_x X_{it} + \epsilon^p_{it} \tag{25}$$

$$d_{it} = \delta'_0 + \delta'_\ell + \pi Z_{it} + \gamma'_w W_i + \gamma'_x X_{it} + \eta^p_{it}, \tag{26}$$

where y_{it} is the postlottery test score of student i at time t, δ_0 is an intercept, δ_ℓ is a lottery fixed effect, d_{it} is an indicator for whether student i appeared outside of their home district under the choice program at time t, W_i is the vector of baseline observables described in the main text, and X_{it} are the relevant margins of heterogeneity as described in the body of the main text.

To recover the center of the non-experimental distribution ($\hat{\beta}^n_0 = \widehat{\beta_0 + b_0}$), I estimate the following model on the non-experimental data using OLS:

$$y_{it} = \delta_{hgt} + \beta^n_0 d_{it} + \theta_x X_{it} + u^p_{it}, \tag{27}$$

where δ_{hgt} is a home district (h) by grade (g) by academic year (t) fixed effect.

Note that I have superscripted the residuals of the regressions in this section by the letter p, which stands for "pooled." This notation will become useful later on to distinguish the

residuals from these regression from the corresponding heterogeneous regressions when I estimate the parent variance-covariance matrix.

Estimating Heterogeneous Effects

To recover the heterogeneous effect estimates, I first estimate the following model on the lottery sample using 2SLS:

$$y_{it} = \delta_0 + \delta_\ell + \alpha_0^e d_{it} + \alpha^e X_{it} d_{it} + \gamma_w W_i + \gamma_x X_{it} + \epsilon_{it} \tag{28}$$

$$d_{it} = \delta_0' + \delta_\ell' + \pi_0 Z_{it} + \pi X_{it} Z_{it} + \gamma_w' W_i + \gamma_x' X_{it} + \eta_{it}. \tag{29}$$

I then estimate the following model on the non-experimental sample using OLS:

$$y_{it} = \delta_{hgt} + \alpha_0^n d_{it} + \alpha^n X_{it} d_{it} + \theta_x X_{it} + u_{it}. \tag{30}$$

$$\tag{31}$$

And from there, I recover the observable heterogeneity as

$$\hat{\beta}^e = \hat{\alpha}_0^e + \hat{\alpha}^e \mathbb{X} \tag{32}$$

$$\hat{\beta}^n = \hat{\alpha}_0^n + \hat{\alpha}^n \mathbb{X}, \tag{33}$$

where $\mathbb{X} = (X_{1t}, \ldots X_{Et})$ is the matrix of support points for covariates found in the lottery sample.

Estimating the Variance-Covariance Matrix of the Parent Distribution

To estimate the variance-covariance matrix of the parent distribution, I calculate $\hat{\Sigma} = cov(\tilde{\epsilon}_{it}^p - \tilde{\epsilon}_{it}, \tilde{u}_{it}^p - \tilde{u}_{it} | d_{it} = 1) \approx cov(\beta^e, \beta^n)$. I calculate the residuals $(\tilde{\epsilon}_{it}^p, \tilde{\epsilon}_{it}, \tilde{u}_{it}^p, \tilde{u}_{it})$ using the estimated coefficients from models (25), (27), (28), and (30) and the support points of the lottery data as follows:

- $\tilde{\epsilon}_{it}^p = y_{it} - (\hat{\delta}_0 + \hat{\delta}_\ell + \hat{\beta}_0 + \hat{\gamma}_w W_i + \hat{\gamma}_x X_{it})$
- $\tilde{\epsilon}_{it} = y_{it} - (\hat{\delta}_0 + \hat{\delta}_\ell + \hat{\alpha}_0^e + \hat{\alpha}^e X_{it} + \hat{\gamma}_w W_i + \hat{\gamma}_x X_{it})$
- $\tilde{u}_{it}^p = y_{it} - (\hat{\delta}_{hgt} + \hat{\beta}_0^n + \hat{\theta}_x X_{it})$

- $\tilde{u}_{it} = y_{it} - (\hat{\delta}_{hgt} + \hat{\alpha}_0^n + \hat{\alpha}^n X_{it} + \hat{\theta}_x X_{it})$

Estimating the Asymptotic Covariance Matrix of the Heterogeneous Effects

As discussed in the main body of the text, I assume the off-diagonal entries of the asymptotic covariance matrix Ω_k are zero. This assumption is reasonable because I have dropped the students appearing in the lottery sample from the nonlottery data when estimating the non-experimental model. To recover the diagonal entries, observe that

$$\hat{\Omega}_k = diag(var(\hat{\beta}_k^e), var(\hat{\beta}_k^n)) = diag(X_{it}' var(\hat{\alpha}^e) X_{it}, X_{it}' var(\hat{\alpha}^n) X_{it}). \tag{34}$$

Hence I recover $\hat{\Omega}_k$ by replacing $var(\hat{\alpha}^e)$ and $var(\hat{\alpha}^n)$ with standard sample analogues.

Calculating Posterior Modes

Let $\theta_k = \begin{bmatrix} \beta_k^e \\ \beta_k^n \end{bmatrix}$ and $\theta_0 = \begin{bmatrix} \beta_0 \\ \beta_0 + b_0 \end{bmatrix}$, and recall that the hierarchical model induces the following Bayesian structure:

$$p(\theta_k|\hat{\theta}_k) \propto p(\hat{\theta}_k|\theta_k) p(\theta_k). \tag{35}$$

Then provided the parent distribution is normal, the posterior distribution is $\mathbb{N}(\mu_k, \Gamma_k)$, where

$$\Gamma_k = (\Sigma^{-1} + \Omega_k^{-1})^{-1} \tag{36}$$

$$\mu_k = \Gamma_k \Omega_k^{-1} \hat{\theta}_k + \Gamma_k \Sigma^{-1} \theta_0. \tag{37}$$

And thus we can plug in in the empirical analogues $\hat{\Sigma}$, $\hat{\Omega}_k$, and $\hat{\theta}_0$ to recover an estimate of the consensus posterior mode $\hat{\mu}_k$.

Jack-Knife and Split Sample Procedures

Ultimately, I recover the posterior modes because I wish to correlate them with the application and take-up behavior of students. For this reason, I estimate the posterior modes using a jack-knife procedure to ensure there is no mechanical correlation between the treatment indicator and heterogeneous effects. The jack-knife algorithm for the experimental data proceeds as follows:

1. Drop all students associated with lottery ℓ from the sample.

2. Estimate the pooled and heterogeneous experimental and non-experimental models.

3. Estimate the joint covariance matrix of the parent distribution.

4. Use the estimated experimental and non-experimental models to predict the heterogeneous effects of the students associated with lottery ℓ along with the corresponding covariance matrices Ω_k.

5. Calculate the posterior modes for the students associated with lottery ℓ.

To estimate the posterior modes of the heterogeneous effects of students not contained in the lottery data but who were eligible to apply, I use a split sample procedure where I divide the pool of eligible applicants in half. I use the first half of the potential applicants to estimate the relevant models along with the joint covariance matrix. I then predict the heterogeneous effects for the second half of students out of sample using the models estimated on the first half and apply the shrinkage.

C.2 Properties

A General Expression for the Posterior Modes

For all empirical results in this paper, I have assumed the off-diagonal elements of the joint covariance matrix Ω_k are zero. This is justified because I do not include the students in the lottery data in the non-experimental models. Under this assumption, we can write the consensus posterior modes as

$$\beta_k^s = \beta_0 + \alpha_k(\hat{\beta}_k^e - \beta_0) + \delta_k(\hat{\beta}_k^n - \beta_0 - b_0) \tag{38}$$

$$\alpha_k = \frac{\phi_n^k - \rho^2}{\phi_n^k \phi_e^k - \rho^2} \tag{39}$$

$$\delta_k = \frac{\rho \frac{(\omega_e^k)^2}{\sigma_e \sigma_n}}{\phi_n^k \phi_e^k - \rho^2}, \tag{40}$$

where $\rho \equiv corr(\beta_k, \beta_k + b_k)$ is the correlation between the experimental and non-experimental estimands, $\phi_e^k \equiv \frac{\sigma_e^2 + (\omega_e^k)^2}{\sigma_e^2}$ is the inverse of the weight you would recover by applying a standard empirical Bayes idea to the experimental data alone, and $\phi_n^k \equiv \frac{\sigma_n^2 + (\omega_n^k)^2}{\sigma_n^2}$ is the inverse of the weight you would recover if applying a standard empirical Bayes idea to the non-experimental data alone.

Now I will provide general expressions for the weights when we relax the assumption on the off-diagonal elements of the variance-covariance matrix. First, I will clarify the notation. Let $\theta_k = \begin{bmatrix} \beta_k^e \\ \beta_k^n \end{bmatrix}$ and $\theta_0 = \begin{bmatrix} \beta_0 \\ \beta_0 + b_0 \end{bmatrix}$, $\Sigma = \begin{bmatrix} \sigma_e^2 & \sigma_{en}^2 \\ \sigma_{en}^2 & \sigma_n^2 \end{bmatrix}$, and $\Omega_k = \begin{bmatrix} (\omega_e^k)^2 & (\omega_{en}^k)^2 \\ (\omega_{en}^k)^2 & (\omega_n^k)^2 \end{bmatrix}$. Observe that we can write the posterior distribution of heterogeneous effect k as $\theta_k \sim \mathbb{N}(\mu_k, \Gamma_k)$, where

$$\Gamma_k = (\Sigma^{-1} + \Omega_k^{-1})^{-1}, \tag{41}$$

$$\mu_k = \Gamma_k \Sigma^{-1} \theta_0 + \Gamma_k \Omega_k^{-1} \hat{\theta}_k = W_0 \theta_0 + W_1 \hat{\theta}_k, \tag{42}$$

where W_0 and W_1 are weighting matrices. The expression for these matrices are

$$W_0 = \begin{bmatrix} \frac{(\omega_e^k)^2(\sigma_n^2+(\omega_n^k)^2)-(\omega_{en}^k)^2(\sigma_{en}^2+(\omega_{en}^k)^2)}{(\sigma_e^2+(\omega_n^k)^2)(\sigma_e^2+(\omega_e^k)^2)-(\sigma_{en}^2+(\omega_{en}^k)^2)^2} & \frac{\sigma_e^2(\omega_{en}^k)^2-\sigma_{en}^2(\omega_e^k)^2}{(\sigma_n^2+(\omega_n^k)^2)(\sigma_e^2+(\omega_e^k)^2)-(\sigma_{en}^2+(\omega_{en}^k)^2)^2} \\ \frac{\sigma_n^2(\omega_{en}^k)^2-\sigma_{en}^2(\omega_n^k)^2}{(\sigma_n^2+(\omega_n^k)^2)(\sigma_e^2+(\omega_e^k)^2)-(\sigma_{en}^2+(\omega_{en}^k)^2)^2} & \frac{(\omega_n^k)^2(\sigma_e^2+(\omega_e^k)^2)-(\omega_{en}^k)^2(\sigma_{en}^2+(\omega_{en}^k)^2)}{(\sigma_n^2+(\omega_n^k)^2)(\sigma_e^2+(\omega_e^k)^2)-(\sigma_{en}^2+(\omega_{en}^k)^2)^2} \end{bmatrix} \tag{43}$$

$$W_1 = \begin{bmatrix} \frac{\sigma_e^2(\sigma_n^2+(\omega_n^k)^2)-\sigma_{en}^2(\sigma_{en}^2+(\omega_{en}^k)^2)}{(\sigma_n^2+(\omega_n^k)^2)(\sigma_e^2+(\omega_e^k)^2)-(\sigma_{en}^2+(\omega_{en}^k)^2)^2} & \frac{\sigma_{en}^2(\omega_e^k)^2-\sigma_e^2(\omega_{en}^k)^2}{(\sigma_n^2+(\omega_n^k)^2)(\sigma_e^2+(\omega_e^k)^2)-(\sigma_{en}^2+(\omega_{en}^k)^2)^2} \\ \frac{\sigma_{en}^2(\omega_n^k)^2-\sigma_n^2(\omega_{en}^k)^2}{(\sigma_e^2+(\omega_n^k)^2)(\sigma_e^2+(\omega_e^k)^2)-(\sigma_{en}^2+(\omega_{en}^k)^2)^2} & \frac{\sigma_n^2(\sigma_e^2+(\omega_e^k)^2)-\sigma_{en}^2(\sigma_{en}^2+(\omega_{en}^k)^2)}{(\sigma_n^2+(\omega_n^k)^2)(\sigma_e^2+(\omega_e^k)^2)-(\sigma_{en}^2+(\omega_{en}^k)^2)^2} \end{bmatrix}. \tag{44}$$

The simplified expression can be found by setting $\omega_{en}^k = 0$ and rearranging the expression for the weights.

Consistency

To see that the estimator is consistent under the same conditions as IV, observe that as the IV sample size (E) becomes large, then given standard regularity conditions and a fixed non-experimental sample size (N), we have

$$
\begin{aligned}
plim_{E\to\infty}\hat{\beta}_0 &= \beta_0 \\
plim_{E\to\infty}\hat{\beta}_k^e &= \beta_k \\
plim_{E\to\infty}\hat{\phi}_e^k &= \frac{plim_{E\to\infty}\hat{\sigma}_e^2 + plim_{E\to\infty}(\hat{\omega}_e^k)^2}{plim_{E\to\infty}\hat{\sigma}_e^2} = \frac{plim_{E\to\infty}\hat{\sigma}_e^2 + 0}{plim_{E\to\infty}\hat{\sigma}_e^2} = 1 \\
plim_{E\to\infty}\hat{\alpha}_k &= \frac{plim_{E\to\infty}\hat{\phi}_n^k - plim_{E\to\infty}\hat{\rho}^2}{(plim_{E\to\infty}\hat{\phi}_n^k)(plim_{E\to\infty}\hat{\phi}_e^k) - plim_{E\to\infty}\hat{\rho}^2} = \frac{plim_{E\to\infty}\hat{\phi}_n^k - plim_{E\to\infty}\hat{\rho}^2}{(plim_{E\to\infty}\hat{\phi}_n^k)(1) - plim_{E\to\infty}\hat{\rho}^2} = 1 \\
plim_{E\to\infty}\hat{\delta}_k &= \frac{\frac{(plim_{E\to\infty}\hat{\rho})(plim_{E\to\infty}(\hat{\omega}_e^k)^2)}{plim_{E\to\infty}(\hat{\sigma}_e\hat{\sigma}_n)}}{plim_{E\to\infty}(\hat{\phi}_n^k\hat{\phi}_e^k - \hat{\rho}^2)} = \frac{\frac{(plim_{E\to\infty}\hat{\rho})(0)}{plim_{E\to\infty}(\hat{\sigma}_e\hat{\sigma}_n)}}{plim_{E\to\infty}(\hat{\phi}_n^k\hat{\phi}_e^k - \hat{\rho}^2)} = 0.
\end{aligned}
\tag{45}
$$

And we can see that

$$
plim_{E\to\infty}\hat{\beta}_k^s = \beta_0 + (1)(\beta_k^e - \beta_0) + (0)(\hat{\beta}_k^n - \widehat{\beta_0 + b_0}) = \beta_k^e. \tag{46}
$$

C.3 Simulations

In this section, I show the results of two sets of simulation. The first simulation shows how the estimator's performance varies with the signal-to-noise ratio under an idealized DGP and when the parent distribution for the treatment effect heterogeneity is normal. I find that the empirical Bayes procedure using all of the data strictly dominates other available estimators for the treatment effect heterogeneity.

The second simulation shows how the estimator's performance performs using a DGP calibrated to match the distribution of observables in the real data and estimated using the linear approximation for treatment heterogeneity applied in the main text. This is important because the model I take to the data in the main text assumes the treatment effect heterogeneity is a linear combination of binary and other nonnormally distributed variables and hence implies the potential for deviations from normality in the parent distribution. In the simulation, I find that despite these deviations from normality, the empirical Bayes procedure using all of the data weakly dominates among consistent estimators under a wide range of values for the amount of selection present in the non-experimental data.

Performance under Normality

To assess the theoretical performance of the empirical Bayes procedure, I generate fake data under idealized conditions as follows:

$$y_i = \alpha + \beta_k d_i + x_i + \epsilon_i \tag{47}$$

$$\beta_k \sim \mathbb{N}(\beta_0, \sigma_\beta) \tag{48}$$

$$x_i \sim \mathbb{N}(0, \sigma_x) \tag{49}$$

$$\epsilon_i \sim \mathbb{N}(0, \sigma_\epsilon). \tag{50}$$

Here y_i is the outcome of interest, d_i is a treatment indicator, β_k is the heterogeneous treatment effect to an individual of observable type $k \in \{1, 2, \dots K\}$, x_i is an unobserved potential confounder, and ϵ_i is a residual. The treatment vector d_i is assigned differently depending on whether the observation is in the experimental or observational sample. Note that I generate the data such that there are J_e observations per type k in the experimental sample and J_o observations per type k in the observational sample, with $J_e << J_o$, so that the total number of observations in each sample is $N_e = J_e \times K$ and $N_o = J_o \times K$, and thus the total sample size is $N = N_e + N_o$.

For the experimental sample, I assign treatment according to

$$d_i = \mathbb{1}(i \bmod 2 = 1) \tag{51}$$

so that all observations with an even index are treated. Since I generate the indices such that every type k has the same number of even values, and J_e itself will be even, half of all observations per type k receive treatment in the experimental sample.

For the observational sample, I assign treatment according to

$$d_i = \mathbb{1}(x_i > 0) \tag{52}$$

so that assignment to treatment is confounded by x_i for the observational sample.

For the Monte Carlo trials, I assume the following values for the parameters:

$$
\begin{aligned}
K &= 100 \\
J_e &= 10 \\
J_o &= 100 \\
\alpha &= 1 \\
\beta_0 &= 1 \\
\sigma_\beta &= 1 \\
\frac{\sigma_\beta}{\sigma_\epsilon} &\in \{.3, .4, .5, .6, .7, .8, .9, 1, 2, 3, 4, 5\} \\
\sigma_x &= 1.
\end{aligned}
$$

I generate data according to this process 500 times for each value of the signal-to-noise ratio $\frac{\sigma_\beta}{\sigma_\epsilon}$ and estimate the vector β_k using the cross sample empirical Bayes strategy discussed earlier in this section. For purposes of comparison, I also estimate β_k using standard empirical Bayes applied only to the experimental data, the unpooled estimator (i.e., the maximum likelihood estimator) applied only to the experimental sample, the unpooled estimator applied only to the observational sample, and the pooled estimator (i.e., the projection assuming no heterogeneity so that $\beta_k = \beta \; \forall k$) applied only to the experimental data. For each trial, I measure the estimator's performance by calculating the mean squared error of the estimates with respect to the true vector of heterogeneous effects β_k, and I report the average mean squared error for each estimator across the 500 trials for each value of the signal-to-noise ratio $\frac{\sigma_\beta}{\sigma_\epsilon}$.

Figure A.3 plots the results. Observe that for small values of the signal-to-noise ratio, the standard empirical Bayes procedure offers substantial gains in mean squared error relative to the unpooled (maximum likelihood estimator) estimate, while the consensus procedure proposed in this paper offers modest gains relative to standard empirical Bayes. As the signal-to-noise ratio becomes large, the unpooled and standard empirical Bayes estimate converge, while the consensus procedure using all of the data continues to offer substantial improvements relative to the unpooled estimate. Intuitively, this occurs because empirical Bayes using all of the data brings new information to the table that is unavailable to the unpooled estimator or to the standard empirical Bayes estimator; synthesizing the experimental and observational data effectively increases the sample size by increasing the pool of information used to estimate the heterogeneous treatment effects.

[Figure A.3 about here.]

Estimator Performance Calibrated to Empirical Data-Generating Process

To assess the estimator's performance for the actual application in this paper, I perform a Monte Carlo simulation with a DGP that matches the actual model I brought to the data and with the DGP calibrated to the data's empirical distribution.

Before describing the DGP and simulation, I first introduce some notation. Let y_{it} denote the math score of student i at time t, and let X_{it} denote a vector of student covariates that includes the same set of covariates used to estimate student heterogeneity in the main text. For ease of notation, also assume that X_{it} contains an intercept. Let w_{it} denote student i's ELA score from the prior period, and let d_{it} denote an indicator for whether student i participates in choice at time t. Let $(y, X, w, d)^e$ denote the entire experimental sample so that a row within the experimental sample is given by $(y, X, w, d)^e_{it} = (y_{it}, X_{it}, w_{it}, d_{it})^e$. Similarly, let $(y, X, w, d)^o$ denote the observational sample.

I assume the experimental sample is generated according to the following process:

$$y_{it} = \beta_k d_{it} + \Gamma_e X_{it} + u_{it} \tag{53}$$

$$u_{it} = \delta_e w_{it} + \epsilon_{it} \tag{54}$$

$$\beta_k = \alpha X_{it} \tag{55}$$

$$(d_{it}, X_{it}, w_{it}) \sim F_e(d, X, w) \tag{56}$$

$$\epsilon_{it} \sim G_e(\epsilon), \tag{57}$$

where α is calibrated to the empirical DGP by projecting the consensus posterior modes for the experimental data found in the main text onto the corresponding covariates X_{it};[43] Γ_e and δ_e are calibrated to the empirical DGP by imposing the appropriate values for α and by recovering the relevant parameters from the model implied by equations (53) and (54), as estimated on the actual lottery data. F_e is determined by the empirical distribution of the lottery data but imposing that $d_{it}|X_{it} \perp w_{it}$,[44] and G_e is determined by the empirical distribution of residu-

[43] To be precise, $\alpha = [(X'_{it}X_{it})^{-1}X'_{it}\hat{\beta}^s_{k(i,t)}]'$

[44] To be precise, I create draws from F_e by resampling with replacement from the actual lottery data within "subgroups," as defined by the interaction of all binary variables contained in (X_{it}, d_{it}). Resampling the data within subgroups ensures that the fraction "treated" within each of these subgroups matches the empirical distribution and hence ensures that all of the necessary interaction terms are identified within any given Monte Carlo iteration. While it is a low probability event, over many iterations, naively resampling from the entire distribution can result in draws where important parameters are not identified. After resampling, I then randomly permute the treatment vector d within subgroups. I do this to break the correlation between treatment and the variable w_{it} that will play the role of the "omitted variable" in the simulation and hence ensure that treatment is "as-if" randomly assigned in the experimental sample conditional on X_{it}.

als found by imposing the appropriate values for $(\alpha, \Gamma_e, \delta_e)$ and by calculating residuals using model (53) as applied to the lottery data.[45]

I assume the observational sample is generated according to the following process:

$$y_{it} = \beta_k d_{it} + \Gamma_o X_{it} + u_{it} \tag{58}$$

$$u_{it} = (\delta_o + \chi) w_{it} + r\epsilon_{it} \tag{59}$$

$$\beta_k = \alpha X_{it} \tag{60}$$

$$d_{it} \sim bernoulli(p_w) \tag{61}$$

$$(X_{it}, w_{it}) \sim F_o(X, w) \tag{62}$$

$$\epsilon_{it} \sim G_o(\epsilon), \tag{63}$$

where α is identical to the parameter values used for the experimental sample; Γ_o and δ_o are calibrated to the empirical DGP by imposing the appropriate values for α and by recovering the corresponding parameters from the model implied by (53) and (54), as estimated using the actual non-experimental data. χ is a free parameter that I use to vary the amount of selection in the observational data; $r = r(\chi)$ is a rescaling factor that, for small values of χ, ensures the variance of y_{it} is not a function of χ.[46] F_o is determined by the empirical distribution of the non-experimental data.[47] The parameter of the Bernoulli distribution is given by $p_w = p(w_{it}) = p_1 \mathbb{1}[w_{it} > 0] + (1 - p_1)(1 - \mathbb{1}[w_{it} > 0])$ so that students of different ability levels have different propensities to take up the treatment, and $p_1 = \mathbb{E}(d_{it}|w_{it} > 0)$ is calibrated to match the corresponding empirical probability in the non-experimental data. G_o is determined by the empirical distribution of residuals found by imposing the appropriate values for $(\alpha, \Gamma_o, \delta_o)$, setting $\chi = 0$ and $r = 1$, and calculating residuals using the model implied by equations (58) and (59), as applied to the actual non-experimental data.[48]

Finally, I set the experimental and observational sample sizes to the actual sample sizes of

[45] To be precise, the distribution of residuals that determines G_e is constructed using the lottery data according to $\epsilon = (I - \Omega(\Omega'\Omega)^{-1}\Omega')y$, where $\Omega = (\alpha X_{it} d_{it}, X_{it}, w_{it})$ with $\alpha = [(X'_{it}X_{it})^{-1}X'_{it}\hat{\beta}^s_{k(i,t)}]'$.

[46] To be precise, $r(\chi) = \sqrt{max[1 - \frac{var(\chi w) - 2cov(\chi w, \beta_k d_{it} + \Gamma_o X_{it} + \chi w)}{var(\epsilon)}, 0]}$.

[47] To be precise, I create draws from F_o by resampling with replacement from the actual non-experimental data within "subgroups" as defined by the interaction of all binary variables contained in (X_{it}, d_{it}). Resampling the data within subgroups ensures that the fraction "treated" within each of these subgroups matches the empirical distribution and hence ensures that all of the necessary interaction terms are identified within any given Monte Carlo iteration. While it is a low probability event, over many iterations naively resampling from the entire distribution can result in draws where important parameters are not identified.

[48] To be precise, the distribution of residuals that determines G_o is constructed with the non-experimental data according to $\epsilon = I - \Omega(\Omega'\Omega)^{-1}\Omega'y$, where $\Omega = (\alpha X_{it} d_{it}, X_{it}, w_{it})$ with $\alpha = [(X'_{it}X_{it})^{-1}X'_{it}\hat{\beta}^s_{k(i,t)}]'$.

the lottery (1,705) and non-experimental (6,549,949) data used for the main analysis. I allow the selection parameter χ to vary from zero to ten in increments of 0.5.

For the Monte Carlo trial, I proceed as follows:

0. Fix the selection parameter χ.

1. Draw the experimental sample $(d, X, w, \epsilon)^e$ according to the distributions F_e and G_e, and then construct the outcome for the experimental sample (y) according to equations (53), (54), and (55).

2. Draw the observational sample $(X, w, \epsilon)^o$ from distributions F_o and G_o, construct treatment status from equation (61), and then construct the outcome for the observational sample (y) according to equations (58), (59), and (60).

3. Delete the unobserved potential confounder (w_{it}) from both the experimental and non-experimental data.

4. Estimate the heterogeneous treatment effects by applying the same model and empirical Bayes procedure used with the actual data in the main text to the simulated data from steps one through three.[49]

5. Calculate the mean squared error of the estimated treatment effect heterogeneity relative to the actual vector of heterogeneous effects in the experimental simulation sample. I perform this calculation for the pooled estimate found using only the experimental simulation sample, the unpooled estimate found using only the experimental simulation sample, the unpooled estimate found using only the observational simulation sample, the standard empirical Bayes estimate found using only the experimental simulation sample, and the consensus empirical Bayes estimate found using both the experimental and observational simulation samples.

6. Store these calculations.

7. Repeat steps one through six for 100 iterations.

8. Average the mean squared error calculations for each estimator over the stored values from the 100 iterations.

[49] I do not include fixed effects in either the DGP or the estimation step of the simulations for computational reasons. Incorporating fixed effects into the simulation DGP would cause the simulations to take weeks to converge.

7. Change the value of the selection parameter χ, and return to step 1.

The results are displayed in Figure A.4.

[Figure A.4 about here.]

At low values of selection, the empirical Bayes procedure using all of the data dominates the corresponding unpooled and standard empirical Bayes estimates. This occurs despite the fact that the treatment effect heterogeneity in this simulation (as it is in the main text) is a linear combination of binary and other nonnormally distributed variables and hence is also not normally distributed. Therefore the empirical Bayes procedure using all of the data can still work well, even under violations of normality.

At large values of the selection parameter, we see that the standard empirical Bayes procedure and the empirical Bayes procedure using all of the data converge. However, it is important to note here that such large values of selection are only possible if we also increase the total variance of y in the non-experimental sample; hence, it is not clear whether the convergence is the result of the increasing bias or the corresponding decrease in the signal-to-noise ratio.

Taken together, the simulation suggests that the empirical Bayes procedure using all of the data weakly dominates among consistent estimators of the treatment effect heterogeneity for this DGP when calibrated to match the actual data.

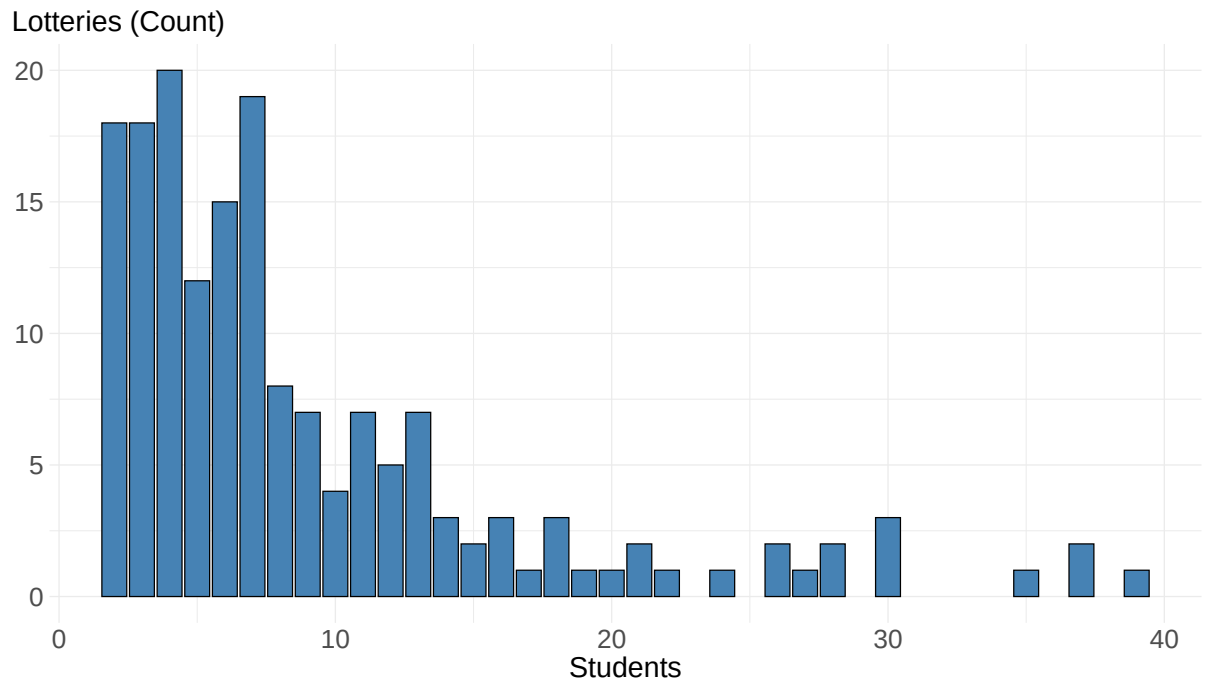

Figure A.1: Lottery Distribution by Size

Note: This figure displays the distribution of lottery size (i.e., the number of students involved in randomization) for the districts where I collected data. I define a lottery as occurring at the entry grade by academic year in the receiving district, provided some form of randomization was used to ration an open seat.

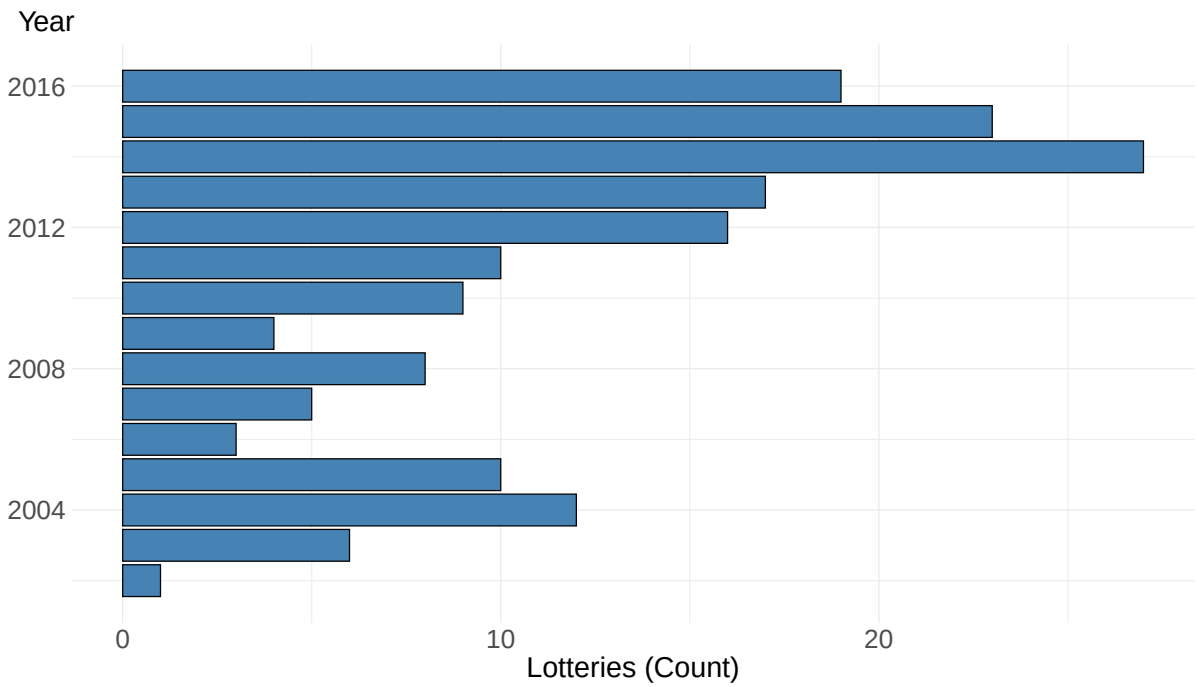

Figure A.2: Lottery Distribution by Year

Note: This figure displays the distribution of lottery years (i.e., the number of lotteries I observe in each year) for the districts where I collected data. I define a lottery as occurring at the entry grade by academic year in the receiving district provided some form of randomization was used to ration an open seat.

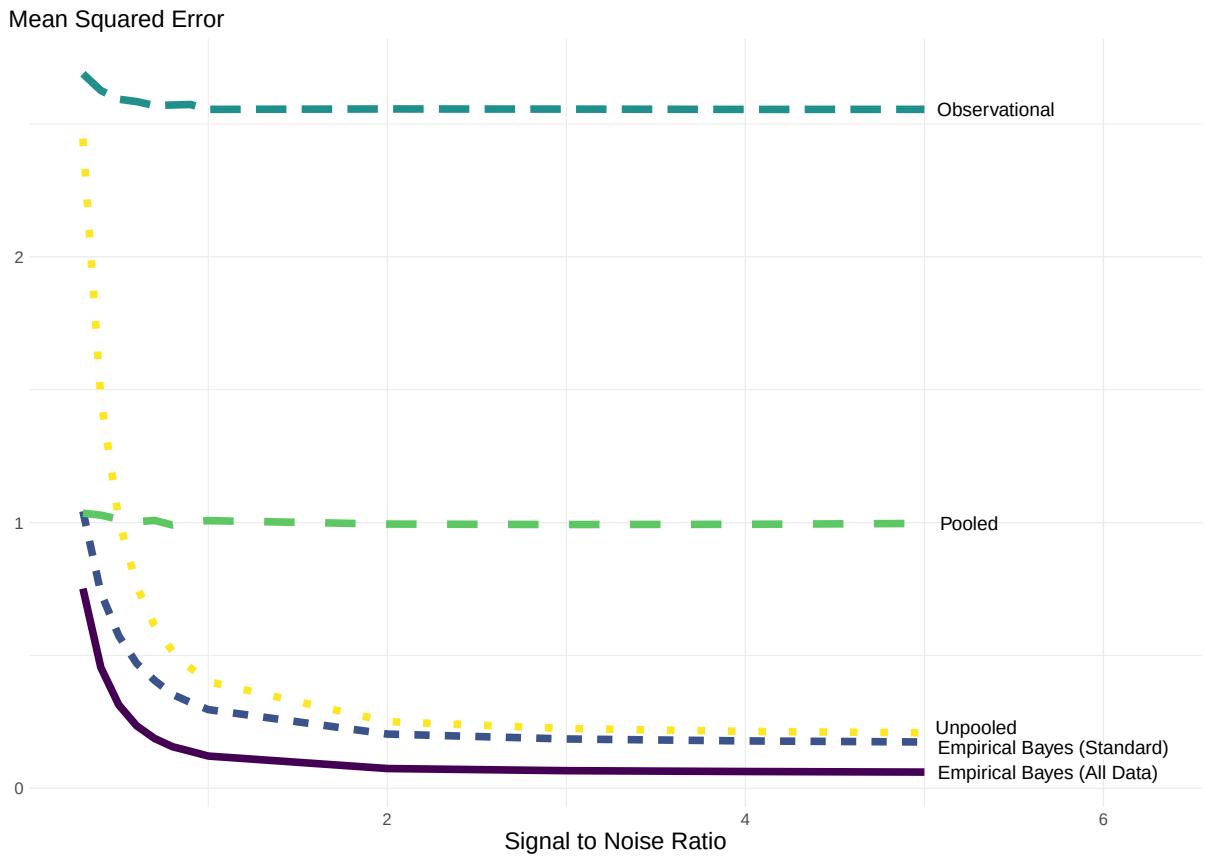

Figure A.3: Estimator Performance Varying the Signal-to-Noise Ratio ($\frac{\sigma_\beta}{\sigma_\epsilon}$)

Note: This figure displays the average mean squared error across 500 Monte Carlo trials for various estimators of treatment effect heterogeneity (β_k) using the data-generating process described in equations (47)–(52) at different values of the signal-to-noise ratio ($\frac{\sigma_\beta}{\sigma_\epsilon}$). "Pooled" refers to the maximum likelihood estimator using only the experimental data and assuming there is no heterogeneity in treatment effect (i.e., $\beta_k = \beta \; \forall k$). "Unpooled" refers to the maximum likelihood estimator using only the experimental data but allowing for treatment effect heterogeneity. "Empirical Bayes (Standard)" refers to the empirical Bayes estimator for the treatment effect heterogeneity applied only to the experimental data. "Observational" refers to the maximum likelihood estimator for the treatment effect heterogeneity using only the observational data and under the (incorrect) assumption that treatment is randomly assigned in this sample. And "Empirical Bayes (All Data)" refers to the empirical Bayes estimator that uses all of the available data, both experimental and observational.

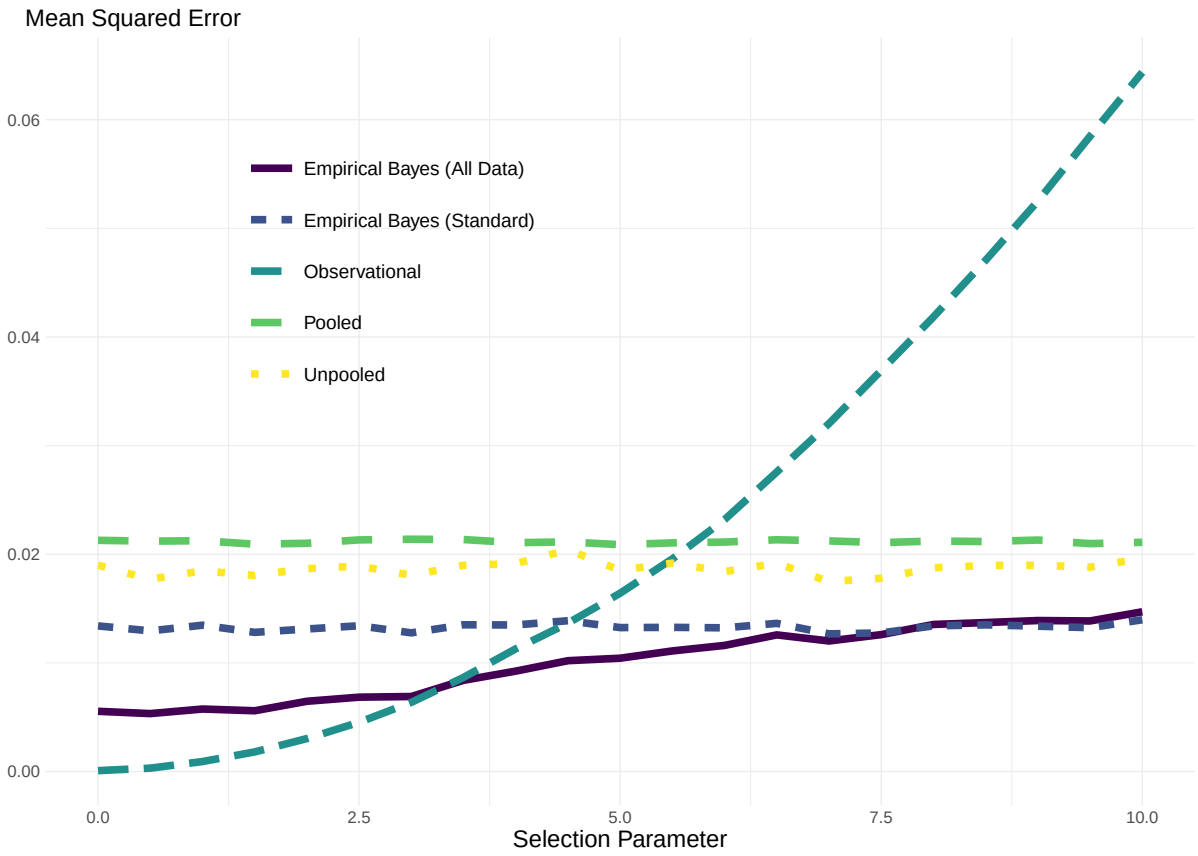

Figure A.4: Estimator Performance Calibrated to Empirical Data-Generating Process

Note: This figure shows that the empirical Bayes procedure using all of the data performs well with a data-generating process (DGP) calibrated to match the actual data from Massachusetts and using a model that approximates the treatment effect heterogeneity linearly with nonnormally distributed variables. For a complete description of the DGP used for this simulation, so the discussion in Section C.3. Each line in this figure gives the average mean squared error over 100 Monte Carlo trials for the indicated estimator at different amounts of selection in the non-experimental data. "Pooled" refers to the maximum likelihood estimator using only the experimental data and assuming there is no heterogeneity in treatment effect (i.e., $\beta_k = \beta \; \forall k$). "Unpooled" refers to the maximum likelihood estimator using only the experimental data but allowing for treatment effect heterogeneity. "Empirical Bayes (Standard)" refers to the empirical Bayes estimator for the treatment effect heterogeneity applied only to the experimental data. "Observational" refers to the maximum likelihood estimator for the treatment effect heterogeneity using only the observational data and under the (incorrect) assumption that treatment is randomly assigned in this sample. And "Empirical Bayes (All Data)" refers to the empirical Bayes estimator that uses all of the available data, both experimental and observational.

Table A.1: Testing for Differential Attrition

	Observed After Choice Year			
	(1)	(2)	(3)	(4)
Initial Offer	−0.001	−0.001	−0.002	0.002
	(0.007)	(0.007)	(0.008)	(0.008)
Waitlist Offer	0.058	0.060	0.043	0.045
	(0.052)	(0.054)	(0.051)	(0.043)
Lottery Number	−0.0003	−0.0003	−0.001	−0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Waitlist Number	−0.002	−0.002	−0.003	−0.002
	(0.002)	(0.002)	(0.002)	(0.003)
Demographic Controls	No	Yes	Yes	Yes
Baseline Controls	No	No	Yes	Yes
Baseline Test Scores	No	No	No	Yes
Probability Observed Post-Choice Year	0.94	0.94	0.94	0.94
Observations (Students)	1905	1905	1571	1292
Adjusted R^2	0.672	0.673	0.737	0.775

Note: This table shows specifications where I predict an indicator for observing a student in the state data in a postlottery year using the vector of lottery instruments. All specifications contain lottery fixed effects.[50] I drop students who I am unable to match to the state data, were not involved in competitive lotteries, received a sibling preference, or applied late. Column (1) presents results for the entire lottery sample, and column (2) is identical to column (1) but controls for demographic characteristics. Column (3) further restricts the sample to students I observe in the data at baseline and includes baseline controls for any disability, subsidized lunch status, and English language learner status. Column (4) further restricts the sample to the set of students for whom I observe at least one baseline test score and includes a test score average observed prior to the lottery as controls.

Table A.2: Falsification Test: Math

	Math Score (Baseline)					
	(1)	(2)	(3)	(4)	(5)	(6)
Initial Offer	0.14	0.06	0.15	0.07	0.01	−0.01
	(0.06)	(0.06)	(0.07)	(0.06)	(0.04)	(0.04)
Waitlist Offer	0.36	0.19	0.32	0.15	0.04	0.08
	(0.27)	(0.27)	(0.25)	(0.27)	(0.18)	(0.19)
Lottery Number	0.01	0.003	0.01	0.002	−0.001	−0.001
	(0.004)	(0.003)	(0.004)	(0.004)	(0.002)	(0.002)
Waitlist Number	−0.004	−0.001	−0.004	0.002	−0.01	−0.005
	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)
Sample	All	All	2-test	2-test	2-test	2-test
Baseline Controls	No	Yes	No	Yes	No	Yes
Pre-Baseline Test Score	No	No	No	No	Yes	Yes
Observations	1,265	1,265	1,025	1,025	1,025	1,025
Adjusted R^2	0.06	0.23	0.05	0.23	0.71	0.72

Note: This table shows specifications that predict baseline student math test scores with the vector of lottery instruments. I drop students who I am unable to match to the state data, did not have a baseline test score, were not involved in competitive lotteries, received a sibling preference, or applied late. All specifications contain a set of lottery fixed effects. Column (1) presents the uncontrolled comparison, and column (2) is identical to column (1) except that it also includes demographic variables as controls. Column (3) restricts the sample to the set of students for whom I observe at least two test scores prior to the lottery and presents results for the otherwise uncontrolled comparison. Column (4) is identical to column (3) except that it also includes demographic variables as controls. Column (5) is identical to column (3) except that it includes the earlier prelottery test score as a control, and column (6) is identical to column (3) except that it controls for both demographic characteristics and the earlier prelottery test score.

Table A.3: Falsification Test: English Language Arts

	ELA Score (Baseline)					
	(1)	(2)	(3)	(4)	(5)	(6)
Initial Offer	0.14	0.08	0.17	0.11	0.03	0.03
	(0.06)	(0.05)	(0.06)	(0.06)	(0.04)	(0.04)
Waitlist Offer	0.13	−0.09	0.11	−0.07	0.05	0.07
	(0.28)	(0.26)	(0.27)	(0.26)	(0.17)	(0.17)
Lottery Number	0.01	0.003	0.01	0.003	0.002	0.002
	(0.003)	(0.003)	(0.004)	(0.003)	(0.002)	(0.002)
Waitlist Number	−0.01	−0.01	−0.02	−0.01	−0.02	−0.01
	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)
Sample	All	All	2-test	2-test	2-test	2-test
Baseline Controls	No	Yes	No	Yes	No	Yes
Pre-Baseline Test Score	No	No	No	No	Yes	Yes
Observations	1,265	1,265	1,025	1,025	1,025	1,025
Adjusted R^2	0.12	0.30	0.12	0.31	0.60	0.63

Note: This table shows specifications that predict baseline student English Language Arts (ELA) test scores with the vector of lottery instruments. I drop students who I am unable to match to the state data, did not have a baseline test score, were not involved in competitive lotteries, received a sibling preference, or applied late. All specifications contain a set of lottery fixed effects. Column (1) presents the uncontrolled comparison, and column (2) is identical to column (1) except that it also includes demographic variables as controls. Column (3) restricts the sample to the set of students for whom I observe at least two test scores prior to the lottery and presents results for the otherwise uncontrolled comparison. Column (4) is identical to column (3) except that it also includes demographic variables as controls, and column (5) is identical to column (3) except that it includes the earlier prelottery test score as a control. Column (6) is identical to column (3) except that it controls for both demographic characteristics and the earlier prelottery test score.

Table A.4: Additional Controls

	Standardized Test Score					
	Math	Math	Math	ELA	ELA	ELA
	(1)	(2)	(3)	(4)	(5)	(6)
Choice	0.19	0.15	0.17	0.01	−0.02	−0.06
	(0.08)	(0.08)	(0.07)	(0.08)	(0.08)	(0.07)
F-Stat Excluded Instruments	117.2	115.8	114.6	114.9	118.2	112.3
Lottery Fixed Effects	Yes	Yes	No	Yes	Yes	No
Baseline Test Score	Yes	Yes	No	Yes	Yes	No
Year Fixed Effects	Yes	Yes	No	Yes	Yes	No
Demographic Controls	No	Yes	No	No	Yes	No
Student Fixed Effects	No	No	Yes	No	No	Yes
Year x Lottery Fixed Effects	No	No	Yes	No	No	Yes
Observations	1702	1702	5113	1702	1702	5167
Observations (Students)	959	959	959	961	961	961
Adjusted R^2	0.65	0.67	0.78	0.56	0.59	0.72

Note: Columns (1) and (4) replicate the 2SLS results from Table 3. Columns (2) and (5) add controls for demographic characteristics and other baseline observables. These include indicators for race, ethnicity, and gender; age (measured in days); whether the student received a subsidized lunch at baseline or had any disability at baseline; number of days suspended at baseline; number of unexcused absences at baseline; and an indicator for whether the student was labeled as an English language learner at baseline. Columns (3) and (6) expand the sample to include all student years observed prior to the lottery and replaces all controls with a set of student fixed effects and year-by-lottery fixed effects. Hence columns (3) and (6) represent an IV difference-and-difference design that generates reduced-form and first-stage estimates by comparing the trends in test scores and choice status across lottery winners and losers relative to the lottery's date.

Table A.5: Initial Offer Lotteries

	Standardized Test Score			
	Math	Math	ELA	ELA
	(1)	(2)	(3)	(4)
Choice	0.19	0.21	0.01	0.01
	(0.08)	(0.09)	(0.08)	(0.09)
F-Stat Excluded Instruments	117.2	231.5	114.9	223.1
Instruments	All	Initial Offer	All	Initial Offer
Observations	1,702	1,556	1,702	1,556
Observations (Students)	959	874	961	875
Adjusted R^2	0.65	0.65	0.56	0.56

Note: Columns (1) and (3) replicate the main specification from Table 3 using all instruments. Columns (2) and (4) drop students who were not involved in initial offer lotteries and only uses the initial offer instrument in the first stage.

Table A.6: Omitting Idiosyncratic Lotteries

	Standardized Test Score					
	Math	Math	Math	ELA	ELA	ELA
	(1)	(2)	(3)	(4)	(5)	(6)
Choice	0.19	0.19	0.19	0.01	0.01	0.01
	(0.08)	(0.08)	(0.09)	(0.08)	(0.08)	(0.09)
Admission Round Lotteries	Yes	No	Yes	Yes	No	Yes
Waitlist Request Lotteries	Yes	Yes	No	Yes	Yes	No
Observations	1,702	1,680	1,606	1,702	1,680	1,605
Observations (Students)	959	937	886	961	939	887
Adjusted R^2	0.65	0.65	0.65	0.56	0.56	0.56

Note: Columns (1) and (4) replicate the specifications from Table 3. Columns (2) and (5) drop the lottery with admission rounds, and columns (3) and (6) drop the waitlist request lotteries.

Table A.7: Predictors of District Participation

	Choice Status			
	(1)	(2)	(3)	(4)
Students (thousands)	−0.04	−0.01	0.01	0.04
	(0.03)	(0.01)	(0.03)	(0.04)
Average Math Score	−0.30	−0.39	0.04	0.05
	(0.16)	(0.16)	(0.06)	(0.07)
Average ELA Score	−0.18	−0.15	−0.04	−0.04
	(0.15)	(0.16)	(0.05)	(0.05)
100x(Fraction White)	0.01	−0.001	−0.01	−0.01
	(0.01)	(0.004)	(0.01)	(0.005)
100x(Fraction Asian)	0.01	−0.002	−0.01	−0.01
	(0.01)	(0.01)	(0.01)	(0.01)
100x(Fraction Hispanic)	−0.01	−0.01	−0.004	−0.004
	(0.004)	(0.003)	(0.005)	(0.004)
100x(Fraction ELL)	−0.001	−0.004	−0.004	−0.004
	(0.01)	(0.01)	(0.003)	(0.003)
100x(Fraction HQ Teachers)	−0.01	−0.01	0.002	0.002
	(0.003)	(0.003)	(0.002)	(0.002)
Student-Teacher Ratio	−0.07	−0.07	−0.002	−0.002
	(0.02)	(0.02)	(0.01)	(0.01)
Per-Pupil-Spending: Instruction	−0.08	−0.05	0.04	0.01
	(0.03)	(0.02)	(0.02)	(0.01)
Per-Pupil-Spending: Pupil Services	0.15	0.23	0.03	0.04
	(0.08)	(0.07)	(0.04)	(0.04)
Per-Pupil-Spending: Teachers	−0.01	−0.001	−0.05	−0.04
	(0.04)	(0.04)	(0.02)	(0.02)
Metco Students (tens)	−0.01	−0.01	0.01	0.02
	(0.005)	(0.005)	(0.02)	(0.02)
Estimation Method	OLS	Post-Lasso	OLS	OLS
Additional Variables	Yes	No	Yes	No
District Fixed Effects	No	No	Yes	Yes
Year Fixed Effects	No	No	Yes	Yes
Dependent Variable Mean	0.55	0.55	0.55	0.55
F-Stat (Projected)	14.6	26.43	0.91	1.28
Observations	2280	2280	2280	2280
Observations (Districts)	285	285	285	285
Adjusted R^2	0.34	0.31	0.88	0.88

Note: This table contains specifications that predict a year-by-district indicator for participation in inter-district choice since 2009 with district-level observable characteristics. Column (1) estimates the model with OLS, and column (2) presents results from a post-Lasso regression where model selection was performed over the set of variables included in column (1). For visual clarity, the only variables displayed in the table are those that were Lasso selected. See the discussion in Section B.4 for a complete list of predictors. Columns (3) and (4) include district and time fixed effects and the time-varying observables from columns (1) and (2), respectively.

Table A.8: Intensive Margin Coursework Results

	Number of Classes			
	AP	Remedial	General	Advanced
Choice	0.14	−0.10	0.35	0.79
	(0.10)	(0.10)	(0.45)	(0.20)
Mean Dependent Variable	0.3	0.23	9.25	0.91
Observations	805	2,413	2,413	2,413
Observations (students)	467	911	911	911
Adjusted R^2	0.27	0.10	0.20	0.32

Note: The results in this table come from the two-stage least squares model outlined in equations (1) and (2) from the main text. All of the table's estimates are from specifications that use my preferred set of controls. These include a lottery fixed effect, a baseline average of test scores observed prior to randomization, academic year and grade fixed effects, indicators for PARCC testing, indicators for whether or not a student was matched to the state data via an exact or fuzzy process, and indicators for waitlist requests or admissions rounds. The sample used for estimation includes all students from the lottery data that I was able to match to the state data who were involved in competitive lotteries, did not receive a sibling preference, did not apply late, and for whom I observe at least one test score prior to randomization. I also drop a small number of students who appear in my data only prior to the period when the department of education in Massachusetts kept records on student coursework. The column labeled "AP" further restricts the sample to observations that occur in grades 11 and 12, since Advanced Placement courses are typically unavailable to students in earlier grades.